Principles for Designing Context-Aware Applications for Physical Activity Promotion

by

Gaurav Paruthi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2018

Doctoral Committee:

Associate Professor Mark W. Newman, Chair
Associate Professor Natalie Colabianchi
Assistant Professor Predrag 'Pedja' Klasnja
Professor Ken Resnicow

Gaurav Paruthi

gparuthi@umich.edu

ORCID iD: 0000-0002-5100-1578

Acknowledgement

I owe the most thanks to my supervisor Mark Newman for his unconditional support and guidance during my Ph.D. I most enjoyed the freedom given to me to pursue multiple intellectual streams. At times, I wasn't sure how things would converge, but Mark's confidence in me allowed me to keep pursuing the things that interest me the most. From choosing research projects to exploring startup opportunities, I was fortunate to have an advisor who encouraged and supported me in my endeavors. This freedom allowed me to explore three state of the art technologies- crowdsourcing, machine learning, and hardware prototyping that I was deeply excited about and then successfully explored through the projects described in this dissertation.

I want to thank my dissertation committee members: Pedja Klasnja, Natalie Colabianchi, and Ken Resnicow. Their support and feedback helped me to focus and take the individual projects to completion. I would also like to note that I wouldn't have had the opportunity to explore the area of Health behavior without Pedja's support. I was always motivated by his excitement and vision of how research and design can have a significant impact in improving the health behavior of people.

I was lucky to have great mentors throughout my Ph.D. and even before I came to Ann Arbor. I feel incredibly fortunate to be advised my Bill Thies at Microsoft Research who inspired me to pursue research and showed how hard work and dedication could have a positive impact on the lives of people. Thanks to Joyojeet Pal, who taught me two of the most memorable courses I took and for his mentoring throughout my Ph.D. journey. Many thanks to Mark Ackerman for his supervision during my earlier years. I also thank Qiaozhu Mei for being a great teacher and thereby influencing my research interests towards machine learning.

I thank my colleagues and friends who helped make this long journey enjoyable. Thanks to my lab mates at the Interaction Ecologies Group- Tao Dong, Rayoung Yang, Yungju Chang, Chuanche Huang and Shriti Raj for being excellent collaborators and providing constant feedback to help me and my work reach the high quality. Thanks to my friends Priyank Chandra and Gaurav Singhal, with whom I brainstormed countless ideas, learned new things and shared some great memories. I also thank Seongseok You, Cindy Lin, Seyram Avle, Hariharan Subramonyam, Tawfiq Ammari, Daphne Chang, my Ph.D. cohort, and many other friends for being around to help out whenever needed.

Finally, this journey wouldn't have been completed without the support of my family. My mom and dad have been a tremendous source of inspiration and support during the long time I have been away from home. My brothers Saurab, Anshul and Adit for always being there and sharing many exciting adventures. Last but not the least a big thanks to my partner Chenfang who kept me going all these years and gave me many reasons to be excited about all that the future holds.

Contents

List of Figures

List of Tables

Abstract

Mobile devices with embedded sensors have become commonplace, carried by billions of people worldwide. Their potential to influence positive health behaviors such as physical activity in people is just starting to be realized. Two critical ingredients, an accurate understanding of human behavior and use of that knowledge for building computational models, underpin all emerging behavior change applications. Early research prototypes suggest that such applications would facilitate people to make difficult decisions to manage their complex behaviors. However, the progress towards building real-world systems that support behavior change has been much slower than expected. The extreme diversity in real-world contextual conditions and user characteristics has prevented the conception of systems that scale and support end-users' goals.

We believe that solutions to the many challenges of designing context-aware systems for behavior change exist in three areas: building behavior models amenable to computational reasoning, designing better tools to improve our understanding of human behavior, and developing new applications that scale existing ways of achieving behavior change. With physical activity as its focus, this thesis addresses some crucial challenges that can move the field forward.

Specifically, this thesis provides the notion of sweet spots, a phenomenological account of how people make and execute their physical activity plans. The key contribution of this concept is in its potential to improve the predictability of computational models supporting physical activity planning. To further improve our understanding of the dynamic nature of human behavior, we designed and built Heed, a low-cost, distributed and situated self-reporting device. Heed's single-purpose and situated nature proved its use as the preferred device for self-reporting in many contexts. We finally present a crowdsourcing system that leverages expert knowledge to write personalized behavior change messages for large-scale context-aware applications.

Chapter 1 Introduction

Mobile computing devices such as smartphones, smartwatches, and fitness trackers wearable devices are carried by an ever-increasing number of people, and provide an unprecedented opportunity for enabling applications that improve people's lives. Such devices are outfitted with sensors that enable them, with appropriate software, to be context-aware. That is, we can build systems that understand the actions of users and situations they are in, which in turn allows applications to deliver functionality that is highly personalized and situation-relevant.

Wide adoption of mobile smartphones equipped with sensors, GPS, camera, and Bluetooth detectors has enabled software companies, as well as researchers to leverage these devices to understand the actions of users and the situations in which they find themselves. Some commonly used applications include Google search personalizing its search to the physical location of the user, or Apple's Siri, when asked about the gas station, being able to find the gas station that is open now and closest to the user. Taking a step forward, the Google Assistant service on Android devices can proactively provide the information that might be useful for the user. For instance, based on the location, nearby traffic and the current time, Google Assistant can estimate the time when the user should leave for the airport. Similarly, The Nest thermostat is able to proactively change the temperature of the user's home based on many different contextual features including user's presence, time of day and activity history.

This widespread adoption of such "smart" systems has made computing increasingly closer to the era of ubiquitous computing envisioned by Mark Weiser (Weiser 1991), where heterogeneous computing devices respond to users' context and activity. An important aspect of such a computing environment is *context-awareness*, which refers to the ability of computing devices to adapt their behavior to sensed context and users' behavior.

As the venues for human-computer interaction are diversifying, designers of context-aware applications are looking at solving important challenges in many different domains including Health and Sustainability. While the use of fitness trackers and health accessories are growing, their effectiveness is in doubt (Goasduff and Pettey 2012). Data indicates that 50% of customers stop using their fitness tracking devices after six months of buying (Ledger and McCaffrey 2014). The dearth of a theoretical bases for building mobile applications for everyday use might be one of the reasons for such failure (Herrmann and Kim 2017). Although health researchers have explored and studied a wide range of strategies and interventions to promote physical activity, development of a more effective intervention to promote participation in regular physical activity remains an important challenge for researchers, clinicians and health authorities.

Identifying correlates and determinants of physical activity is critical for the development of effective programs and policies. Much research has been done towards understanding cognitive, social, and environmental influences on physical activity behavior (Bauman et al. 2012). Variables such as attitudes, intentions, outcome expectancies, social support, recreational facilities have been observed to have a modest association with physical activity in adults (Trost et al. 2002). What has been lacking is the knowledge temporal patterns in activity context. These could be useful in promoting physical activity. Recently, it has been found that factors such as perceived environment (Hekler et al. 2012), affective state (Dunton et al. 2014), social and physical contexts (Dunton et al. 2007) have been shown to have a strong effect. In line with the recent research efforts (Hekler, Michie, et al. 2016), the goal of this research is to fill the gaps where existing theoretical frameworks of behavior change are lacking.

This research delves more deeply into the potential use of context-aware systems for Health and Behavior Change. In this section, I first outline our vision of the higher-level architecture of a context-aware behavior change system. The system, at a higher level, consists of three main steps- gathering context, modeling behavior and generating a personalized intervention. Following this high-level description, I will describe some of the key challenges in the latter two steps, that will remain as the focus of the rest of the dissertation.

Vision of Context-Aware Systems for Behavior Change

This subsection presents an abstract context-aware system with the goal of highlighting the key challenges. In the literature, the phenomenological view of context treats human activity as an ongoing process of interpretation where the interaction between context and action shape each other (Dourish 2004). On the other hand, seeking a more operational definition, the positivist perspective defines context in terms of objective representations that form attributes of a computational system (Bauer, Newman, and Kientz 2014). This point of view aims to support the development of computational systems and thus is more relevant to our current discussion of the vision of context aware system for physical activity promotion.

When seen from the positivist point of view, Dey et al. 2001 provided a broad definition of context that incorporates other narrower definitions (Dey, Abowd, and Salber 2001): *"Context is any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. (Abowd et al. 1999)"* Following this definition, Dey et al. 2001 defined context-awareness of system as *"a system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task. (Abowd et al. 1999)"*

For software systems, researchers have proposed many life cycles of context (Perera et al. 2014). For instance, Ferscha et al. 2001 proposed a lifecycle which begins with sensing, then data representation, then rule base and then actuation. (Ferscha, Vogl, and Beer 2001). In another system, Bernardos et al. proposed three steps in such a system: context acquisition, information processing, and reasoning and decision (Bernardos, Tarrio, and Casar 2008). Inspired by such systems, we describe a context-aware system to comprise of three main steps (Figure 1) – data acquisition, context management and generate a personalized intervention.

Data Acquisition

We now have a large number of sensors deployed in mobile devices. Moreover, due to the growing popularity of household internet connected devices, such numbers are predicted to increase drastically over the next decade (Sundmaeker et al. 2010). Such deployment of sensors generate big data about user's context (Zaslavsky, Perera, and Georgakopoulos 2013). The usefulness of this data for any application will depend upon the techniques develop to analyze, interpret, and understand it. The research in context-aware computing applications has focused on solving these very challenges of potential applications. With a large amount of data being collected from many sensors dedicated to each user, one main challenge that arises is the lack of a solution that addresses aspects such as device management, interoperability, platform portability, and security and privacy (Perera et al. 2014). Although these aspects are very important, the challenges scoped in this dissertation do not relate to sensor data aggregation. Our focus remains on the latter two aspects of the system.

Context Management

Context management step is associated with the steps of the lifecycle where sensor data is modelled and represented in a meaningful way. During this step, more context features might be computed from other, lower-level contextual data. Researchers have used this distinction between sensed data and derived data to categorize context as well (Perera et al. 2014). Perera et al. 2014 identified primary context as that data which is received directly from a sensor (e.g. GPS coordinates as location information). Secondary context is any information that is computed from other contexts (e.g. distance between two sensors calculated from their GPS coordinates). Alternatively, Schilit et al. 1994 categorized context into three conceptual questions of Where you are (gps, location names, etc.), Who you are with (people present near the user), and what resources are nearby (nearby objects and services) (Schilit, Adams, and Want 1994). An ideal context management system may utilize a combination of such schemes.

Once context has been gathered, machine learning techniques such as logical rules, supervised learning, unsupervised learning, and fuzzy logic are leveraged to make decisions on the collected data. Such techniques are especially useful in contextual reasoning when there is a large amount of sensed data and contextual features. Using behavior modelling approaches, high level behavior patterns such as routines may also be inferred (Banovic et al. 2016).

Intervention Generator

The user-interaction offered by the system can determine the level of context-awareness of a system (Barkhuus and Dey 2003). Barkhuus and Dey 2003 identified three levels of context-awareness- personalization, passive context-awareness, and active context-awareness. A system may *personalize* the content it provides based on the preferences or context of the user, for a physical activity promotion application, if the user has been sedentary for a while at work, the system may send a message to the user encouraging her to stretch. A system may also be *passively context aware*, when it monitors the context information and offers recommendation to the user when the user wishes to interact. An example of this could be a navigation app which suggests nearby gas station when requested by the user. Finally, a system may be *actively context aware* if it automatically takes actions based on contextual reasoning. A typical example for such a system is a smart-thermostat which constantly monitors environment temperature to set heating/cooling of the air-conditioner.

The user-interaction offered by the system is largely determined by the intended application of the system. For physical activity promotion, such applications may be motivational, i.e. it may motivate the user to be more physically active by sending personalized content like quotes, or by nudging the user to set physical activity goals to be accomplished, or by presenting visualization of user's physical activity data. The interventions that personalize the content usually depend on computational logic (machine learning, rules, etc.) to figure out the right content for a user.

Challenges of Designing a Context-aware System for Physical Activity Promotion

In this thesis, I will contribute to the literature supporting the design of context aware systems for physical activity promotion in three main ways.

Better models need better understanding of specific behaviors

First, I will argue that our current theories about behavior are limited for building effective context-aware systems for physical activity promotion (Chapter 2). I will posit that there is a need to improve the predictability of computational models to support planning of physical activity. In chapter 3, I will present a qualitative study that supports the notion of sweet spots, a phenomenological account of how people create physical activity plans. This work posits that the notion of sweet spots increases the predictability of computational models that support creation of actionable plans.

Better understanding of behavior need better tools to support researchers

Second, in order to gain a deeper understanding of people's behavior, researchers need the right tools that provide rich data about people's behaviors within their temporal context. Experience Sampling Method, an approach involving frequent self-reporting prompts has been commonly used to get a deeper understanding of people's behaviors. Towards improving this method, in Chapter 4, I will present the design space for using single purpose devices for experience sampling. More specifically, we designed and evaluated Heed, a physically situated self-reporting device. Our findings show that Heed devices expand the coverage of captured experiences by complementing the use of Phones.

Better content generation

Thirdly, I will argue that contextualization requires personalized content in a scalable way. Depending upon the application, a system may require personalizing its interaction with the user to the given context. A subset of such applications is one that send contextualized suggestion to the user. Given the large number of possible situations, systems intended for such applications require a scalable way for content generation. In Chapter 5, I will present a crowd-sourcing system that generates physical activity promotion messages more cost-efficient manner than experts.

The Thesis Statement

In order to design context-aware applications for Physical Activity Promotion, it is necessary to:

- Understand the role of context in planning desired physical activity in order to develop the right abstractions that can be computationally modeled. *Sweet spots*, as a unified representation of converging factors, provides a novel and useful perspective to (a) better understand the nature of physical activity planning and execution, (b) bridge the gap between phenomenological and positivist perspectives of context to allow for computational support, and (c) inform the design of systems intended to support physical activity planning. (Chapter 3)

- Design data collection tools to gather behavioral traces that can be used to design better context-aware applications for physical activity promotion. We designed the HEED system comprising of portable light weight single purpose self-report devices that lower the burden of self-reporting in specific contexts. (Chapter 4)

- Design crowdsourcing systems that can generate personalized content for context-aware applications in a scalable way. We designed a crowdsourcing system that leverages expert knowledge to guide crowd workers to write messages for physical activity promotion. (Chapter 5)

Chapter 2 Related Work

Lack of physical activity is one of the most important behavioral risk factors for chronic diseases such as diabetes and coronary heart diseases. Yet, less than 20% of North Americans attain the recommended amount of physical activity (Troiano et al. 2008). The prevalence of chronic diseases continues to rise and is now responsible for over 70% of U.S. healthcare expenditures (Hoffman, Rice, and Sung 1996). To address this problem, researchers have sought to leverage technologies such as mobile phones, web applications, and social networking tools to encourage physical activity (Consolvo et al. 2006) due to their low cost, high penetration, and integration in people's everyday lives.

Increasing ownership of mobile phones and wearables (Smart Wearables Vendor Strategies, Opportunities, Forecasts n.d.), and the growing interest in deployable sensors (Internet of Things Consumer Devices: Ownership and Preferences n.d.) open new opportunities for Physical Activity promotion. People spend about 90% of their time in close proximity of such devices (Dey et al. 2011), thus enabling applications to potentially interact or intervene with the user many times within a day, with content that is personalized to the user. Moreover, the sensors used by these technologies enable rich streams of continuous data, for e.g. GPS, accelerometer, etc. The data streams along with statistical modeling techniques, provide a unique opportunity to understand user behavior in context of their fast changing social, biological, personal and environmental states. Moreover, these techniques can potentially allow applications to learn and respond to rapid changes in behavioral states and related contextual factors.

Interventions that are personalized to user input or sensed data are termed, elsewhere, as digital behavior change interventions (DBCIs) (Yardley et al. 2016). One class of such interventions are "just-in-time" adaptive interventions (JITAI). Just-in-time adaptive interventions (JITAI) seek to provide individuals with the right type of support at the right time and in the right context (Intille

2004; Patrick et al. 2008; Spruijt-Metz and Nilsen 2014). JITAIs are increasingly being explored in support of health behavior change in multiple domains, including physical activity and smoking cessation, while addressing issues related to mental illness and other chronic disorders. An imagined example of a DBCI is a personalized assistant for planning physical activities. The assistant should learn the preferences of the user over the multiple dimensions such as desired activities, weather, time of day and locations. Moreover, it may synthesize real-time data streams for the different contextual factors to provide the recommendations for when and where to exercise.

DBCIs require theories that take into account within-person variations in individual characteristics and dynamic contexts (Hekler, Klasnja, et al. 2016). Existing theories of behavior provide only inadequate understanding the inherent complexity of real-world behavior change and must be understood more in order to develop such interventions (Riley et al. 2011). Furthermore, development of such interventions requires the development of “precise, quantifiable computational models” that can represent the user’s context in a meaningful way (Hekler, Michie, et al. 2016).

This chapter lays the foundation for the rest of the work in this dissertation. In the following sub-sections, I first present existing evidence to the claim of why it is important to understand context. I then present the results from a few but important studies that unpack intra-individual differences in how contextual factors affect physical activity decisions. In the next sub section, I present existing arguments for the need of developing new theories and computational models for behavior change. I then present the few existing works that have modelled behavior in a dynamic way and also some applications that have applied machine learning techniques on data streams associated with contexts towards promoting physical activities. Finally, I summarize the claims made with this thesis with respect to the presented related work.

Role of Context in Physical Activity Promotion

Social cognition models, such as the Theory of Planned Behavior (Ajzen 1991), have characterized motivational factors, such as people's intention to perform a particular behavior, as the most proximal determinants of that behavior (Darker et al. 2010). Although important, meta-analytic work suggest that intention is not sufficient for promoting increased physical activity (Hagger, Chatzisarantis, and Biddle 2002). One potential explanation for this may the importance of establishing intentions within a given context. Implementation intentions is a concept that involves the articulation of intentions linked with particular situations (Gollwitzer 1999). A number of meta analytical works have highlighted that participants who form implementation intentions are significantly more likely to carry out a desired behavior such as exercise (Milne, Orbell, and Sheeran 2002; Prestwich, Lawton, and Conner 2003).

Positive outcomes for implementation intention interventions are, however, not shown in research done in naturalistic setting (Vet et al. 2009). Moreover, it was found that promoting more 'planful' actions, i.e. making daily plans for exercise, may hinder engagement in health behaviors based how stressed the individual is in the given context (Payne, Jones, and Harris 2010). Payne et al. also found that participants who were less stressed performed even more physical activity than participants who participated in planning. Such results indicate that contextual factors may play an even more important role than intentions in promoting physical activities.

Study of the interaction between context with intentions to influence physical activity has mostly used cross-sectional methods that measure the constructs at large time intervals such as months (Hekler et al. 2012). Recent growth in the use of experience sampling method (ESM) has allowed researchers to collect self-reports from participants, repeatedly, at a high frequency (Shiffman, Stone, and Hufford 2008), thus allowing analysis of not only between-person levels of physical activity but also within-person levels. Such analysis can provide better insights into the decision making and fluctuations for physical activities in participants (Dunton et al. 2009).

Most theories of health behavior change (such as the Theory of Planned Behavior (Ajzen 1991) and Social-Cognitive Theory (Bandura 2002)) aim to understand the role of psychosocial and environmental factors between people (i.e. inter-individual variation). The need for a more idiographic approach to behavioral science seeking the description of within-person processes (Lamiell 2003; Molenaar 2004) and the recent availability of methods such as ESM can allow researchers to study time-varying factors such as cognitions, mood, physiological states, and contexts (Dunton and Atienza 2009). Recent studies have uncovered cognitive factors (Dunton et al. 2009), affective factors (Dunton et al. 2009; Dunton et al. 2015), mood (Gauvin, Rejeski, and Reboussin 2000), social (Salvy et al. 2007; Salvy et al. 2008) and perceived environmental factors (Hekler et al. 2012) to influence physical activity. Yet the role of such factors over the course of the day is quite less (Liao, Shonkoff, and Dunton 2015) leading to a disconnect between health management practices and the context of health related activities, such as time of day, location, and daily activities (Klasnja and Pratt 2012).

An important contextual factor that has gained the interest of researchers is the Affective state of an individual. Earlier research hypothesized that affective states experienced during a behavior reinforce properties of a behavior such as physical activity [Loewenstein 2000, Updegraff 2004, Leone 2005]. Such works motivated lab studies that highlighted the direct relationship between affect and physical activities. More recently using ESM, researchers have studied how affective states differ across specific social and physical contexts. Evidence showed that exercising with people improves affective response while being outdoors reduce negative affective response (Dunton et al. 2015). Moreover, positive affective states can predict physical activity levels in children, and physical activity can in turn predict less negative affective states (Dunton et al. 2014). In a study exploring the relationship between cognitive, social, affective, and physiological factors to physical activity in older adults (50+ years) showed that self-reported measures of self-efficacy and control predicted higher physical activity. Moreover, positive social interactions also had an effect on physical activity levels of participants.

Context might include neighborhood characteristics (access to walking paths) or opportunities for utilitarian walks. Hekler et al. 2012 showed that within day variations in perceived access to environment that are supportive of physical activity showed a positive association to actual physical activity (Hekler et al. 2012). Moreover, increased walking is correlated with "access to walking paths, enjoyable scenery, good weather and nice locations".

Computational representations of Context

Context in literature has been viewed from a number of lenses. The phenomenological view of context defines context as interactional. Context shapes one's actions, is dynamically defined by actions, and is scoped by its relevance to the action concerned (Dourish 2004). The positivist

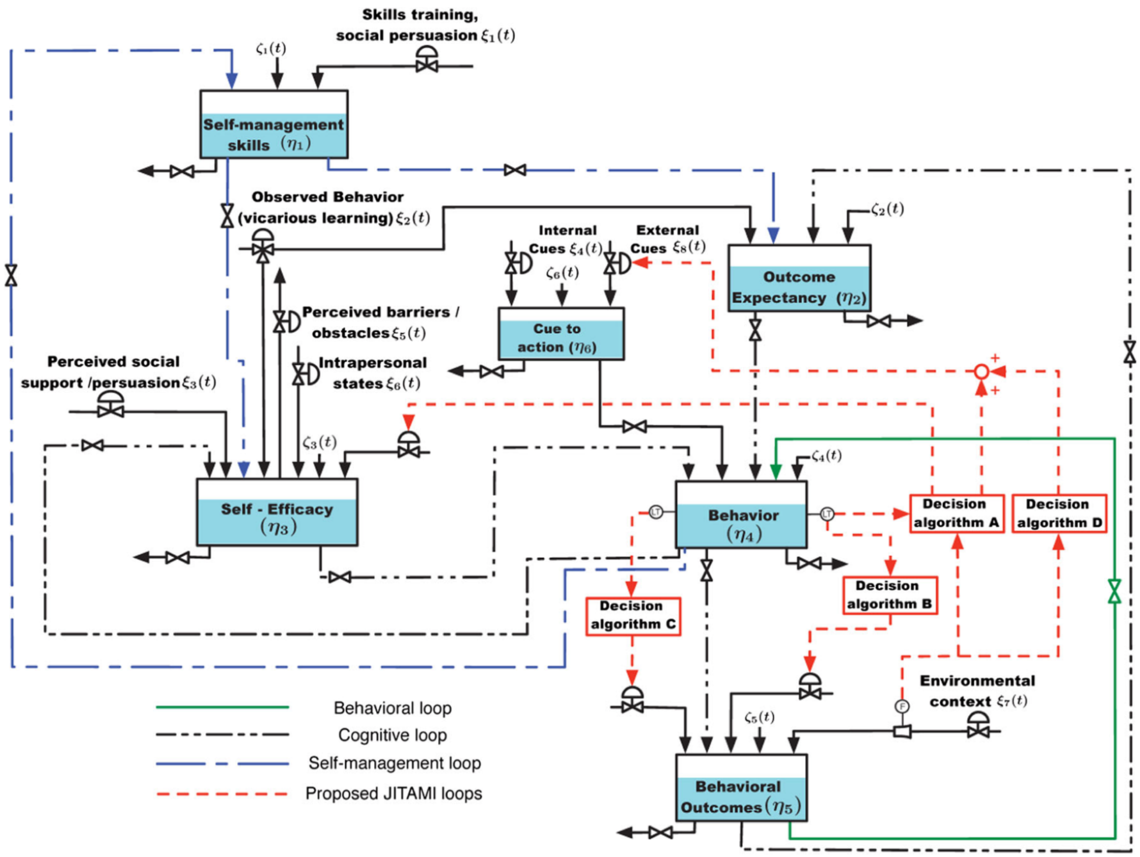

Figure 1. Fluid analogy of social cognitive theory represents the inputs to a model as values that fill a reservoir representing an aggregated factor (Martin et al., 2014).

perspective of context defines context in terms of concrete attributes to computationally represent and utilize context in applications (Bauer, Newman, and Kientz 2014). In order to design context-aware systems for behavior change, it is not only important to understand how significant the role of context is, but also represent the context such that it is amenable to computational reasoning. An example of a dynamical model utilizes a “fluid-analogy” to articulate an abstract model structure. In this analogy (Figure 1), the inputs to a model are represented as values that fill a reservoir representing an aggregated factor (Martin et al. 2014). The analogy is able to depict how the physical activity behavior is influenced by “inflows, outflows and feedback loops” between the various concepts (Spruijt-Metz et al. 2015). The model presented by Martin et al. 2015 is limited to understanding the relationships between constructs at daily timescales. To overcome this limitation, another model emphasizes how timescale impacts constructs (Spruijt-Metz et al. 2015). The authors note that model proposed is far from actually being applied to many constructs but is one way in which researchers can map out the dynamics of this “complex and nested” problem.

A transdisciplinary approach such as the one used by Hekler et al. 2016 can lead to the articulation of behavioral constructs in the language of control systems, however, such an expert driven approach may be too abstract and can easily be perceived to be distant from the lived reality of people. An opportunity, one that this dissertation also tries to explore, lies in bridging the gap between computational models and the lived experience of people’s lives by using qualitative methods. This approach can help ground the abstract notions of construct dynamics to recorded instances of how people experience challenges in their real life, thus providing the designer with a strong understanding of how the model might work in real world.

Prior work on computationally extracting causal relationships between people’s context and actions that describe their routines (Banovic et al. 2016) could provide a basis to model the effect of context on specific behaviors, such as making plans for physical activity. Such models improve the accuracy of systems to represent routines and behavior, which could potentially lead to applications that help people improve their routines (Davidoff et al. 2011; Davidoff, Zimmerman, and Dey 2010), including those that support physical activity.

On the more applied side, applications that have exploited the dynamic nature of context are few. Health mashups presented its users correlations between pairs of contexts, such as “On days when you sleep more, you get more exercise”. The inferred insights about participant’s behavior showed positive outcomes in physical activity of its users as well improved long-term engagement (Bentley et al. 2013). Representing the relationship between contextual factors is needed to facilitate self-awareness about how one's contextual factors come together to affect people’s physical activity. Prior work around this has looked at the use of contexts, such as events, places, and people, to support reflection on the factors that affect physical activity (Li, Dey, and Forlizzi 2012). (Li, Dey, and Forlizzi 2012) found that making associations between contextual factors and physical activity visible allow users to increase their awareness of the factors that influence their physical activity. Moreover, they showed that reflecting on physical activity and context allowed their participants to find more opportunities for physical activity. Being more aware of such opportunities associated with context can lead to better planning and hence, suggests the need to further understand the relationship between context and physical activity.

In exploring the space of generating recommendations for physical activity and eating behavior, MyBehavior system leveraged two contextual variables- activity and food logs to provide them with actionable suggestions (Rabbi et al. 2015). Specifically, they used "a standard machine-learning, decision-making algorithm, called multi-armed bandit (MAB), to generate personalized suggestions that ask users to either continue, avoid, or make small changes to existing behaviors". Although the study suffered from a small sample size (n=17), the participants who were shown the system generated suggestions walked more than the participants who were shown generic suggestions over the course of 3 weeks. Participants also rated the system generated suggestions higher than the generic suggestions.

In the related space of generating contextual nudges, Heartsteps tailored activity promotion messages to not only on personal characteristics but also to the context, i.e. to provide individuals with actionable support in the situations in which they find themselves (Smith et al. 2017). This personalization could potentially cover a large set of situations, for each individual. Table 1 shows how a subset of these contextual variables can lead to 36 different combinations. A system therefore needs a large bank of contextually relevant messages. Moreover, the messages must also be tailored towards stable individual characteristics (e.g. roles, goals, stage of behavior change) of every user.

Table 1. A sub-set of contextual features from Heartsteps (Smith et al. 2017) lead to a large number of situations for only one user.

Location	Time-of-Day	Weather (Intended Location)
Home	Morning	Outdoor
Office	Lunch	Indoor
Other	Afternoon	Outdoor Snowing
	Evening	
	After Dinner	

Total number of situations = 3 * 5 * 3 = 45

Scalable message generation

Mobile interventions for promoting physical activity have attracted significant attention from researchers. As it is known that well-designed, personalized messaging to remind and motivate people to integrate physical exercise into their daily activities can be effective (Kreuter et al. 2013). Recently, there has been a rapid increase in the use of behavior change applications such as Fitbit, Jawbone, and RunKeeper. In the future, we foresee this trend to continue in places all around the world, where people from diverse cultural backgrounds and native languages would like their mobile phone to support them in behavior change. With the increasing use of such applications, there is an even greater need for experts who can author these tailored messages. In this research, we ask the question: Can we support experts by scaling the process of creating tailored health communication messages?

Recent works in the domain of crowdsourcing show the promise of supporting complex activities similar to the one of our interest, such as annotating documents (Deng et al. 2014), proofreading texts (Bernstein et al. 2010), providing empathetic reappraisals (Morris and Picard 2012). Moreover, Hong at al. explored how individuals with autism can use crowdsourcing for advice in social situations (Hong et al. 2013). Another set of researchers used crowd workers to identify obstacles on sidewalks to help a person with physical disability navigate a city (Hara et al. 2015; Hara, Le, and Froehlich 2013). The techniques proposed have shown the viability of crowdsourcing as a scalable alternative to existing expert-based approaches.

Conclusion

To summarize, the literature lacks finer grained understanding of context and researchers have articulated the need for better theories and the potential solutions (e.g. multi-disciplinary collaboration) towards reaching this goal (Hekler, Michie, et al. 2016). There is a need to unpack the complexity in the interaction between contexts and individual differences. In the next chapter, I present a phenomenological account of how people make actionable plans within their temporal contexts. Namely, our qualitative findings support the notion of sweet spots, which may improve the predictability of computational models that support the creation of action plans for physical activity promotion.

To build such models and theories, researchers need the right tools to collect information about the users. Experience Sampling Method has generally been used with mobile phones to get a deeper understanding of participant behaviors. Although the use of smart phones has many advantages, existing works have noted its many limitations, such as the loss in participation due to heavy burden (Choe et al. 2014) and its limited coverage of situations, population, and research questions. In Chapter 4, I delve into this topic and the related works. I will present the single purpose devices we designed and studied with participants in their naturalistic setting.

Finally, a third research direction is revealed when we extrapolate the current progress of personalized mobile interventions to future applications that would need to personalize content in a scalable way. Given that health communication is known to be an important area, with much positive impact, a class of personalized applications can be hypothesized to persuade and nudge the user with personalized messages. In Chapter 5, I present a scalable way to generate personalized messages for physical activity promotion. The system uses expert generated micro-tasks to guide online crowd workers to write personalized messages for personas. I present the results of this study.

Chapter 3 Finding the Sweet Spot(s): Understanding Context to Support Physical Activity Plans

Creating actionable plans has been shown to be helpful in promoting physical activity. However, little research has been done on how best to support the creation and execution of plans. We conducted a study with 16 participants to study the role that context plays in the formulation and execution of plans for physical activity. Our findings highlight nuanced ways that contextual factors interact with each other and with individual differences to impact planning. We propose the notion of *sweet spots* to encapsulate how particular contextual factors converge to create optimal states for performing physical activities. The concept of sweet spots helped us to better understand the creation and execution of plans made by our participants. We present design guidelines to show how sweet spots can help support physical activity planning and guide the design of context-based tools for planning support.

Introduction

Researchers have highlighted the need for better models that unpack the complexity in the interaction between contexts and individual behavior (Hekler, Michie, et al. 2016). Improvements in modelling routines and human behavior (Banovic et al. 2016) can potentially lead to applications that support users in achieving their desired behavior. Design of tools that leverage such models, however, require a deeper understanding of the complexity involved in achieving a desired behavior such as physical activity. Towards this goal, this study contributes towards the design of computational systems that support people in translating their intention to behavior.

Social cognition models, such as the Theory of Planned Behavior (Ajzen 1991), have characterized motivational factors, such as people's intention to perform a particular health behavior, as the most proximal determinants of that behavior. However, it is a challenge for many people, even when motivated, to adopt and regularly perform a desired health behavior. To overcome this “intention-behavior gap,” specifying plans about how to enact one’s intentions and how to deal with difficulties in one’s goal pursuit is a promising strategy (Sheeran 2002). Specifically, for physical activity, planning is a key predictor of physical activity maintenance over time (Huberty et al. 2013; Mailey and McAuley 2014).

While tools have been extensively used to encourage physical activity using a variety of techniques (e.g., motivation, information, reflection, reminders, and social influence)(Dombrowski et al. 2012), there is limited support for specifying plans in such tools. That is, the design space for planning support as part of tools to promote physical activity remains largely unexplored and hence, creates an opportunity for researchers to contribute. Understanding how plans are made and when they fail or succeed could lead to the incorporation of improved planning support in tools to promote physical activity.

To better understand how people make plans for physical activity and how tools can be designed to support planning, this study aims to answer the following research questions:

- What factors do people consider while specifying plans for physical activity?
- What are the challenges of creating and executing plans for physical activity?

We interviewed 16 participants who demonstrated considerable motivation to perform physical activity but struggled to fit exercise into their lives. We found that multiple contextual factors such as weather, location, time, social interaction, and affect were considered by our participants in making physical activity plans. Moreover, these factors interact with each other and with the individual preferences of our participants to influence their plans for physical activity. Acknowledging the complexity that the consideration of multiple contextual factors and individual differences brings to planning, we present the notion of the *sweet spot*, a phenomenologically grounded construct to understand the role of context on plans.

While connections between context and activity have been noted in previous work (Liao, Shonkoff, and Dunton 2015; Humpel, Owen, and Leslie 2002), the notion of sweet spots helped to shed light on the nuanced interplay between these factors and the challenges they create for plan creation and execution. Specifically, we found that plan creation was challenging because of the need to consider multiple contextual factors, the constraints imposed by one's preferences and priorities that affected how participants weighed these factors, and the need to coordinate with others. Plan execution was challenging because of the difficulty in anticipating favorable and unfavorable contextual conditions, and in sustaining engagement in the face of fluctuating motivation. Additionally, we report that in dealing with these challenges, a few participants tried to strategize for improved plan execution.

We demonstrate how *sweet spot*, as a unified representation of converging factors, provides a novel and useful perspective to (a) bridge the gap between phenomenological and positivist perspectives of context to allow for computational support, and (b) inform the design of systems intended to support physical activity planning.

Related Work

In this section, we describe prior work relevant to this study. We first provide a background on existing planning strategies and interventions to support planning. We then provide an overview of research on understanding factors affecting health behaviors, which is pertinent to our research question about understanding factors that affect physical activity plans.

Types of Planning Strategies

Creating plans is found to be one of the most effective ways of reaching behavior goals (Latham and Locke 1991). A plan for a behavior consists of different elements, such as specifying a behavior ('what'), a time ('when'), a place ('where'), and elaborations on execution ('how'). Gollwitzer (Gollwitzer 1999) defined *implementation intentions* as explicit *if-then* plans that link anticipated critical situations (i.e., when, where) to goal-directed responses (i.e. what, how). This form of an explicit plans has been identified as an effective strategy for bridging the intention-behavior gap (Prestwich, Perugini, and Hurling 2010). Two distinct types of implementation intention strategies have been studied: action planning and coping planning (Carraro and Gaudreau 2013). *Action planning* involves specifying the action (when, where, and how) to act in the service of one's intentions. *Coping planning* involves specifying the anticipated response to potential barriers or obstacles that could get in the way of the intended action. Preliminary evidence has shown that high-quantity planning (creating more action plans) and high-quality plans (plans with higher specificity (Ziegelmann, Lippke, and Schwarzer 2006)) may lead to higher levels of physical activity (de Vet, Oenema, and Brug 2011).

Interventions to support planning

Although many interventions have been designed to incorporate implementation intentions as one of the intervention components to support physical activity (Hurling et al. 2007; Prestwich, Perugini, and Hurling 2010), little is known about lived experiences around how implementation intentions are formed and how technology could support people in setting implementation intentions by specifying high quality plans. Interventions using technology to support planning have mostly focused on motivating users to make more and higher-quality action plans by sending text messages (Mistry et al. 2015; Sweet et al. 2014). Although systematic reviews generally support such planning strategies (Gollwitzer and Sheeran 2006; Kwasnicka et al. 2013), their effectiveness to behavioral interventions was not evident in all settings (Paay et al. 2015; Vet et al. 2009).

Given the difficulty in creating high quality plans, many people resort to the Internet for help (Lewis et al. 2011), where the quality of plans may be contentious (Saperstein, Atkinson, and Gold 2007). Moreover, plans found online may fail to account for the individual preferences and opportunities (Paay et al. 2015), and are less likely to be tried than personalized feedback (Agapie et al. 2016; Rabbi et al. 2015). In order to seek personalized plans, people sometimes seek help from experts in the form of coaching. Coaching, in general, shows increased likelihood of achieving ones goals (Luszczynska 2006; Rosal et al. 2001), however the expense (>$18 per hour) (May 2016 National Occupational Employment and Wage Estimates n.d.) and time commitment required for personalized coaching are not feasible for everyone.

Therefore, we sought to understand situated planning for physical activity and how the design of tools to promote physical activity can incorporate support for making high quality plans.

Understanding the Context of Health Behavior

In order to design context-aware systems for behavior change, it is not only important to understand how significant the role of context is, but also represent the context such that it is amenable to computational reasoning. Inherent complexity in real world behavior change makes those theories unusable, that have “poor specification, both in construct definitions and in the relationships between them” (Hekler, Michie, et al. 2016). Hekler et al. posit the need for knowledge about "ongoing, dynamic feedback loops of behavior in response to ever-changing biological, social, personal, and environmental states" to articulate an abstract model structure.

An opportunity, one that this paper also tries to explore, lies in bridging the gap between computational models and the lived experience of people’s lives by using qualitative methods. This approach can help ground the abstract notions of construct dynamics to recorded instances of how people experience challenges in their real life, thus providing the designer with a strong understanding of how the model might work in real world.

On the more applied side, applications that have exploited the dynamic nature of context are few. Making associations between contextual factors and physical activity visible allow users to increase their awareness of the factors that influence their physical activity (Li, Dey, and Forlizzi 2012). The inferred insights about participant’s correlations between pairs of contexts showed positive outcomes in physical activity of its users as well improved long-term engagement (Bentley et al. 2013). Moreover, reflecting on physical activity and context leads to finding more opportunities for physical activity (Epstein et al. 2015). In exploring physical activity and eating behavior, MyBehavior system leveraged users’ context to provide them with actionable suggestions, which was found to be more effective than generic suggestions (Rabbi et al. 2015).

This body of prior work suggests that making people aware of the context in which health practices take place could improve those practices. Leveraging context information for planning tools is an area that is relatively unexplored. While automated planning and scheduling approaches have been proposed (Wikipedia 2017b), there is an opportunity for a user-centered approach towards designing such a system.

In this work, we delve into 1) what context information is considered relevant while making plans for physical activity, 2) how contextual factors challenge or support planning, and 3) how contextual factors can undermine existing plans. Given the importance of planning in predicting adherence to physical activity, understanding the role of context can inform how context information could be incorporated in system design to aid planning.

Methods

Our goal was to understand the factors that people consider while planning physical activity and the challenges they face in creating and executing their plans, that is, how people plan for performing physical activity and how those plans materialize or fail to materialize. We do not seek to make generalizations about how particular factors impact activity, but to qualitatively understand how people take context into account, alongside other factors, when making and executing plans.

We specifically studied people who were motivated to exercise more than they currently do and who had a (self-described) busy work schedule. In doing so, we were guided by prior work that has found lack of time to be a barrier for physical (Choi et al. 2017; Trost et al. 2002). We expected that a person with a somewhat full schedule would have more constraints around planning physical activity as compared to a person with a flexible routine with little time pressure. Because we wanted to maintain consistency around the nature of planning issues faced by our participants, we chose to study people who had similar work schedules. Our participants' work hours were approximately 9-5p, Monday to Friday.

Participant and Recruitment

We interviewed 16 adults who self-reported that their satisfaction with their physical activity is low (they are motivated to exercise more) and who worked at least 40 hours per week. The interviews were conducted between December 2015 and April 2016. We recruited participants through mailing lists at the authors' university. The list consisted of the departments' students, alumni, staff, and faculty. Respondents were asked to fill out a screener with 13 short-answer questions asking about their daily routine, typical physical activities and demographic information. Potential participants indicated their score of overall satisfaction with their current level of physical activity on a Likert scale from 1-5. The participants with satisfaction score less than 3, were chosen after the screening for in-person or online audio interviews. Interviews lasted 50 minutes on an average. The participants were located in Michigan, California, Maryland and Illinois. Out of the 16 participants, 3 were male, and 13 were female. Eleven participants were aged 25-35 while 5 were 35-45.

We analyzed data as we collected, to improve our understanding and also improve the probes we use in future interviews. We noticed data saturation occurring with our last few participants, realize that interview responses became repetitive, so we stopped enrolling more participants (although we had people wanting to participate).

Data Collection

The first author conducted all the interviews using a semi-structured interview protocol. At the beginning of the interview, each participant reported their activity levels on a standard Activity Scale (NASA JSC Physical Activity Status Scale). The average physical activity level for the participants was found to be 3.8 on the scale of 7. During each interview, participants were asked about their daily routines, whether and how they made plans for exercise, the challenges they faced, and how they dealt with those challenges. We further probed participants on some important incidents to explore participants' experience of successful and unsuccessful plan execution in depth. Many of our participants identified such incidents of plan creation and execution. Participants also described the context around which their physical activity happened. Although our interest in this study was to understand what contextual factors play a role in making and executing plans for exercising, it is important to note that contextual factors are difficult to be asked about individually. For instance, when directly asked about weather a participant may mention their overall preference but would find it hard to articulate the many different ways in which weather interacts with another contextual factor like location. Hence, we sought to understand the role of context by asking about participant's daily routines, incidents around physical activity, and the relation of these incidents with the plans that participants made. Contextual factors were often discussed organically by participants, and were further elicited through probes.

Data Analysis

All interviews were audio-recorded and transcribed. The interviews were coded and analyzed using a mix of structural coding and in vivo coding (Saldaña 2015) consisting of an iterative process of generating, refining, and probing the themes that emerged. Codes were initially drawn from research questions and then supplemented with those that emerged from the interviews. The research team performed data analysis, starting with a few interviews to reach an agreement on codes, and understanding of concepts. The coding process was further supplemented with group discussion with the other members of the research team as a form of peer debriefing (Creswell 2013) to develop the understanding of emerging themes.

In the another round of analysis, coded data were grouped under themes using affinity diagrams (Wikipedia 2017a). Coded data was analyzed to compare and understand the differences between participants and how they make and follow through on their plans with regards to the types of challenges faced and the types of coping strategies.

Findings

Our findings shed light on the challenges of plan creation and execution arising out of multiple contextual factors that affect physical activity plans. Contextual factors including the time, location, type of activity, duration of activity and presence of others during activity influenced choices relating to physical activity such as when, where and how. Additionally, we found that participants weighed these contextual factors differently depending on individual differences, such as priorities and preferences, as well as affect and motivation. While the role of context in physical activity promotion has been reported in prior work (Bentley et al. 2013; Li, Dey, and Forlizzi 2012; Liao, Shonkoff, and Dunton 2015), our work adds to the literature by showing how multiple aspects of context interact with each other to influence physical activity planning and execution. Moreover, we show how individual differences affect the role of specific contextual factors in planning physical activities.

Operationalizing contextual factors and personalizing content and services to the user are two main aspects of a context-aware system. A deeper understanding of context is especially important to understand when seeking to inform the design of context-based systems that can support planning, which is our ultimate goal.

Successful planning requires finding “sweet spots”

Table 2. P14's sweet spot was found to be a convergence of multiple contextual factors. P14 considered these factors in creating her physical activity plans.

Contextual Factor	Values
Activity	Exercise at gym
Location	Nearby gym
Time	Morning
Social preference	Likes the social environment of gym
Social preference X Time	Wants evenings free for socializing
Affect after exercise	Feels very satisfied
Affect on missing exercise	Feels regret

Our findings show that all our participants engaged in planning for physical activity by considering certain contextual factors. Although the relevance of contextual conditions varied across participants, they had to co-occur for participants to find an opportune circumstance that supported a given physical activity. That is, a right combination of time, location, activity, affect and social preferences was needed to perform a physical activity successfully. A state when these conditions were satisfied simultaneously was described by P3 as a "sweet spot":

"But the problem is, again with my schedule, really that 5:30 to 6:30. If it ends before 7:00, that is my sweet spot. But so many of the classes started at 7:00, and I just can't do a weekly commitment. I can't take two days a week, two evenings a week from the other work I have to do."

P3's work schedule and the schedule of martial arts classes determined her sweet spot for exercise. Such reports from our participants demonstrate that different factors are considered in planning for exercise. They also show that these factors converge to create conditions that support or deter the execution of those plans. Inspired by this understanding, we propose a notion of *sweet spots* for physical activity, that is *a state that supports a desired activity, formed by a favorable convergence of perceived contextual factors, and sustained for a sufficient period of time to successfully perform the activity.* The construct of sweet spots helped us better understand how contextual factors interacted to influence plans. While every participant had their own way of formulating plans, the concept of sweet spots unifies the diverse approaches adopted by our participants: planning can be seen as a matter of making choices around factors that form one's sweet spots.

In what follows, we describe our findings with the help of the notion of sweet spots. Our participants considered multiple factors to create plans based on their own preferences and priorities. It should be noted that much of the consideration and weighing of factors was not explicit—that is, these were not the terms or framing that participants used, but rather the analytical categories imposed by us to understand the role of context in planning. Additionally, we describe our findings under the two main themes – challenges of plan creation and challenges of plan execution. Plan creation and execution are complex and often intertwined in the lives of people (Epstein et al. 2015). We make a distinction here with the goal of unpacking the complexity and thereby extract the distinct challenges associated with them. Once articulated, the challenges of planning can be well scoped for the designers of context-aware system to tackle. Our primary goal is to unfold the complex role of context (e.g., time, location, social factors, affective state, and activity) and individual differences in planning physical activity.

Challenges of Plan Creation

As noted above, planning is a critical component for health behavior change. All our participants described some degree of conscious, explicit planning to try to incorporate physical activity into their routines. Although many health interventions tend to treat creation of plans as straightforward, we observed three ways that participants faced challenges in plan creation: considering multiple contextual factors, being influenced by the interaction between their individual preferences and relevant contextual factors, and coordinating with others.

Plan creation requires considering multiple contextual factors

Planning generally included choosing an activity, deciding on a time, and identifying a suitable location and environment. Other factors were sometimes included, such as coordinating with others or securing resources (e.g., registering for a class). Our participants considered many such contextual factors simultaneously. Participants had to make decisions about the type of activity they would perform, the place and its facilities, the duration of activity, their commute to the location, and constraints associated with certain aspects of their routine, such as work engagements and family obligations. For example, the kind of plan participants could create was restricted by the facilities accessible at a location (e.g., gym, shower) as mentioned by P3, *"when you work out at work on our lunch hour, or ride your bike in, we don't have a shower there. The* *smell is not terrible, [chuckle] but I would rather be clean when I go to work."* She did not have a shower at her workplace so she could not bike to work or exercise during office hours. Thus, any plans that she made to exercise were usually after work hours.

Some of these factors challenged plan creation more than the others. For example, work and family responsibilities limited the scope for plan creation. Lack of time, family obligations, and social engagements are known barriers for physical activity (CDC, 2015). While the participants could better control a few factors, other factors, such as competing priorities at work and home, were more difficult to navigate.

Individual differences impose constraints on plan creation

Although the set of contextual factors that were considered for plan creation were mostly consistent across participants, individual differences governed how each participant preferred to weigh these factors, which in turn affected their plans. One individual difference was around time preference. For example, different participants felt a sense of accomplishment by exercising at different times in their daily schedules. Some participants liked to exercise before work, while others preferred exercising after work. Some participants preferred to exercise at the same time across all days of the week, whereas some chose variable times for different days of the week. Depending on the time of the day, the reason for feeling accomplished varied across participants. For example, P14, a single female, professional, who was socially outgoing**,** thought that morning was better because it gave her free time during the rest of the day to do other things, such as socializing: *"If I've gone in the morning, I'm really satisfied because I did my workout, and now I'm going to have the rest of the evenings to do whatever I want. My first preference is always going in the morning because something can come up. I could have a team dinner or a team outing."* On the other hand, P1, a single female graduate student who spent most of her time alone, preferred evenings because her sense of accomplishment was driven by the feeling of being done for the day. She went for a walk after finishing the day's work and before sunset. At this time, there was still natural light in the park and more importantly she felt satisfied knowing that she can go home, take a shower, eat dinner and relax after completing her walk. As seen in the above two examples, *time of the day*, one's *work schedule* and *individual differences* over what led to satisfaction, determined the plans for physical activity. The favorable conditions of contextual factors such as time of day, work, daylight, and social factors resulted in sweet spots for P14 and P1. P14's sweet spot was bound by multiple contextual factors as shown in Table 1. While creating physical activity plans, P14 considered these factors and how they interacted, which was particular to social preferences of P14. Anticipating contextual conditions for successful plan execution may thus be seen as anticipating these sweet spots (that is, making plans that lead to sweet spots) and then choosing one (or more) of them for executing physical activity.

Another individual difference observed was around the desire to be with other people. Consistent with the above example, we found that this factor affected participants differently. As noted by prior work (Berkman 1995), for some participants in our study having people to exercise with was a necessary motivation. That is, without others they were less likely to create plans for physical activity and, if they did plan, their plans for exercise were less likely to materialize. For example, P4, who had been motivated to lose weight for many years, only decided to exercise when her friend at her workplace invited her to join a weight loss program*: "It was not until my colleague asked me to lose weight with her together, at that point I really started to do the exercises… After she left the office. I do not know a reason I stopped going to the gym. I think she played a very important role in my exercise program."* P4 was aware of her preference and acknowledged that it inhibited plan creation, as it was difficult to find people with whom she could plan.

In contrast to P4, the presence of too many people in the gym was a deterrent for P10 and she preferred to plan exercise around times when the gym would be relatively less crowded. Time in the gym was her "me time," which she wanted to spend with as few people around her as possible. Consequently, she made observations about the crowd patterns in the gym to conclude that her gym got very crowded after Christmas holidays. Using this information, P10 avoided the times when the gym was crowded and planned exercise sessions so that she could exercise at home on those days when the gym was crowded. P4's anticipation further required awareness about the various contextual factors that make up her sweet spot and how variations in each factor affected her plans.

Competing priorities in the context of work and home obligations was another factor that interacted with individual differences. Although competing priorities inhibited plan creation for all participants, we found that the relevance of barriers associated with competing priorities was different for different participants. For example, participants living with spouses and children had priorities dominated by family obligations, whereas participants living independently reported the desire for socialization (e.g., going out with friends) as a competing priority. This eventually determined the plans they would make. For instance, P14, whose priority was socialization, preferred not to create evening plans for exercise as that was a time when she socialized with friends and colleagues. The same contextual factors thus affected each participant differently depending on the individual differences in their preference for people, priorities, and sense of achievement and satisfaction.

Plan creation requires coordination with others

Participants living with their family needed to coordinate with others and plan around others' activities while making their own plans for exercise. This was because their schedules were directly influenced by other people's schedules. For example, P9 looked at her husband's schedule to plan her exercise. The couple could not go to the gym together because they shared childcare duties. In this case, presence of family members creates a need for coordination. In response to how she planned her exercise, she reported: *"I pretty much have it planned. The days that I work, I have it planned that I get home, rest for half hour or something, change, and go to the gym. I do it that way because my husband also likes to go to the gym, so I try to go so when I come back and he can go before dinner time."*

The effects of social aspects on plan creation varied from participant to participant. Most participants with families described schedules that were interdependent with those of other family members, thus having continuous social impact, as evident from P9's case. But for P14, who lived alone, vulnerability arising out of interdependent schedules was rare. Only when her friend visited her did she have to give up on her exercise plan as she wanted to be with her friend.

Challenges of Plan Execution

Even when the challenges of plan creation could be overcome, most of our participants reported having difficulty in executing physical activity plans once they were made. Unexpected and unplanned events can disrupt the execution of planned physical activity during the sweet spots, resulting in plan failures. That is, the sweet spots that could have been utilized for the desired behavior proved to be infeasible because of unfavorable conditions. Whether their plans were executed as expected was dependent on: a) the near-term anticipation of favorable and unfavorable contextual conditions that would support and deter their plan execution respectively (e.g., unexpected events), and b) fluctuating motivation levels. In the rest of this paper, we refer to the failure to execute plans as plan failures.

Anticipating unfavorable contextual conditions

When anticipating the conditions for physical activity, it was difficult for participants to foresee contextual conditions that would ruin their plans, e.g., unexpected and unplanned events. Most of the participants reported instances when a planned physical activity failed to happen because of an unforeseen event. For example, one of the participants described these events as out-of-the-ordinary things that came up and threw her off routine, affecting her physical activity plans: "*sometimes there are things, you know, like maybe somebody is sick or something like that too. I mean that would be something that would throw off my routine and something out of the ordinary."* Feeling sick was one the most common reasons participants mentioned by participants, as also mentioned by P12: "*Then the last few days I was sick, so I didn't get to do my weights. I did running except for yesterday because I wasn't feeling good. Obviously, I was not happy with those.*" Other than being sick, such unexpected events also involved impromptu activities needing attention. Most commonly these impromptu activities included errands, as reported by P15: "*I planned to run on Monday evening, after work, but then I ended up having to meet my husband somewhere to do some errands. And then we didn't get home until it was too late. Because I had planned to do that [run] from 6:00 to 7:00 but then we didn't get home until 7:30 or maybe 8:00."*

Another participant (P4) blamed impromptu socialization for disrupting her plans for exercise. Her friends would ask her for a meal together in the evening and her plan to exercise got deprioritized because of her time with friends. She mentioned: "*a friend would say 'hey do you want to have dinner together? And then if I need to hangout first, then I will socialize first and* *then think about exercise… I think it is something I need to balance my daily life on.*" It is also important to note that some of these unfavorable conditions have less probability of occurrence, such as a sick kid, and some are more predictable for some people, such as an outing with friends or a long workday.

Sustaining engagement despite fluctuating motivation levels

Some participants reported that lack of novelty can lead to boredom, which demotivated them to follow through on their plans. For example, P10 found that routine can engender boredom, which eventually would lead her to fall off the routine, "*The thing that I have a problem with is I get bored easily so I know some people that like go to the gym and run on the treadmill like I hate and I don't think I can do that. So, I have to like mix it up and mix up where I'm at... I try not to get bored because if I get bored then I don't stick with it or if it becomes too routine, I get bored with it."*

Lack of novelty discouraged sustained engagement in physical activity to the extent that it also led a few participants to completely discontinue planning physical activities. They stopped planning for and performing any physical activity. For example, P7 said she was bored of doing the same set of things as physical activity: "*I mean I would love to get back into a routine. I've tried workout videos at home as I do it for about a week and that's about it… So I go on spurts. I do it then I stop. I have a Wii. Done that. Then, get bored of that. I have a treadmill at home. Just never find time to use that either.*" As mentioned in the interview quote, for each new mode of exercise the motivation depleted over time, and required planning new activities for continued engagement.

Dealing with Challenges

In dealing with the challenges of plan execution, a few of our participants strategized for improved plan execution by creating backup plans or by mixing up activities.

Strategizing the execution of plans

While most of the participants acknowledged the presence of vulnerabilities because of poor anticipation of contextual factors, we found that only some of the participants had developed strategies to plan for physical activity and work around plan vulnerabilities. Such strategies resulted in more robust plans that led to improved physical activity execution. For example, one of the participants (P15) had created multiple exercise plans and she chose to execute one of them depending on the contextual conditions for any given day. This decision was governed by when she got back from work and what available time she had for exercise: "*I've got couple different routes [to run]. If I don't have a lot of time, I'll go on my shortest route, which is maybe 20, 30 minutes. Then if I have a longer time, I've got a route that takes about an hour. And then if I don't have an hour but I have more than just the shortest route, I've got sort of a 45 minutes one that I can do. And so I usually just rotate between those options.*"

The above-mentioned strategy followed by P15 made her plans less vulnerable to being disrupted. That is, she could deal with a situation in the week when she would end up getting home late from work, giving her less time for exercise. Because she could adjust to a plan that required less or more time to exercise, she did some exercise instead of missing it altogether. For P15, forming stable exercise routines was a result of being able to anticipate potential plan disruptions. Another planning strategy was to mix up different types of activities. For example, P8 was doing Yoga classes, but also did strength training with arm-weights because of the ease with which she could plan this activity Having multiple desired activities increased the scope for execution as the participants could execute one of the activities depending on the contextual conditions.

Discussion

This study explored how people seek to make and carry out plans for physical activity. Findings from this study bring to light two aspects of physical activity plan creation and execution—plans encompass multiple contextual factors (weather, social factors, affect, time, other activities) that interact with individual differences (preferences, priorities), and planning requires acknowledging and understanding the transient nature of contextual conditions. These two characteristics are indicative of the complex ways in which context affects the creation and execution of physical activity plans of our participants. Considering this complexity, we identified the concept of sweet spots, a favorable convergence of multiple contextual factors, to better understand the influence of context on physical activity plans. Given the central role of context in affecting plans, in this section, we briefly discuss the implications of our findings and the concept of sweet spots when viewed through the lens of the dichotomous view of context that exists in current literature on context-awareness, that is the phenomenological and the positivist views of context. We argue that the model of sweet spots is an initial step towards bridging the gap between these two views to aid the development of applications that use context to support a situated behavior. We further demonstrate how the results of this study inform the design of context-based tools to support planning for physical activity.

While the phenomenological view is less clear in terms of operationalization, the positivist view oversimplifies the complexity and nuances of context. Both these conceptualizations have their own merits and limitations and we do not aim to argue for the importance of one over the other. While the socio-technical gap cannot be completely bridged (Ackerman 2000), designing context-aware applications requires one to apply phenomenological understanding to system design. That is, understanding how context can be best incorporated in system design to support a situated behavior requires understanding the nuanced role context plays in relation to the behavior being studied and requires a construct to computationally utilize context in designing tools to support the behavior. In this study, we provide a deeper understanding of the role context plays in creation and execution of physical activity plans, and offer a phenomenologically grounded construct (sweet spots) to understand and computationally model the effect of context on plans. Sweet spots as a computational framework is rooted in phenomenological understanding, and hence, bridges the two approaches to understand context and its role in affecting a behavior.

Although intuitively the concept of sweet spots may seem straightforward, the merit of this construct becomes clearer once we consider the complexity that the phenomenological view of context brings to the design space of context-aware systems. For planning physical activities, findings from our empirical investigation unpack this complexity by a) highlighting that it is not just isolated contextual factors that matter but that their combination must be considered as well, b) anticipating changes across multiple contextual factors and their effect on actions is challenging, and c) individual differences affect what specific factors or combination of factors matter for whom and how, indicating that there is no "one size fits all" solution. This points to the diversity in the design space of context-based applications to support a behavior such as exercising. Although context is unique for each user, to design applications using context we need a framework that offers some degree of generalization while being flexible to the unique needs of users. Sweet spots, as described above, is one way to achieve this goal.

Our findings show the diverse interactions between context and individual differences, implying that a universal mapping between context and planning is not entirely possible. What might work for one user would not work for another. This is where a phenomenologically grounded notion of sweet spots becomes more useful: as an abstraction that captures the importance of context in planning without pre-defining rigid and specific relationships between context, and plan creation and execution.

Predicting sweet spots

Sweet spots provide one way to understand the effect of context on plans. We hypothesize that the notion of sweet spots can potentially have greater explanatory and predictive power for the creation of actionable plans. Computational models of contextual factors and desired behaviors, that leverage the notion of sweet spots could then be used to predict likely sweet spots and suggest opportunities for being active. When plans fail, a predictive model could help repair the plan by suggesting alternative sweet spots. Automated prediction and reasoning about human behavior is a challenging research problem given the inherent uncertainty in human behavior (Banovic et al. 2016). However, recent work on extracting routines from behavior logs holds promise. The work by Banovic et al. (Banovic et al. 2016) on routine modeling can be leveraged to define sweet spots in terms of a computational model. Given the scope of this paper, our intention is to convey the intuition of a plausible approach rather than providing the details of the actual implementation. In the following paragraphs, we first briefly outline the approach used by Banovic et al. We then describe how their approach may be extended to include the notion of sweet spots. Finally, we provide an example to help understand this computational model as a basis for predicting sweet spots and for predicting the plans that would result with those sweet spots.

Banovic et al. present an approach to model the causal relationship between contexts and the actions that people perform in those contexts. Their model consists of contextually-defined states mapped to actions that a person can take in that state. They calculate the probability distribution of actions, given the states and the state transition probability distribution from behavior logs. This allows their system to extract routines, as well as variations in those routines. They model routine behavior using Markov Decision Processes (MDP) framework and leverage MaxCausalEnt, an algorithm based on inverse reinforcement learning approaches to calculate the probability distribution of actions given states (Ng and Russell 2000).

Translating their model to our intended use, a state represents combination of contextual factors such as recent activity, location, time and weather. Actions represent the activity (behavior) that can be performed in the given context. Based on the knowledge of states and the state transition probabilities, it is possible to extrapolate future states from any given state. Considering the diversity in context and behavior data, it is possible for a model like this to end up with large number of states. However, highly predictable events from a user's day (sleeping, office meetings) could effectively prune the state space, reducing the number of likely subsequent states. In this model, each state includes an activity feature that tells us the predicted behavior (physical activity in this case). The likelihood of a desired behavior can then be calculated by aggregating the likelihood of all states, at a given time, where the desired behavior/activity exists. Situated in such a model, a sweet spot for a person can be defined as *a state when the*

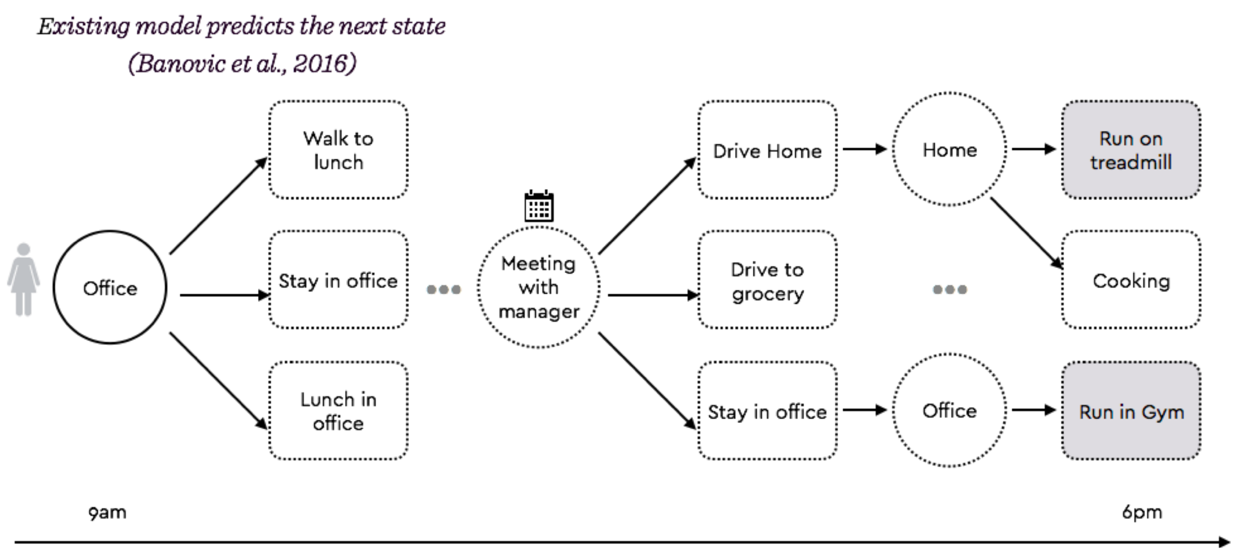

Figure 1. An example of the state-action space for a person, simplified to convey two points: A) A high probability state restricts the state space of future states. B) Exercise being the desired behavior is identified as the sweet spot. Although a state captures multiple contexts, state in this scenario is only represented by the location or the activity of the state for simplicity.

probability of a desired behavior happening is high, considering constraints of multiple factors including location, activity, and user's preferences (e.g. desired behavior is in the top 3 most probable behaviors for a given set of factors.)

We now present an example of a routine day for a hypothetical user. Assuming the user's current state is (Activity=Work, Location=Office, Weather=Nice), the future state space (a collection of possible states) maybe predicted as shown in Figure 1. The state space is only a simplified version of the state space that might exist in reality. This state space is constricted by highly likely states, such as a recurring meeting with one's manager. We consider two predicted states with the desirable behavior "Run outside" and "Run in gym" to exist at around 6pm. The likely time for desired behavior for the user would then be predicted by the model as consisting of these two desirable states (run outside, and run in gym), inclusive of three factors (Time=6pm, Location=Home, Activity=Run Outside and Time=6pm, Location=Office Gym, Activity=Run in Gym), and the paths that lead to those states. The predicted states are sweet spots, and the paths are plan simulations, which when followed could result in successfully executed sweet spots.

Design Implications

Our elaboration of sweet spots as a construct to model context data makes it amenable to computational support. While prior computational work on context data to support health-related needs has suggested ways to discover correlations between pairs of contextual factors (Bentley et al. 2013) or suggest activities (Rabbi et al. 2015), sweet spots offer a way to improve the explanatory and predictive power of existing models by holistically accounting for multiple contextual factors affecting a behavior/activity to a) present context as information to the user to support awareness around a behavior, and b) use context to trigger services/suggestions in support of a behavior. Considering these broad use cases for incorporating context in designing tools, we now provide design directions for tools to support planning as implied by the findings from our study, using the construct of sweet spots as a basis.

Support Creation Of Plans By Learning User's Behavior

Anticipating favorable and unfavorable contextual conditions is a challenging aspect of planning. For our participants, considering multiple factors in their plans introduced challenges for plan creation and execution. A system supporting plan creation and execution could suggest potential opportunities for exercise and reduce the need for explicit planning. Although existing systems do this by considering the user's calendar and prior behavior (Google 2016), we propose that systems need to consider multiple contextual factors simultaneously while suggesting potential opportunities to the users.

In terms of sweet spots, based on the factors and preferences relevant to the user, the system could predict and suggest sweet spots– the states when the user is likely to execute physical activity plans successfully. A system could, for instance, provide a look ahead into the user's day or week suggesting all the slots on the user's calendar when a specific physical activity is feasible. Alternatively, the system could recommend multiple sweet spots having considerable probability to occur and let the user choose a primary sweet spot and one or more backup sweet spots in case the primary sweet spot becomes infeasible.

Suggest paths of successful plan completion

The dynamic nature of contextual conditions made our participants' plans fragile. At the same time, sweet spots become *missed opportunities* for exercise when the conditions are favorable but the user did not realize the opportunity to do the desired behavior. Therefore, there is a need and an opportunity to support the creation and modification of plans to fit the changing contextual conditions to – a) deal with potential failures arising out of unfavorable conditions and b) take advantage of emergent sweet spots that represent unplanned opportunities for exercise.

Owing to the fragile nature of plans, it is important to have strategies that accommodate the changing nature of contextual conditions. We observed that only three participants employed such planning strategies to ameliorate plan failures by creating flexible or backup plans. We believe that such strategies would benefit other users as well.

Sweet spots, as a phenomenological account can support the use of existing computational models for suggesting multiple paths (series of state transitions) in the predicted state space that lead to a sweet spot/desirable behavior. The presence of multiple paths can help the user choose paths to achieve the desired behavior. These paths are useful in themselves as they represent the changes in context and the potential actions that could be taken by the user in order to reach a certain behavior. They can thus be used to guide plans for that behavior as they are essentially detailed simulation of plans for a behavior/activity. Paths that reach the desired behavior are “successful” transitions while the paths that do not reach the desired behavior are "unsuccessful” paths. Visualizing and recommending successful and unsuccessful paths can make the user aware of unforeseen and disruptive contexts that may occur.

Help the user reflect when failures happen

Not understanding how different factors affect one’s plans challenged plan creation and execution for our participants. Planning tools could represent the relationship between contextual factors and outcomes to facilitate self-awareness about how contextual factors come together to affect plan creation and execution. The framework of sweet spots, encapsulating the knowledge of desirable and undesirable states and paths that lead to those states, can aid designers in highlighting patterns of plan failures - i.e. highlighting the paths that have less probability of reaching desired behavior. Consequently, causal contextual factors, i.e. the states/events along the path that lead to deflection from the desired behavior, can also be highlighted to promote reflection over one's plan failures. For example, a system could nudge users by highlighting the differences between conditions when they fail and succeed. Such information can be presented as natural language statements that associate context data with a particular behavior/activity (Bentley et al. 2013). This may help the user draw actionable insights to avoid failures. In doing so designers should be aware of the discouragement and any undesirable emotional response that assessing one’s own failures could engender.

Help the user explore new ways of achieving one’s goal

As found in this study, lack of novelty in some cases discouraged sustained engagement in making plans for physical activity. To deal with this issue, tools to support planning should suggest new options from time to time, such as a new physical activity or a new location. Such suggestions can be triggered using context-based cues, that is, changes in context that lead to the emergence of a sweet spot. For example, repeated plan failure for a given activity that the user has been trying to perform can trigger suggestions for new activities that result in new sweet spots for the user. Alternatively, changes in seasons can be used as cues to trigger suggestions for activities more appropriate for the season. The new options may also be learned from trajectories of other users who are similar to the user. Since incorporating new activities into a daily routine requires planning afresh for those activities, using a sweet spot representation could be used to help the user become aware of the relevant contextual factors that the system considered while suggesting new options or new sweet spots, which would also need to be considered while making plans for the suggested activity.

Limitations

Our study has limitations that may be addressed in future work. We used interviews to understand how participants made plans and executed them. Although this allowed us to gain deep insights into participants' daily routines and the nuances of how they plan, future work could incorporate methods like Ecological Momentary Assessments (Dunton et al. 2009) to gain deeper insight into creation and execution of plans in their actual context. Using interviews, our goal was to provide a rich contextualized understanding of how people create and execute plans, through the intensive study of particular cases. The generalization is therefore in relation to knowledge acquired through the analysis of the interviews.

Furthermore, our methodology and recruitment strategy imposed restrictions on the type of participants we interviewed. All our participants belonged to relatively high socio-economic status. Most of our participants were women and fairly young and healthy. Additionally, we scoped our data collection to explicitly planned activities and thus, our findings may not generalize to spontaneous physical activities. Although these factors might have restricted our understanding about the role of context and the challenges with planning to a sub-set of all possible roles, we believe that the notion of sweet spots provides a useful intuition, backed by empirical evidence, for designing tools for planning support for a wider audience. Future studies aimed at understanding planning behavior can utilize the concept of sweet spots to design and evaluate planning support tools.

The notion of sweet spots emerge from findings from interviews where we were especially interested in within-day timescales. Based on the evidence we presented, we cannot claim that sweet spots would be generalizable to other time scales such as weekly, or monthly. Intuitively, the computational model of sweet spots model determines states where the probability of achieving a desired physical activity is high. Such states are not restricted by time duration and may occur over days or weeks, depending upon context features that vary over days. An example where sweet spot for a person might occur at a weekly timescale is when a person is work schedules vary over days, for e.g. a user may have a week where she is travelling most of days but have one day that is optimum considering her workload, weather, and social factors. Instead of suggesting an hour, the system may nudge the user to make plans for that day.

Conclusion

In this chapter, we reported results of an interview study with 16 participants to understand the impact of contextual factors on physical activity plans. We found that our participants faced certain challenges in creating and executing physical activity plans. This was because their plans encompassed multiple contextual factors, and planning required acknowledging and understanding the transient nature of contextual conditions. The primary findings of our study lead us to conclude that context and individual differences play a complex role in affecting people's physical activity plans. To ameliorate such complexity, we use the phenomenologically grounded notion of sweet spots- states that represent a favorable convergence of contextual factors to support a desired behavior. Sweet spots can be utilized to improve the predictability of context-based tools or models that support creation and execution of plans for physical activities. This work suggests new opportunities for research in context-aware systems for physical activity promotion and for exploring the role of contextual information in creating and executing physical activity plans.

Chapter 4 Heed: Exploring the Design of Situated Self-Reporting Devices

In Chapter 1, we introduced the *context management* step of a context-aware system and in Chapter 2, we argued the need for better computational models for human behavior. In order to develop theoretical knowledge about human behavior, researchers have used a wide variety of methods to gather and analyze data from their participants. In-situ self-reporting is one such widely used data collection technique for the assessment of the behavior and the context of its users. It involves soliciting information from the user at a relatively high frequency. Characteristics of smartphones such as their high proliferation, close proximity to their users, and heavy use make them an ideal choice for such tasks. Newer device categories such as wearables, voice assistants, augmented reality and the Internet of things (IOT) offer their own advantages, providing an opportunity to explore a wider range of self-reporting approaches. In this paper, we focus on exploring the design space of Situated Self-Reporting (SSR) devices. We present the Heed system, consisting of simple, low-cost, and low-power SSR devices that are distributed in

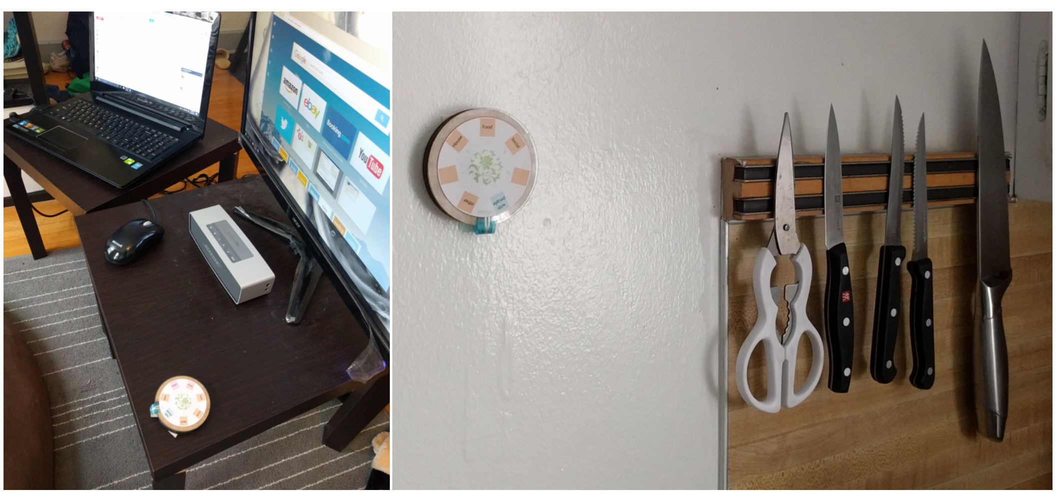

Figure 2. Heed self-reporting devices placed in participant's environment. (left) On a desk where the participant usually works, (right) Stuck on the wall in a frequently accessed area in a participant's kitchen.

the environment of the user and can be appropriated for reporting measures such as stress, sleepiness, and activities. In two real-world studies with 10 and 7 users, we compared and analyzed the use of smartphone and Heed devices to uncover themes of use, such as situational and social context, notification types, and physical design. Our findings show that Heed devices complemented smartphones in the coverage of activities, locations and interaction preferences. While the advantage of Heed was its single-purpose and dedicated location, smartphones provided mobility and flexibility of use.

Introduction

To develop models and theories about a behavior, researchers first need the right tools to collect reliable information about the users, following which they analyze and discover patterns of use. Self-reporting tools such as mobile devices are common in the assessment of behavior and context in various disciplines, and is referred to by variety of names: the experience sampling method (ESM) (Conner et al. 2009; Hektner, Schmidt, and Csikszentmihalyi 2007), diaries (Palen and Salzman 2002; Pielot et al. 2015), ecological momentary assessment (EMA) (Shiffman, Stone, and Hufford 2008) and ground-truth labeling (Chang, Paruthi, and Newman 2015). Its popularity in the HCI and Health community is noted for its use in studies related to personal informatics, the Quantified Self, lived informatics, and self-monitoring of health and wellness. ESM allows for soliciting of information from its users about their behavior and contexts at a relatively high frequency (often 6-10 times in a day). Immediate responses reduce recall biases that may occur when using retrospective methods such as diary study (Beal 2015:2).

The use of smartphones for ESM studies followed its increasing ownership, improving ease of use, and increasing use of physical sensors. Moreover, smartphones are found to be proximal within the same room as their users almost 90% of the time, and within arm's reach almost 50% of the time (Dey et al. 2011). A recent increase in the adoption of smartwatches and the growing interest in wearables has further intrigued researchers about the possibility of using these as self-reporting devices. Smartwatches by their nature are always within arm's reach when used, allowing for *micro-interactions* that lower the interaction burden, thus increasing reporting frequency (Ponnada et al. 2017). Yet, smartwatch interfaces come with their own set of disadvantages relative to smartphones, such as a high abandonment rate (Gartner Survey Shows

Wearable Devices Need to Be More Useful n.d.), additional effort required to learn their management (Min et al. 2015), and relatively limited screen real-estate.

The interest in exploring newer device categories is guided by the goal of reducing the burden of high-frequency self-reporting and thus increasing the temporal density of reports. Increased temporal density may also imply greater coverage over time and contexts. One of the main strategies used is to reduce the time a user spends accessing the device (access time) while aiming for a high compliance rate. Multi-device experiences that take advantage of the complementary nature of different device types (Grubert, Kranz, and Quigley 2016) aim for similar goals by allowing the user to choose the most convenient device to achieve a task. A similar opportunity exists for self-reporting applications, where a wide range of multi-device approaches can make it easier for researchers and designers to choose a multi-device strategy that best fits their intended self-reporting application. For instance, in a hypothetical ESM study with children, it may not be wise to use smartphones at schools as their use may not be permitted. Instead, such a study might leverage single-purpose devices designed specifically for the intended application of self-reporting by children.

The Internet of things (IOT) paradigm in itself represents many trends, one of them being the ability to design and build physical computational devices at low cost. This opens up an opportunity to build new forms of self-reporting devices that may come with their own sets of tradeoffs. In this paper, we present a class of self-reporting devices that are situated in the environments of their users, providing a more convenient way to self-report in certain contexts. Existing Situated Self-Reporting (SSR) devices have been shown to be unobtrusive and convenient tools that allow users to log repeating actions (Yekeh et al. 2015).

In this paper, we first articulate the design space for SSR devices, enumerating key design dimensions other than their *situatedness,* such as their mode of interaction (e.g. touch, voice), complexity of supported constructs, and energy requirements (e.g. battery-powered, connected source). Building upon this, we present the design and implementation of Heed, an instantiation of an SSR device that is distributed, low-power, low-cost, supports multiple simple constructs, and has a wooden enclosure with embedded sensors that support touch interactions.

To explore the tradeoffs offered by SSR devices, we conducted two studies of one-week duration, one with 10 and one with 7 users in their natural environments, with two distinct types of device interaction. The qualitative and quantitative findings from the paper will inform the future design of situated self-reporting devices.

In the rest of the paper, we review existing literature and highlight the opportunity to the design space for Situated Self-Reporting (SSR) devices. We then present the design and implementation of Heed, an instantiation of an SSR device. Based on the findings from a one-week pilot with 10 users and another with 7 users, we uncover insights on the characteristics and advantages of Heed. Qualitatively we uncovered themes of use along dimensions of user preference such as situational and social context, notifications, and location of use. We show that Heed devices were complementary to smartphones in nuanced ways. While the advantage of Heed was its single-purpose and dedicated location, smartphones provided mobility and flexibility of use. Finally, we discuss what we learned about the design of SSR devices from the design and evaluation of Heed. The lessons we learned have implications for the design of SSR devices that may be deployed in future real-world studies. Moreover, we present some unintended consequences of SSR devices that researchers should be aware of.

Related Work

Self-reporting involves manual work that imposes a high burden. Self-reporting by Quantified Selfers is noted to cause fatigue that may lead users to abandon self-tracking (Choe et al. 2014). HCI researchers have thus explored multiple approaches to how the burden of self-reporting can be reduced. Such approaches include: utilizing novel user interfaces (Truong, Shihipar, and Wigdor 2014:2; Zhang, Pina, and Fogarty 2016), understanding the role that choice of strategy plays in answering research questions (Lathia et al. 2013), and understanding the effects of choices of devices (Hernandez et al. 2016; Ponnada et al. 2017).

Moreover, self-reporting approaches are commonly subject to bias due to the times available to a person to respond. For instance, if smartphones are used, users can respond only when they are able to engage with them. A user who is away from the smartphone may not wish to or be able to engage. Similar to the choice of device, choice of trigger strategy also leads to biases in reporting. Lathia et al. 2013 demonstrate that the specific design of an experience sampling approach (e.g. random sampling, contextual sampling) is linked to bias in the resulting data (Lathia et al. 2013). Furthermore, it can be argued that infrequently occurring events of interest may occur at times when either there is no trigger for response (no notification or the user doesn't have access to the smartphone, or the user doesn't remember to initiate an interaction), or when a user perceives the smartphone to be intrusive or stress-inducing and decides not to interact with it.

The difficulty in accessing the smartphone further contributes to the burden of using it to answer questions, especially when smartphones are often not within arm's reach (Dey et al. 2011). This perceived barrier to the behavior of self-reporting may result in lower compliance with ESM studies or a lack of motivation to use quantified-self applications. ESM studies have found the optimal frequency for prompts to mitigate annoyance (Wen et al. 2017). Intrusiveness is also known to be alleviated by the use of decision theory methods to generate prompts at opportune moments and reduce the overall number of prompts (Mathur, Lane, and Kawsar 2016).

The earliest ESM studies used dedicated, single-use reporting devices such as pagers to remind user to initiate paper reports (Csikszentmihalyi and Larson 2014). These devices were designed for the single purpose of generating self-reporting reminders. Although the same kind of visual triggers continue to be used in smartphones to solicit reports from users, the single-purpose nature has been lost. It may be argued that the multi-purpose nature of smartphones may have reduced the perceived importance of the self-reporting task.

One way to minimize access time while keeping the single purpose is to design dedicated self-reporting devices placed in the environment of the user. SAL (a simple, situated ambient logger) was found to be unobtrusive and convenient for logging progress towards behavior goals (Yekeh et al. 2015). The lower cost of IOT provides an opportunity to design SSR devices that are personalized for an intended user and application. Inspired by similar comparative studies of self-reporting devices that quantified the trade-offs of using earlier approaches such as PDAs (Burgin et al. 2013), pen-and-paper (Berkman, Giuliani, and Pruitt 2014), and more recently smartwatches (Intille et al. 2016) and head-mounted displays (Hernandez et al. 2016), we sought to explore the design space of low-cost self-reporting devices that are situated in the environment of the user.

Design Process and Implementation

Situated Self-Reporting Devices

Situated devices that are commonly found in any modern home (e.g. thermostats, home assistants, refrigerator) are placed by users in their environments to fit certain physical constraints or satisfice to meet their personal preferences. The *situatedness* of such devices provide the user with ease of access in the relevant context. We expect that the ease of use offered by situated devices can also be leveraged to lower the burden of self-reporting in certain contexts. In the following paragraphs, we first define situated self-reporting (SSR) devices and briefly discuss some of the design dimensions that came up from our brainstorming. We then present an instantiation of SSR device, that we designed, built, and evaluated in order to explore the design space of SSR devices.

SSR Device is a situated device intended to be placed in a location to optimize user's self-reporting efficacy. Smartphone and watch interfaces for self-reporting have seen much interest by researchers in their design-space explorations. The design of SSR devices remains largely unexplored, however, with the notable exception of SAL (Kummerfeld et al. 2015), a small, situated, ambient logger designed for personal goal tracking. SAL, inspired by Weiser's original vision for ambient, calm, and peripheral computing (Weiser and Seely Brown n.d.), was found to be an unobtrusive and convenient way for users to log repeated actions that they intended to perform. Our abstract notion of SSR devices extends these ideas with the aim of systematically exploring some of their key design dimensions.

The design of SSR devices can vary along any of the design dimensions associated with situated devices, only some of which may directly affect the self-reporting behavior of users. For instance, interaction with SSR devices is meant for the specific purpose of self-reporting on a given measure. The type of interaction (e.g. touch, gesture) supported by a device may thus depend on what is to be measured by it. For instance, if the goal of an ESM study is to measure the stress level of the user while driving, an SSR device in the car may use voice interaction

Figure 3. (Left) Individual components of each device laid out. (right) Final devices

rather than touch interaction. In the same scenario, an SSR device in the car may utilize the car's power source to charge itself rather than relying on its own battery.

In this paper, we aim to illuminate the design space of SSR devices using Heed, a system consisting of a simple instantiation of SSR devices. Heed devices are intended to enable general users to report simple measures of activities, stress, sleepiness and social context. Heed devices are low-cost SSR devices, distributed in certain chosen locations by the user to optimize the user's self-reporting behavior. Table 3 describes Heed's characteristics, along with other enumerated design dimensions of SSR devices. A Heed device is designed to complement or challenge a user's preference for the smartphone as a self-reporting device. We evaluated the use of Heed devices in a real-world study to investigate the influence of key characteristics such as the location and context of its use on users' willingness to report.

Implementation

Heed Devices

Heed devices are to be located in the most visited spaces of the user. Moreover, the form of a Heed device must evoke a sense of duty to report; yet, being a physical part of user's décor, it must not stand out. Such opposing intentions directed our initial exploration of materials and the physical design of Heed such that the devices were neither too ambient nor too distracting. Each device was intended to be placed in a location of the user's choice. This constrained our design to a device that must run continuously, without needing the user to charge it, for the duration of the study. Furthermore, we designed the device to sync with the user's smartphone via Bluetooth for two reasons: a) it allows the device to trigger notifications only when the user is nearby and b) it allows the device to sync in real time, thus avoiding unnecessary lag.

We used a circular soft-potentiometer that consumed relatively less power than capacitive touch. The device consists of a microcontroller, a Bluetooth Low Energy (BLE) module, a linear touch sensor, and an LED. We used an off-the-shelf low-power BLE + microcontroller module (BLE Nano n.d.). Figure 3 shows each of these components laid out on a table. The devices were optimized for low power, going to sleep at night and after a report was made. To minimize the space requirements, we designed our own Printed Circuit Board, integrating all the components in a small form factor. The code was written and uploaded to each device using the Arduino IDE.

We chose the form of the device to be compact and have a round shape for aesthetic reasons. A minimal form factor allowed us to fit up to seven touch points, thus supporting interactions necessary for simple constructs such as a five-point Likert scales. We chose wood as the material for the enclosure for two main reasons. Firstly, wood is reported to inspire a perceived sense of durability (Odom et al. 2009). Secondly, we received positive feedback about the wood during our iterative design process. The wood enclosure was overlaid with a button map printed on glossy sticker paper, allowing it to be easily customized for the intended self-reporting task. A laser cutter was utilized to prepare the wood and paper enclosures.

The final design of the Heed device offers personalization for the user, the specific application of self-reporting, and the intended context of the interaction. For instance, the device interface can be easily appropriated to the intended research question by allowing a different paper enclosure to be pasted on the top. A disadvantage of Heed's design is the lack of flexibility offered by the interface with respect to the dynamic content that can be shown on a smartphone screen. We note that such constraints are mostly due to the cost and energy requirements of Heed and may be circumvented by other SSR devices.

Table 3. Selected design dimensions of SSR devices and the instantiation of Heed on those dimensions.

Design Dimension	Description	Heed's characteristics
Mode of Interaction	How does the user interact with the device? Possibilities include voice prompts, touch or gestures.	Touch
Notification type	How is the user reminded to self-report? Possibilities include haptic, sound and lights	Single light notification
Construct complexity supported	What kinds of self-reporting constructs are supported by the device? Device may support a one or more than one simple construct items (e.g. stress level), or may support multi-dimensional constructs (Mood-Affect) or composite constructs.	Multiple simple constructs
Context Awareness	How much does the device adapt to its sensed context? For instance, its notification is triggered by user's proximity or other contexts.	Notifications are shown when the user was nearby
Number of users	How many users does the device serve? Possibilities include a single user device or a multi-user device.	Single user
Distributed	How many devices does the user use? A system may consist of one or more than one self-reporting devices	Heed is a single user - many devices
Material	What materials were used for the device? Materials affect the durability as well as the feelings that may evoke in participants. Possibilities include wood, ceramic, and plastics.	Wood along with a glossy paper overlay for the interface.
Cost of device	How much does it cost to make the hardware?	Very low cost (~$5)
Energy requirements	How is the device powered? Possibilities include an always connected power source, user is asked to charge, and self-contained power source.	Runs without charging for ~7 days
Connectivity	How does the device send or receive data? Possibilities include direct Wi-Fi connection, user's phone, etc.	Uses Bluetooth to connect to phone
Other reporting features	What are the ways user can report on the device? Possibilities include participatory, in-situ and post-hoc. Does the device provide the ability to edit/ undo reports previously made?	Only in-situ
Purpose	What is the purpose of the device? Is it a single-purpose device? Or is it multi-purpose? If a conversational home-assistant is used for self-reporting, it will be a multi-purpose SSR. SAL had two main purposes of logging and reflecting on one's progress towards a goal.	Single-purpose

Heed Phone App

The app is instrumented to perform four fundamental tasks: a) allow self-reports to be made on the smartphone, b) collect the user's location, c) collect smartphone and Heed device usage data, and d) manage Heed devices. Notifications to self-report remain visible for one minute and then disappear (Figure 4). The app tracks the user's location every five minutes in addition to whenever an action is performed. The app connects with and syncs data with nearby devices. It could trigger a pre-programmed or a dynamic notification on Heed devices. The app also allows the user and researchers to configure the time window within which notifications can be made, and other study-relevant parameters.

The software was written in HTML/JavaScript using Cordova libraries (Apache Cordova n.d.). Although the app works on multiple mobile platforms, Android was targeted as the distribution platform because it allowed the application to run in the background. The software required Internet connectivity to transmit participants' reports and other system logs.

Evaluation

The design process outlined above culminated in two one-week-long studies, one with 10 and the other with 7 externally recruited participants. The rationale for using two studies was to explore the use of devices for reporting different constructs that required different kinds of interactions.

Stress and

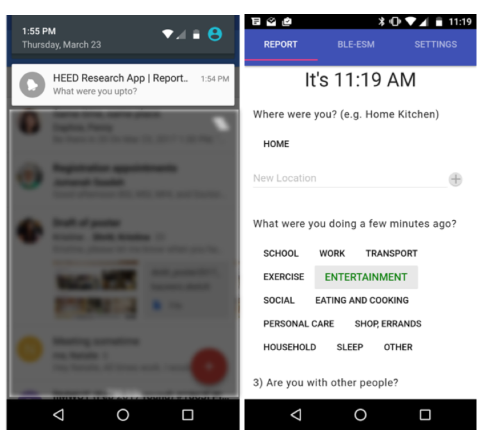

Figure 4. (Left) Notification from the Heed app. (Right) The reporting interface on the app.

Table 4. Details of the two studies

	Study 1	Study 2
Self-reporting measures	Activities and Social	Stress and Sleepiness levels
Interaction type	Activities: Select from options Social: Tap if with people	Likert scale (1-5)
Study design	2 days only Phone 2 days only Heed 3 days both Phone and Heed	7 days both Phone and Heed
Number of Heed devices per participant	5	3

sleepiness may be reported on a Likert scale, while activity reporting requires multiple options. All participants were selected for the one-week study based on their smartphone's compatibility with our apparatus (any version of Android 4 or later was allowed).

In both studies, participants were asked to use their smartphones as they normally would. They were asked to choose either the smartphone or a Heed device based on what they found convenient. The incentive for participating in either study was $65 for seven days of reporting.

Web-diary interface. At the end of the day, we asked participants to use a web-based diary tool that we developed to verify the data they provided throughout the day, as well as share any overall comments or feedback. This data was used to guide our final interview. The web diary is essential for this particular study as it allows us to measure the accuracy of Heed and smartphone reports.

Final Interview. At the end of the study, all participants were asked to return to the lab for a follow-up interview. At the end of the final interview, we asked participants to answer a post-study questionnaire that asked them to evaluate certain characteristics of the two device types, including the comfort of use, the effect on social interactions, the effect on their stress levels, and the likelihood of future use.

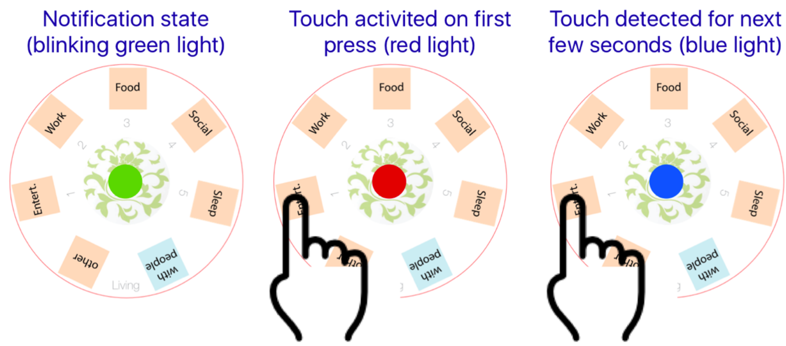

Figure 5. In Study 1, users reported their activities by pressing the button with the label of their activity. They reported their social context by pressing the button with the label "with people" if they with other people.

Study 1: Activity and Social Context Reporting

In the first study, we recruited 10 participants for a one-week field evaluation. The study was conducted between April and May, 2017. Each participant was provided with five Heed devices, out of which one was to be carried around. We asked participants to report on two measures: a) the activities they were doing, and b) if they were with other people. Activity tracking is commonly used in time-use studies. Moreover, reported activities provide deeper insights into the context in which a device or smartphone was used.

During the initial interview, we asked participants to reflect on their geo-spatial patterns and space use in their most visited spaces. We then guided them to choose the locations where they would place the devices (kitchen, bedroom, etc.) and the activity labels on the devices, relevant to those locations (e.g. bedroom-> sleep, personal care, entertainment, etc.). We then customized the five devices for each participant by printing the overlay for each device on sticker paper. Each overlay, consisted of seven buttons, 5 buttons were labelled as the activities that user decided for the location of that device, one as "other", and one as "with people". In our backend, each device mapped to a single location and the buttons on the device to a list of activities as printed on the overlay (that are likely to happen in that location). Participants were asked to report either during an activity or at the end of an activity. If they had begun a new activity and it

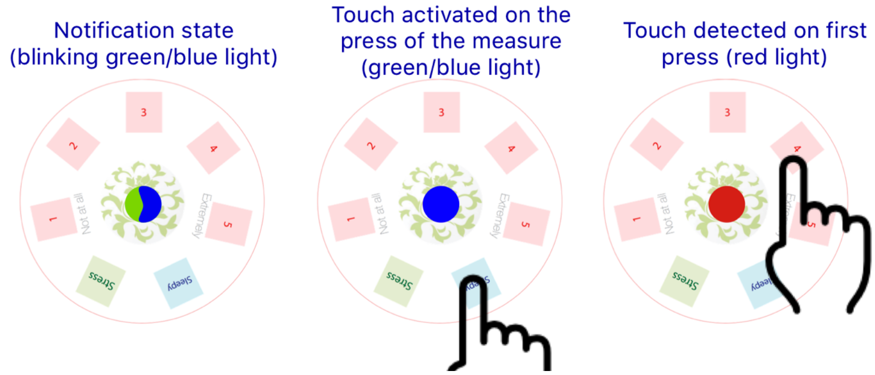

Figure 6. In Study 2, users reported their stress and sleepiness reporting by first pressing the measure and then a number indicating their level

had been less than 15 minutes, they were asked to report the previous activity. Participant could also report their "Sleep" right after they wake up.

On Heed, the user reports their activity by pressing the button with the label of their activity (Figure 5). The activities were chosen and adapted from the literature (Ainsworth et al. 2000). On the phone, participants were instructed to report their location with room-level granularity (e.g. home-bedroom) as we are specifically interested in indoor activities.

Participants were asked to use their smartphones as they normally would. We divided the study duration into three phases. For the first phase (~2 days), participants reported only via their smartphones. In the second phase (~2 days), participants reported only via the Heed devices. In the third phase (~3 days), participants chose either the smartphone or the devices when reporting. The distribution of days was selected for qualitative purposes, to help participants distinguish between the use of smartphones and Heed. We assigned the maximum number of days to the simultaneous use condition in order to study the use of the devices as intended.

Study 2: Stress and Sleepiness reporting

In the second study, we recruited 8 participants for a one-week field evaluation. The study was conducted between May and June, 2017. Each participant was provided with three Heed devices to be placed in three of their most-visited spaces. We asked participants to report on two measures: stress and sleepiness. Stress and sleepiness are commonly tracked in ESM studies. We used a five-point Likert scale for the two constructs as suggested by other works (Hernandez et al. 2016; Wild-Hartmann et al. 2013). The interfaces for activity reporting on Heed and smartphones are shown in Figure 6. Similar to the first study, during the initial interview, we guided participants to choose locations where they might place the devices (kitchen, bedroom, etc.).

Method

Data Analysis

In-depth semi-structured interviews were conducted by the first author with all 17 participants. Each interview lasted approximately 45 to 60 minutes. The qualitative data analysis harnessed open coding and thematic analysis techniques but partly relied on a phenomenological analysis strategy during the first stage of reading the transcripts (Creswell 2013; Lopez and Willis 2004). Such a strategy allowed us to construct participants' daily experience with the Heed devices and the smartphone application. Three members of the research team read transcripts while listening to audio-recordings in an effort to reflect the on-site presence and participants' context in the analysis process. We highlighted meaningful statements and noted spontaneously emerging thoughts in the margins of the transcripts (bracketing (Shosha 2012)). The purpose of this practice was to help us reflectively set aside our prejudices or preconceptions regarding Heed and the smartphone application usage and to attain a certain level of "neutrality" during the analysis process (Guba and Lincoln 1985).

During this process, we extracted 156 meaningful statements and created memos on each of them. Each memo contained participants' feedback, memorable experiences, and thoughts for further improvement, with regard to the Heed device and the smartphone application. In another round of analysis, coded data were grouped under themes using affinity diagrams (Wikipedia 2017a). This initial construction of the data structure went through a couple of verification processes of active reflection on the interview transcripts and the audio-recordings, and it was reconstructed through the data reduction. Five theme clusters are presented as a result of the final data reduction: 1) Reporting behavior, 2) Role of notifications, 3) Complementary nature of smartphones and Heed, 4) Location of Heed Devices, and 5) Unexpected social events.

Table 5. Demographics of the study conducted.

	Study 1	Study 2
All	10	7
Male	7	4
Female	3	3
Age 18-25	3	0
Age 26-35	7	4
Age 36-45	0	1
Age 46-55	0	2
Occupation: Academic	10	1
Occupation: Professional	0	6

Data Overview

One participant dropped out of Study 2 as she had privacy concerns about leaving her Bluetooth and GPS switched on at all times. Two participants reported having participated in ESM studies before. Five participants reported having done some kind of self-reporting in the past for personal reasons. Almost all participants reported that they kept their smartphones in silent or do-not-disturb mode in their natural use. One participant reported changing the notification mode for the study so that she could report more. Most participants in Study 1 mentioned that they would like to decrease the time they spent on the phone, and would not like to use it while working. In the pre-study questionnaire, all users mentioned that they generally kept their phones in silent or Do Not Disturb mode. Table 5 shows the distribution of our participants in terms of some demographic variables.

In the post-study questionnaire, we compared participants' scores for Heed and smartphone for the following characteristics: the comfort of use, the effect on social interactions, the effect on their stress levels, and the likelihood of future use. There was no significant difference between the scores. We observed an increase in participants' overall scores on quality of Heed reports from Study 1 to Study 2 (p=.08). This was probably because a number of technical issues related to the unresponsiveness of devices were fixed prior to Study 2, thus making the Heed devices more reliable in the second study.

We received much feedback on the physical form factor and choice of materials of Heed. The overall feedback was positive, with a few exceptions. The choice of LED light for notification elicited feedback from participants, with many of them suggesting the use of sound in the future. We do not go into detail regarding such findings as we wanted to restrict ourselves to those findings that are most relevant to SSR devices.

Results

Our findings shed light on the use of smartphone and Heed devices for self-reporting in the two studies we conducted. Overall, we found that both device types presented some advantages over the other. For instance, almost all participants felt the smartphone was more convenient to report from while on the go. On the other hand, the Heed devices were preferred for their simplicity of interaction when they were within reach and visible.

In the following subsections, we first briefly discuss descriptive quantitative data on the reporting behavior for the two studies. We support some of the analysis with quotations from participants to further clarify the findings. Following the quantitative analysis, we present the most important themes that emerged from the analysis of the qualitative data. These themes included a) use of notifications; b) the complementary nature of smartphone and Heed devices in terms of their mobility, use in different situations, and overall use; c) use of smartphone and Heed devices in different locations; and d) the impact of the physical characteristics of the Heed devices. Each theme is supported by participant quotes. The participant ids are in the order of their participation, so P1-10 were in Study 1 and P11 to P18 were in Study 2.

Reporting Behavior

An important goal of an ESM tool is to ensure a high frequency of reporting, ensuring coverage over different situations. This section compares the report frequency when using Heed devices with the frequency when using a smartphone. Since our study intentionally allowed participants to report at their own discretion, we separate the analysis of those reports made in response to notifications from those that were made by participants at their own discretion. Moreover, for Study 1, our analysis only compares the condition when both the smartphone and Heed devices were in use simultaneously.

Overall reporting frequency

Our analysis showed no significant difference in the report frequency for smartphones and Heed devices, suggesting that overall the use of Heed was comparable to that of the smartphone. The report frequency is calculated by dividing the number of reports made by the total duration of the study condition. Total duration is calculated as the number of days that passed between the first and the last report. For instance, participant 1 made 19 reports in 2.8 days, making the report frequency equal to 6.7.

Effect of (not) having notifications

One of our design motivations was to know if physically situated Heed devices would elicit more responses because they act as triggers for users to report without the need for notifications. In Figure 7, we compare the report frequency of the respective devices in the two notification conditions. We observed a significantly higher report frequency via the Heed devices when users reported at their own discretion ($p=.02$) in Study 1. There was no significant difference in the report frequency or the compliance rate of making reports when user were notified. We did not see a significant difference in Study 2.

In Study 1, we also see that the report frequency for the smartphone was significantly higher when users were notified than when users reported on their own. The opposite was true for Heed devices; that is, the report frequency was significantly higher when users reported on their own than when they were notified ($p=.02$). This indicates that Heed devices work better when users are asked to report on their own accord. For Study 2, we did not observe any significant effect.

Effect of activities, time-of-day, and social context

Knowing the relationship between activities and social contexts of a user can give us further insight into the use of SSR devices in conjunction with smartphones. Participants reported on their activities and social situations in Study 1, and their stress and sleepiness in Study 2. This additional data, along with the timing of each report, provided us with additional context for their reports. We observed significantly higher report frequencies from Heed devices when users reported the activities food (p=.03) and sleep (p=.01), and when they reported being alone (p=.01), while there was no significant difference when they reported engaging in other activities or being with other people. For Study 1, we observed significantly more reports in the morning (p=.02) and afternoon (p=.01), while there was no such difference in Study 2. For Study 2, we see significantly more reports from the smartphone when participants reported their stress to be at level 2 and 3 (5 being the highest). We didn't see any significant difference in the sleepiness reports.

Role of Notifications

Perceptions of phone use affect use of Heed

Our quantitative analysis of Study 1 showed that users reported on Heed significantly more often than on smartphones when reporting on their own accord. One reason for this is suggested by the difference in participant responses in the initial and post-study questionnaires when asked about the future use of each device. Participants in Study 1 scored smartphones significantly lower than participants in Study 2 (p=.04). Moreover, in the pre-study questionnaire, Study 1 participants (mostly academic) qualitatively reported their perceptions of using smartphones as something they would like to avoid more in their daily life. For instance, P6 said, "*Yes. I try to stay away from it during office hours: 8 am - 5 pm, Monday – Friday."* This view on phone use was absent from participant responses from Study 2. For instance, P14 said, "*I look at my phone at least every 15 minutes, except for during some blocks of time at work when I am in meetings.*"

Heed devices serve as a physical reminder

Qualitatively, our expectation about Heed devices serving as a physical reminder was confirmed. Several participants reported responding to Heed devices for this reason. P7 mentioned this explicitly while comparing the access times between the two devices: "*I see them around the house, which is, I guess, a more physical reminder. On the phone, I had to physically open the app. I had to navigate through the screen to get there. With this device, it's just there.*"

Notifications can be ignored for multiple reasons

Notifications are an integral component of any self-reporting application. Participants recalled many instances when notifications worked and when they did not work. Most interesting were those that talked about when notifications were ignored.

Several participants noted the lack of importance they attached to smartphone app notifications, leading them to completely ignore the notifications in many cases. A reason often mentioned was the smartphone notification being one of many other notifications the smartphone showed them. P7 said, "*On the phone, (ignored) quite often. It would tell me, "Why don't you report now?" And I'll be like, "It's okay. I can do it later." Probably because the notification on the phone is less intrusive, I would say.....Yeah. I would say I ignored it pretty often on the phone.*"

Additionally, being involved in another high-priority task on the smartphone also affected the likelihood of responding to self-reporting notifications on the smartphone. P16 reported ignoring the notification when involved in another task: "*On the phone it was usually like*, '*Oh, I see this notification, but I have another thing I need to use my phone for right now so I'm going to ignore it.' I'm trying to do something at work or I'm trying to get ahold of my partner or something like that. I would just ignore it…* "

We expected that participants would be less likely to ignore notifications from the Heed devices; however, several participants reported ignoring those notifications as well. Participants

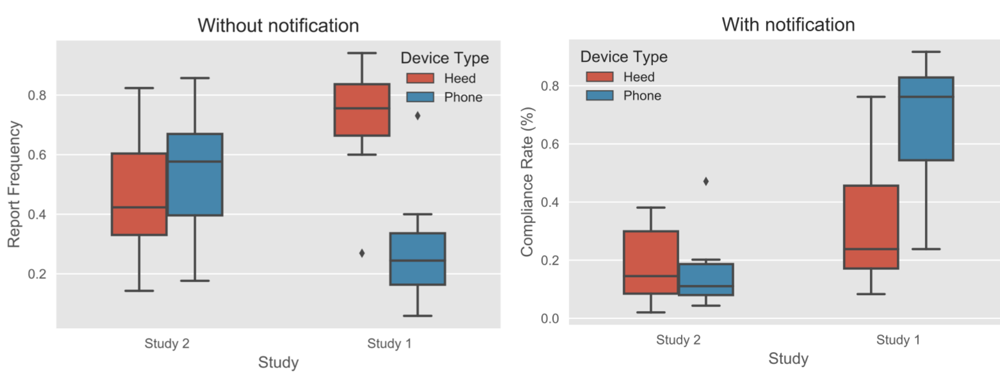

Figure 7. (Left) Comparison of report-frequency between devices when the user reported at their own discretion in the two studies. (Right) The compliance to notifications for each device type in the two studies.

mentioned the repetitive nature of manual self-reporting, indicating a decreasing level of interest as they continued using the devices. P15 noted ignoring Heed notifications when he felt that nothing had changed since the last report. He said, *"Even though I see a notification I wouldn't report it because I would have sent it just one hour before, so I would still wait for half an hour to do it. It's not that I don't see the notification. I see it, but I ignore it. I don't ignore it, I just procrastinate it because I am very concentrated, so my levels wouldn't have changed."*

We also saw notifications on both smartphones and Heed devices being ignored when participants felt tired or lazy, making users procrastinate when tired. P9 said he felt a lack of energy at night to report: *"I was like doing something, "Oh, yeah. I'll do that later," or something. It's not like consciously avoiding that kind of thing, but being reluctant to actually do the job. Yeah, and especially late at night. After 10:30 and that kind of thing, I would say."*

Complementary Nature of Smartphone and Heed Devices

Participants' preference between the self-reporting devices varied from time to time, with preferences over mobility, interaction and activities being the most important factors.

Smartphones when moving, Heed when stationary

In Study 1, we expected that participants, even while moving around, might find using Heed devices to be more convenient than using smartphones because of the shorter access time. Although we found this to be true in some cases where participants noticed the device in their pocket and were reminded to report, we found that most participants found carrying an extra device for self-reporting to be redundant. P5 said, *"Like, yeah. 'Cause the thing is what makes the device redundant is because I'm always having my phone when I'm outside my room, so I just find it like painful to like report on a separate thing as opposed to a phone, but then like in the bedroom, I'm almost like, I don't sit with the phone in my pocket, so like the device is more convenient."* Although Heed devices were not preferred for carrying around, our assumptions about the tradeoffs offered by SSR devices was confirmed as participants preferred Heed when they were stationary and smartphones when they were moving. P16 stated this in her own words: "*I think they're really complementary in these being stationary somewhere and then the phone being the piece if I'm like away for an extended period of time. They're extremely* *complementary. Where I would probably be less prone to use them is if this was to travel around with me.*"

Heed's low interaction burden in certain situations

Heed, by its design, required less time to access and start interacting with. One of our assumptions about Heed devices imposing a lower interaction burden was confirmed by participants' reactions as they pointed out many reasons. For this analysis, we ignored those that were related to the nature of micro-interactions, because smartphone interfaces could have used lock-screen widgets or gestures. Instead, we analyzed other instances which highlight some of the SSR characteristics of Heed, such as its single-purpose nature. For example, P7 pointed out that the interface of Heed differed from that of the smartphone in one important way: the Heed interface never goes away, while it's easy to "leave" an interface on the phone. She said, "*Maybe it feels less effort to report on device. ... Maybe because all the choices are presented in front of you, again, in a physical sense. And then you don't have to leave forever. I guess, if I'm using my phone and I need to report, I have to, if I'm not ... If I'm using it, that means I have to leave whatever I'm doing. Or if I'm not using it to do the whole motion of opening it up and everything again. Sometimes it takes a while for the screen on the phone to load, whereas the device is always there*".

Heed's shorter access time is only available to its user when the participant is near the device. This meant that only certain situations (activities and locations) allowed Heed's advantages to be successfully leveraged. We note some of these in the following paragraphs.

Heed when engaged in Focus work

We found certain situations to influence the preference for one device over the other. One such situation that several participants pointed out was that of being engaged in focused work. During this activity, participants were more likely to be distracted by a smartphone than by Heed. P2, a Ph.D. student, highlighted the less disruptive nature of Heed: *"I really like entering on the device like this. It's not as disruptive to my workflow compared to the phone, which usually leads me to check other notifications and emails… (Heed) is already on my desk, I can see the light flashing sometimes, and if I'm focused on work, it doesn't distract my attention by leading me to check other activities on my phone."* Moreover, she described the smartphone as distracting when she was trying to focus on her work: *"Yeah, it depends on my work. If I have a meeting that day, or if I'm just feeling really focused, then I liked using the [Heed] device, because it's not as distracting. You know, once you check on the phone, or once your interest is on the phone, you end up checking other stuff on the phone as well."*

Heed when with friends

Our quantitative findings showed greater use of Heed when participants were alone. Some participants noted that this happened because when they were with friends, they were likely to zone out and not look at the devices. P18 stated that he preferred to use the smartphone when he was with his friends but he reported much less often in such situations, while he reported using Heed a lot more at work when alone: "*So when I am always out of my desk, like I was hanging out with my friend within this weekend, at those times I had to report with the phone... It's like I report way less compared to when I do it in front of my desk. When I'm doing it in front of my desk, whether I busy or not I don't care. It's just very easy to do. On a phone probably I will not do it… I think for work this (Heed device) is amazing.*"

Heed when phone is not near

Situations when users preferred Heed also included those when the phone is not available nearby. Several participants pointed out cases when their phone was charging and Heed happened to be nearby. In P15's case he had to choose between going downstairs or choosing the nearby device: "*I was seeing some TV series on my couch, and the device was on the coffee table and I saw it blinking. And my phone was downstairs. This walking distance is less than going down. I took this device and just recorded from there."* Similar instances of the smartphone not being around may also occur when a participant has moved from one place to another: "*Like when I'm at my desk, my phone might be on the bed. But, I tended not to move the [Heed] device around then. So the device was good then, in such situations."*

On the other hand, there were several cases where participants preferred to use smartphones. P2 noted that when she was already engaged with the smartphone, she preferred that. P2 reflected on this in her daily diary: *"I'm starting to think that whether I record on the phone or the device more often is mostly driven by whether I have a lot of work that day (device) or if I'm checking my phone a lot (phone). It's just whichever is more readily available."* The smartphone being a multi-purpose device allows users to engage in many different activities. Although smartphones were preferred during less important activities, notifications were found to be ignored when the activity on the smartphone was too important to be distracted from. For instance, P16 reported ignoring notifications when he was trying to reach his partner or get some work done: *"When I noticed it on the device I would almost always report. On the phone it was usually like, "Oh, I see this notification, but I have another thing I need to use my phone for right now so I'm going to ignore it." I'm trying to do something at work or I'm trying to get ahold of my partner or something like that. I would just ignore it, blaze past it and go into whatever task I was doing. I would find I didn't feel like I had to return to report on that."*

Location of Heed devices

To be useful, self-reporting devices must have good coverage of contexts such as location. Our quantitative analysis shows a noticeable difference in reports made in the bedroom (p=.07) with the report frequency via Heed being higher than via smartphone. There was high variability in the report frequency across participants across indoor locations (Figure 4). Qualitatively, we found the following reasons for high variability: the type of room, the space use patterns of the user, and the visibility of the Heed devices.

We found that the locations of Heed devices was regarded as good when the devices were easy to notice, especially when they were in the field of view of the participant. P4 reported that the Heed devices were easy to find when they were in the line of sight: *"The one on my desk is right in front of me just below the monitor, I'm always seeing it. The one in the bedroom is right next to the bed, next to the mirror and I use the mirror … It was in an ideal location so I always saw it."* P2 mentioned that devices were also likely to be found when they were within the peripheral vision: *"Basically I will see it out of the corner of my eye, if I'm looking at my second screen.(I have two screens at home.)"*

Heed devices were reported to be used more often because of their relative proximity to the user. For instance, P7 found Heed devices more accessible than her smartphone when she was at work, as she would often forget her smartphone in her jacket pocket in another corner of the room: "*Oh, I guess the phone is more often out of the field of view. Sometime I will leave it in my* *pocket, in my jacket just leaving it hanging there, not near me but at the corner of the room. Yeah, but the device, the BLE device is always there at the corner of my desk."*

Heed devices were also noted to be used more when placed in a frequently accessed space. For instance, P15 mentioned using his device at his desk more as it was placed in an area that was accessed a lot and also had good visibility: "*[Heed device] was on my table… I keep my laptop here, and on the left these are my snacks and everything else… Having the device in an area where I would generally go to control stuff or probably do something, as I thought that would the most obvious place that I would take notice of the notification, so I kept it there."* Furthermore, spaces that were visited frequently, especially during the transition from one room to another, also made good locations for Heed devices. P16 said, "*Essentially I have a little ledge that separates our living room from the hall but we still have our bedrooms. That ledge also has a staircase next to it, which leads down to my maker space. So I pass that ledge when I'm at home like 50 times every day, not every hour. So I'm always walking by it so it's really easy to just always log everything from that."*

A characteristic of the room that was found to influence device preference was the size of the room. It was noted by some participants that it was easier to use Heed in smaller rooms than in larger rooms. P5 said, *"… especially like my living room and kitchen and stuff were just like much bigger, right? So it's like hard to have a stationary device and all that, like I won't remember to see in which case the phone worked better… But then in my bedroom because like it's much smaller and I could see the device from almost any angle and the green light. So the device worked better in like the bedroom…."*

Unexpected Social Events

The physical presence of the Heed devices was reported by several participants to have attracted the attention of their friends, family and officemates. For instance, P14, who lived with her family, mentioned that the blinking notification on the Heed device turned into a family event: "*One of the times I was outside doing yard work over the weekend, and my partner ran outside with one and was like, "The light is flashing." It was so funny. So the whole family got involved. So I pressed the button and the device was brought back inside.*"

The social nature of Heed also led to some unintended social situations. In some instances, a participant's acquaintance acted as a proxy for the reminder intended for the participant. In the case of P16, his partner noticed the blinking notification light on the device and let him know: "*The other thing that was kind of funny is as it was lighting up, not only did I notice it, but then other people would notice it. So, if I had my partner was over she would be like "Hey, your thing is lighting up". I was like "Okay, I got to go touch it", and then I would go do it. Then I would explain to them what it was and they would like it and they would think that's kind of cool."* In one case, the participant's partner noticed the device notification and then reported on behalf of the participant. "*Oh, yeah. Actually, one time my boyfriend was in the bathroom and he was like, "Your device beeping." So I actually told him, "Why don't you just report for me so it stop beeping." Yeah. So I just told him, "Why don't you hit the ...," So I basically I told him to choose this option, the one that I was doing."* Here, the participant referred to the "blinking" notification light as "beeping."

In some cases, having novel physical devices created unintended tension and added stress to participants, mainly due to concern about how the devices might be perceived by others. This concern was shared by three participants, all of whom were foreign students. P12 was hesitant to place the device where everyone could see and instead placed it to the side where it was less noticeable to others: *"Electronic device. You would see people putting tapes on their webcams these days. People are so conscious of their privacy and all that, so I didn't want them to suspect that I have some device capturing anything in the office."* A similar issue seemed to cause a

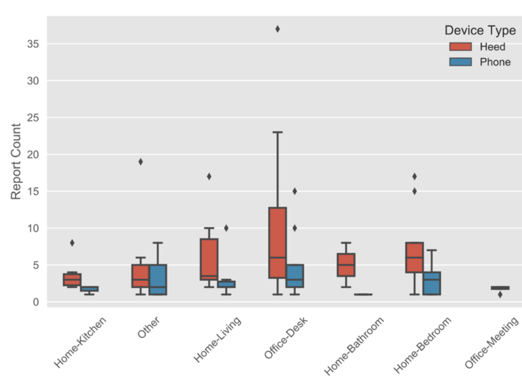

Figure 8. The distribution of reports in Study 1 across different locations. Locations that were not in one of the above categories were categorized as Other.

severe reaction from P6, who was nervous about carrying or using Heed devices in public spaces due to the fear created by the socio-political climate at that time: *"... then I had also this panic attack. Completely unreasonable, I guess, but I had this panic attack. I just want to remember when was it. I can't remember if it was when I was on the bus, or when I was in the workshop, but I had this moment of a sort of a panic attack. It was me thinking, "I am basically using this conspicuous device and pressing it in the middle of a crowded room, and I am a brown man. Are people going to think I am trying to set off a bomb here? I hope not."* A similar sentiment was shared by another foreign student from South Asia (P5): "*I was being very conscious about was like especially yesterday, I was initially going to take the device on the train, but then like because this is like this handmade device, and like you know, considering that someone freaked out about a bomb scare with like the university professor solving a math equation, I was kind of like a bit worried about carrying it with me."*

Discussion

Our findings show that Heed devices are a viable tool to have in a toolbox of self-reporting approaches. Overall, the report frequency of Heed devices is comparable to, if not more frequent than, the report-frequency of smartphones (as in Study 1). We saw that Heed devices benefit from their physical presence in the environment as users reported significantly more via Heed than via the smartphone when they reported on their own accord. Our qualitative findings show that Heed devices and smartphones complement each other in the contexts (activities and locations) in which they are preferred to be used. While the smartphone was preferred in situations where the user was moving around or already engaged in certain smartphone activities, Heed devices were preferred when the user was stationary and engaging in focused work (e.g. when users were engaged in a task and didn't want to be distracted by email or other social communication.) Moreover, we saw a significantly higher use of Heed when participants reported food and sleep as their activities. Anecdotal evidence suggests the perception of the smartphone being too intrusive, increased the likelihood of the use of Heed devices during such activities.

In Table 3, we presented some design dimensions that may be appropriate for SSR devices appropriate and the characteristics of Heed devices in terms of those dimensions. The findings of our study highlighted the effect of certain design choices on the overall use of Heed devices. For instance, the *minimal form factor* and the *low-power* design of Heed devices allowed users to place them in their surroundings with ease. This, in turn, permitted the Heed system to be *distributed*. This led us to some key findings regarding preferences for locations of Heed devices that are generalizable to many SSR devices; that is, SSR devices are preferable to smartphones when they are in the field of view, within arm's reach, in infrequently visited spaces, and in small spaces.

Findings that relate to the situated nature of Heed devices have implications for a range of SSR devices. We noticed that the placement of Heed devices affected how aware users were of a device. This awareness, in turn, reminded users to initiate self-reports. We also noticed that the context of use affected the use of Heed devices. For instance, several participants preferred reporting on the smartphone when they were moving around and when they were already engaged in certain activities. On the other hand, Heed was preferred during activities that led users to be inclined to stay away from their smartphones, such as engaging focused work.

Design Implications

Many themes emerged after characterizing how our participants used the Heed devices and their smartphones to complete the self-reporting tasks. The ways in which they used these devices are indicative of how self-reporting devices should be designed, keeping in mind the actual practices of users. We present these themes as four key lessons learned in the design of SSR devices:

Design or choose self-reporting tools that are suited for your participants

A user's perception of a self-reporting device significantly affects the preference for that device for self-reporting. ESM studies focusing on special populations such as children and elders could consider such preferences before deploying their tools for ESM studies. As an extreme case, doing a study with children may be impossible with smartphones given that smartphone use is often restricted during school hours. However, using an SSR device would allow researchers to solicit information from children on simple constructs in a relatively less distracting way.

We also see that the patterns of smartphone use can affect when users may choose to respond or ignore notifications. Such factors could play an important role, especially when the measure is directly related to such patterns, for instance, if a researcher wants to study the relationship between anxiety and social media using a smartphone-based ESM tool. The self-reporting behavior may be biased depending on the engagement level participants have during social-media consumption, thus affecting their likelihood of ignoring the ESM notification. Ignored notifications may be programmed to return, but such behavior of the app may annoy a user, affecting future participation in the study. Although SSR devices have their own set of biases, they may complement smartphone use in such moments of interest. One may imagine the use of SSR devices in this hypothetical situation. Assuming children spend a significant amount of time in their bedrooms, they may be less likely to ignore a particular ambient light on a device soliciting them to touch one of the buttons.

Leverage multi-device strategies for relevant measures

Self-reporting applications such as quantified-self or ground truth labeling applications target general users. For such applications, a multi-device approach may be more suitable. For example, an activity-labeling study, similar to ours may leverage phone or smartwatch for activities during movement, while it could use SSR devices for indoor activities, especially when the devices are likely to be in the field of view of the user (e.g. on the desk at work, in the bedroom or in the bathroom). These kinds of multi-device self-reporting approaches complement each other. For multi-device studies that span multiple locations, it may be best to use SSR devices in conjunction with smartphones or wearables.

Multi-device experiences intend to provide users with the most convenient way to accomplish a task within their current context (Grubert, Kranz, and Quigley 2016). For instance, smartwatches work best for people who wears them in situations where taking out a smartphone and navigating through its interface is too much of a burden. The notion of SSR devices intends to build on the multi-device paradigm, imagining their use specifically for self-reporting tasks. Home assistant devices such as Alexa and Google Assistant are increasingly being adopted in users' homes. Such situated devices may be used as effective SSR devices, providing advantages over smartphones as shown in the findings of our paper.

Deploy SSR devices for real-world studies

The use of SSR devices for real-world studies depends greatly on where they are placed in the environments of users. SSR devices may be well used when participants expect to spend time in a stationary environment and the devices can be found in frequently accessed or frequently visited spaces. Such placement of the devices can leverage an important advantage of such devices over smartphone; that is, their physical form serves as a reminder to initiate a self-report. Although such constraints may add some burden on the researchers, understanding such constraints for any self-reporting tool can greatly benefit ESM researchers.

Unintended consequences may or may not have a positive impact on the reporting behavior

The physical presence of SSR devices makes them visible to anyone in the space of the user. This may lead to unintended consequences for the user's reporting behavior. Certain social situations may have positive impacts, as occurred in our study; for instance, the partner of a user may notice a notification on an SSR device and then, in turn, remind the user. A user's acquaintance may even report for the user. Although we mainly saw such positive incidents in our study, it is possible that some of the reports may have been made unintentionally by acquaintances. The ability to know if the user is nearby, using the proximity of a smartphone, could be a useful feature in such situations.

The social use of SSR devices is also affected by the socio-political climate around the user and may hinder its intended use. The look and feel of a device may then affect how SSR devices may be perceived by users and others around them. Systems could pay specific attention to the form of devices, such as making the look of the device more polished under such circumstances.

Limitations

Our evaluation of Heed devices with the goal of illuminating the design space for SSR devices has many limitations that may be addressed in future work. A one-week duration of study cannot assess the impact of device novelty. A longer duration is necessary to study the novelty of Heed.

Our emphasis on qualitative analysis allowed us to gain deep insights into the use of Heed devices. Our findings also include results from some basic quantitative analysis. Such results provide us with better insight but are limited due to the small number of participants, the short duration of the study, and a large number of contextual variables.

We also wish to note some limitations of using Heed for other ESM studies. Heed places an extra burden on researchers to build and maintain the devices, thus requiring extra effort from researchers. Another downside of Heed worth noting is the additional effort required in placing the devices for the participants, as it requires thought and effort to understand participants' spatial patterns. We hope that future studies can leverage tools and best practices to accomplish this in an easier way.

We also note that we do not claim an overall lower burden of Heed devices in comparison to smartphones, as assumed in the design of the app. The smartphone app did not use lock-screen widgets or unlock-gestures, which could have significantly decreased the access time. Future studies may further unpack the situated nature of SSR devices separately from the interface elements that lower the burden on users to report.

Conclusion

In this paper, we presented a design exploration of situated self-reporting (SSR) devices. We designed and built the Heed system, consisting of an instantiation of SSR devices that are low-cost, low-power, and have a simple form factor. Overall, we show that Heed devices complemented smartphones in their coverage of activities, locations and interaction preferences.

Chapter 5 Crowdsourcing Messages for Behavior Change Interventions

In Chapter 1, we introduced the *intervention generation* step of a context-aware system. The interventions that rely on sending personalized content usually rely on computational logic to know the content of the intervention. However, current computational techniques fall short in generating certain types of content that are known to improve physical activity in people. One such type of content is physical activity promotion messages. Messages sent to mobile phones have been found to be effective for improving a range of health behaviors, including physical activity, medication adherence and weight loss (Wei, Hollin, and Kachnowski 2011). However, personalizing such messages in a scalable way remains a known challenge. Long-term interventions generally require a large number of messages written by experts. Reliance on experts limits the scalability of such interventions mainly due to the cost of hiring professional writers. Moreover, with the growing use of mobile phones, it is possible to develop applications that personalize messages to aspects of a user's context, such as location, time of day and weather, thus creating a need for an even larger set of messages targeting many possible situations.

We designed a crowdsourcing system to generate high-quality messages for physical activity promotion by employing a three-step pipeline that provides expert-written guidelines for lay writers. The system employs a strategy of generating a large number of messages and then filtering them based on their quality. In this paper, we present the evaluation of this system for generating physical activity promotion messages and empirically evaluated the quality of the messages generated.
We conducted an empirical evaluation to assess quality and cost of crowdsourcing physical activity promotion messages generated by lay writers. We compared crowd-workers with

minimal training to professional writers in terms of the quality and cost of messages they generated as well as their ability to accurately rate message quality. In a blind evaluation, two professionals then evaluated the quality of messages written by experts and online crowd-workers on a 10-point scale.

Overall, as judged by the two experts, crowd-generated messages were as good as expert-generated messages in about 65% of the cases. Moreover, we found that the per-message cost of using crowd-workers was about 15% less than the per-message cost of using experts, showing that the proposed crowdsourcing system is more efficient than experts in generating physical activity promotion messages.

Introduction

Messages sent to mobile phones have been found to be effective for improving a range of health behaviors, including physical activity, medication adherence and weight loss (Wei, Hollin, and Kachnowski 2011). Researchers have laid out guidelines for writing such messages to improve their effectiveness, such as tailoring messages in terms of type of framing (e.g. gain and loss)(Ledford 2012) and supporting participant autonomy (Ryan and Deci 2000). Even though such approaches have been observed to be effective, their reliance on experts limits their use to well-funded or short-duration applications. Experts in health communication usually spend years in training and the cost of hiring them ranges around $30 per hour (Salary: Health Communications Specialist n.d.). This high cost particularly affects tailored physical activity promotion interventions that aim to be used over the long term as they often require a large number of messages to support each user (Patrick et al. 2009).

Advances in technology have opened the possibility of tailoring messages not only to individual characteristics such as the personality (Resnicow et al. 2014) and goals of the individual (Resnicow et al. 2008), but also to more dynamic traits such as the stage of change in a behavior change model such as the Trans-Theoretical Model (TTM) (Prochaska and Velicer 1997), the day of the week, and specific user preferences regarding motivational messages (Morandi and Serafin 2007). Furthermore, the possibility of using user modeling and machine learning (Akker, Jones, and Hermens 2014; op den Akker et al. 2015) allow for personalizing messages to even more dynamic factors such as affect (Liao, Shonkoff, and Dunton 2015), location, time of day,

Table 6. A sub-set of contextual features from Heartsteps (Smith et al. 2017) lead to a large number of situations for only one user.

Location	Time-of-Day	Weather (Intended Location)
Home	Morning	Outdoor
Office	Lunch	Indoor
Other	Afternoon	Outdoor Snowing
	Evening	
	After Dinner	

Total number of situations = 3 * 5 * 3 = 45

weather and the activities of the user (Smith et al. 2017). This level of tailoring creates an explosion in the number of situations for which messages can be tailored.

The situations arise from a combination of different contextual factors. To give a better sense, only three sensed contextual-features- location, time-of-day, weather can enable systems to infer a variety of situations a user is likely to be in. For instance, the situation of the user being in the Office (location), on an afternoon (time-of-day), and it being sunny outside (weather) is one of the situation that is inferred from sensing only three contextual features. Given the large number of situations that can arise from just three features (Table 6), we can expect that the number of situations increases quite rapidly as we add more sensed features. To provide contextually-relevant messages, the system needs a large bank of messages for each of the inferred situation that the intervention aims to address. Moreover, the messages should also be tailored towards stable individual characteristics (for e.g. roles, goals, stage of behavior change) of every user.

The need for a large set of messages creates an opportunity to use novel techniques like crowdsourcing. While health communication researchers are domain experts, crowdsourcing researchers have leveraged paid online crowd-workers on platforms such as Amazon Mechanical Turk [1] (AMT) for applications that are limited by human or technical resources. Crowdsourcing is defined as “a type of participative online activity in which an individual, an institution, a nonprofit organization, or company proposes to a group of individuals of varying knowledge,

[1] https://www.mturk.com/

heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task" (Estellés-Arolas and González-Ladrón-de-Guevara 2012).

Our work exploits this potential for crowdsourcing to write high-quality health promotion messages, thereby allowing scalable generation of messages. Our high-level goal is to design a system that can generate a tailored message library, using crowdsourcing, to support long-term and contextual behavior change that is personalized to the user and leverages the well-established principles of writing content for motivating behavior change.

In this paper, we draw on the health communication and crowdsourcing literature to build a crowdsourcing system that generates high-quality messages for behavior change applications. We present the results from a study that we conducted to investigate the use of the crowdsourcing system we designed for a hypothetical mobile intervention. Our intended application encourages people to walk more. Our main hypothesis is that crowd-workers can generate messages that are close in quality to those generated by experts. Our secondary hypothesis is that crowd-workers can generate high-quality messages at a lower cost than experts. While support for context-aware interventions is a primary motivator for the work in this chapter, we would like to note that this study doesn't evaluate the use of crowd-sourcing for context-aware applications.

We conducted a blind evaluation using two experts to assess the quality and cost of physical activity promotion messages generated by crowd-workers compared to those generated by experts. Overall, we found that crowd-messages were as good as expert-messages in 65% of the cases. Moreover, we found that the per-message cost for crowd-messages was about 15% less than the per-message cost of experts.

Related work

Message Tailoring

Tailored messages sent to mobile phones have been found to be effective for improving a range of health behaviors, including physical activity, medication adherence and weight loss (Wei, Hollin, and Kachnowski 2011; Kreuter et al. 2013). Health communication researchers have highlighted ways to make messages effective, such as tailoring messages in terms of gain vs. loss framing (Ledford 2012) and supporting participant autonomy (Ryan and Deci 2000).

For more than a decade, health researchers have used individually tailored interventions to support a wide range of behavior changes, including a large number focusing on physical activity (Kroeze, Werkman, and Brug 2006; Bull, Kreuter, and Scharff 1999). Tailored messages are linked to increased recipient attention, perceived message relevance and perceived message salience (Kreuter et al. 2013). Before eHealth and telemedicine, tailoring techniques were used with print-based health behavior change interventions. As the internet became popular, communicating via email became a more effective way of reaching out to recipients. This trend continued with the proliferation of mobile phones driving the researcher to use more dynamic features such as stage of change, the day of the week, and specific user preferences to tailor motivational messages that are sent to the users' mobile phones (Morandi and Serafin 2007).

More recently, tailoring has taken advantages of varied theoretical perspectives such as Self-Determination Theory as well as personality factors such as decision-making style, and communication preferences such as autonomous communication, associated with a "pull" tone, and directive communication, associated with a "push" tone (Resnicow et al. 2014). Real-time tailoring of physical activity behavior change messages is an emerging area and can leverage techniques in user modeling and machine learning (Akker, Jones, and Hermens 2014; op den Akker et al. 2015).

Tailored physical activity promotion interventions that aim to be used over the long term often require a large number of messages to support the user. Researchers may use tailoring systems (The Michigan Tailoring System (MTS) n.d.) to automatically create a large number of messages by using rule-based templates created by experts. Such rules use generic templates and conditional statements to generate a final message. Given a template such as "You want to feel more ___, maybe try walking outside," a naïve rule could then be: *If* the user's goal is "to feel more energetic" *then* fill the blank in a given template to be "energetic." However, such reliance on templates makes the messages very similar, with little variety, and hence they are likely to be perceived as monotonous. Moreover, because of their reliance on expert-generated templates, these systems can only cover a limited range of topics for which rules are set.

Crowdsourcing Systems

A crowdsourcing system, TurKit (Little et al. 2010) introduced the idea of using crowd-workers as part of algorithms such as iteratively improving the description of photos. Researchers are taking advantage of crowdsourcing as a practical platform to accomplish micro-tasks (i.e. small units of work that contribute towards a larger goal) (Cheng et al. 2015). Crowdsourcing of micro-tasks has been applied to problems in diverse crowdsourced applications, including those that require a short return time, such as copy editing (Bernstein et al. 2010), A/B testing (Kohavi et al. 2009), taxonomy creation (Chilton et al. 2013), and writing behavior change messages (de Vries et al. 2016). The prevalence of crowdsourced work structures (Difallah et al. 2015) enables teams of people to complete large tasks in piecemeal fashion and to do it at times that are convenient for the workers. Crowdsourcing systems that use micro-tasks perform better than those that use larger, more comprehensive tasks in terms of quality and ease, and they also support recovery from interruption (Cheng et al. 2015). Moreover, micro-tasks can be designed to be more efficient by limiting the amount of information involved and providing workers with actionable steps (Cheng et al. 2015).

Researchers have studied the use of crowdsourcing for writing tasks such as proofreading (Bernstein et al. 2010), writing prose (Kittur et al. 2011), story-writing (Kim, Cheng, and Bernstein 2014) and theory-based message-writing (de Vries et al. 2016). Soylent (Bernstein et al. 2010) used crowdsourced micro-tasks to accomplish common writing tasks including providing suggestions, shortening text, and proofreading. CrowdForge (Kittur et al. 2011) showed that high-quality writing can be done by workers completing tasks such as outlining, fact-gathering and simple prose writing. The authors of Ensemble (Kim, Cheng, and Bernstein 2014) found that assigning specific writing roles to workers, based on their writing skills, produces better content. MicroWriter divides the writing process into three stages of idea generation, clustering ideas and writing paragraphs from related ideas (Teevan, Iqbal, and Von Veh 2016).

Crowdsourcing systems have also been studied for the purpose of eliciting structured feedback from crowd-workers for creative tasks such as generating visual designs and improving the user experience of visual interfaces. In Voyant, Xu et al. (2014) showed that paid non-expert crowd-workers can generate structured feedback on visual designs by breaking down the feedback process into micro-tasks such as identifying first-noticed elements, sharing impressions, and judging how well the designers reached their goals and followed visual design guidelines (Xu, Huang, and Bailey 2014). Other researchers used a rubric-based approach to elicit high-quality design critiques (Yuan et al. 2016; Luther et al. 2015). Expert-generated rubrics helped novice crowd workers provide feedback on visual designs that was rated as valuable as that of experts (Yuan et al. 2016). We apply insights gained from these studies to writing short messages tailored to promote health behavior changes.

Crowdsourcing services have been used to provide empathetic reappraisal to individuals experiencing stressful thoughts or situations that could affect emotional health (Morris and Picard 2012). Providing domain knowledge understanding of a topic helped crowd-workers to generate better responses. Moreover higher-quality messages are produced by crowd-workers when they are provided with brief guidelines that include criteria for quality messages and general behavior change theory (Morris and Picard 2012). Related to health communication, researchers used TTM to guide online crowd-workers on Mechanical Turk to generate messages tailored to the stage of change (de Vries et al. 2016). Following the generation of messages, the authors conducted an evaluation with another set of crowd-workers to assess whether people perceived these messages as motivating. They found that the messages tailored to a given stage of change were in fact more motivational than messages that were not. The work suggests the feasibility of a crowdsourcing approach to generate physical activity promotion messages. In this paper, we conduct a quantitative assessment of crowd-generated messages in comparison to expert generated messages, to verify the claim that crowd-workers can generate quality messages.

System Design

At a high level, the crowdsourcing system we designed leverages crowd workers to first generate many messages, then filter the messages based on their ratings, and then fix any grammar problems in the messages (

Figure 9). The proposed approach is an iterative crowd algorithm (Little et al. 2010). Our assumption is that even if only a small percentage of worker responses are of high quality and we can identify those high quality messages, it will still be possible to build a scalable pipeline.

Crowdsourcing systems depend on the quality of crowd-worker responses. To elicit high-quality responses careful consideration must be given to the design of the task. The task's structure and content can cause drastic differences in how crowd-workers respond (Kittur 2010). We designed our crowdsourcing tasks to be brief and to encourage workers to follow guidelines based on best practices for persuasive health communication.

We chose 140 characters as the maximum length of the messages, a limit followed by users on many platforms, such as Twitter. This limit also satisfies the 160-character limit of Short Message Service (SMS) messages. Short messages also limit the duration of the crowdsourcing task. Shorter duration tasks are perceived as being more approachable and incentivize crowd-workers to keep messages concise, preventing over-zealous responses (Bernstein et al. 2010).

To better guide a crowd writer to connect messages to key aspects of a person's life, such as that person's roles and goals (Resnicow et al. 2008), we created hypothetical participants in a message-based intervention. More specifically, we started with a persona (Figure 10) that has the following content: age, picture of the person, marital status, roles in life (e.g. father, office worker, etc.), and goals in life such as spending more time with family.

We provided writing tips, as shown in Figure 10, based on key principles of Self-Determination Theory (Ryan and Deci 2000) and Health Communications Theory. First, we offered a tip to avoid overly directive messages. Overly directive messages are linked to a sense of loss of control or resistance by the reader and hence not recommended by experts. We therefore

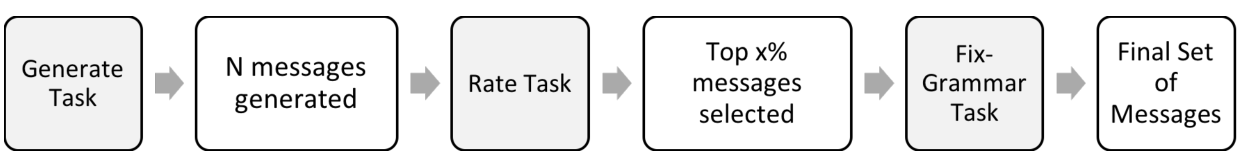

Figure 9. Steps of the crowdsourcing pipeline. Each step is performed by a unique set of crowd workers.

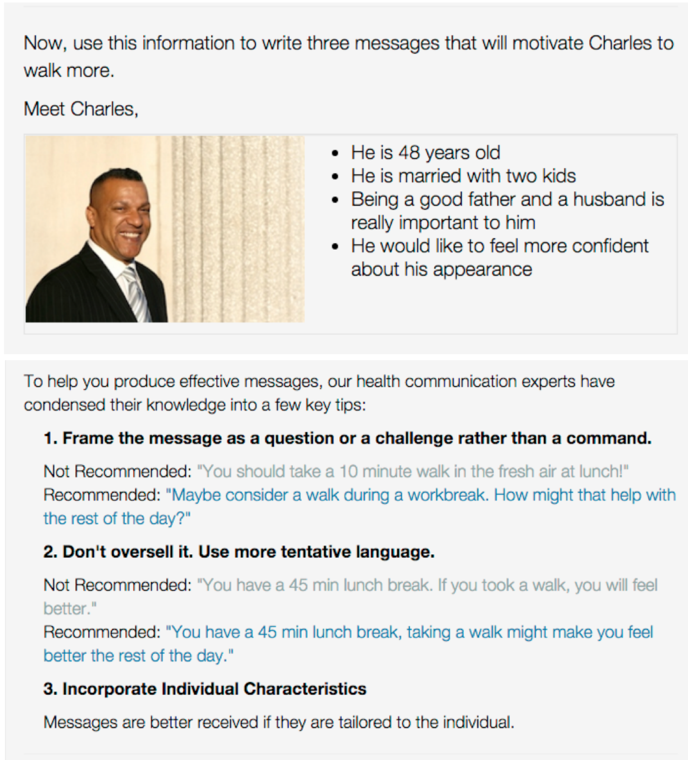

Figure 10. The message-generation task consisted of a hypothetical person created for the task and message writing tips.

included a tip about framing the message as a question or a gentle challenge, rather than as a command. Second, we counselled against overselling. Overselling is counterproductive; we therefore included a tip nudging the crowd-worker to "not oversell” and instead use tentative language. Lastly, we nudged the worker to incorporate individual characteristics such as roles and goals to write better messages (Resnicow et al. 2008). Each tip contained a short sentence such as “Try not to oversell,” accompanied by two examples, one recommended and the other not recommended.

In the following section, we describe the stages of the pipeline and how it works in practice.

Generate Task

The Generate task presents the worker with task instructions, a persona, and tips. The worker is requested to generate three messages that will motivate the persona to walk more. It also provides message-writing tips as discussed in the previous section. The system prompts workers to generate three messages per task instance at an incentive of $0.16 per message, keeping the average hourly wage at about $10 as recommended by others (Williamson 2016).

Rate Task

The Rate task shows the persona, tips, and a set of messages created in the Generate task. The Rate task instructs the worker to rate the overall quality of messages on a scale from 0 to 10, with 0 being poor and 10 being excellent. Each worker is asked to rate at most 13 messages per task, and each message is presented to at least five workers who each give independent ratings. We select the best messages based on the overall score, calculated from the mean of workers' ratings. Additionally, each message is accompanied by a checkbox that asks if the message requires grammatical cleanup.

As a quality check, 1 of the 13 messages in the rating task was selected from a set of low-quality crowd messages received from initial experiments. The writers of these messages ignored the writing instructions, making the messages of significantly lower quality. These low quality

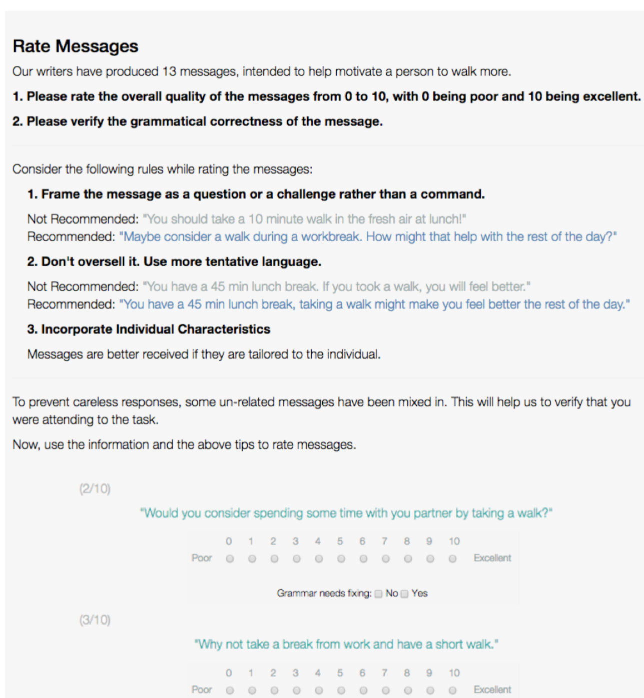

Figure 11. The Rate Task asked crowd workers to rate the overall quality of messages on a scale of 0-10. The task showed the same tips that were shown to the message writers in the Generate task. Crowd workers also marked the grammatical correctness of each message.

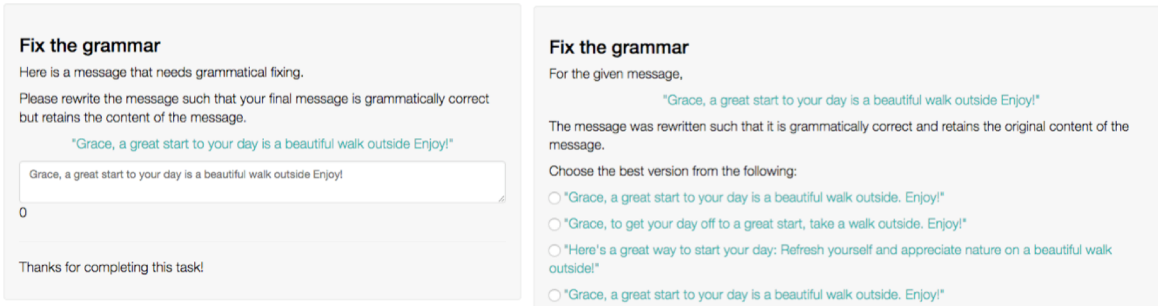

Figure 12. The two Fix-Grammar Tasks: a) fix the grammar b) choose the best amongst the given candidates.

messages were included as a way to screen for "cheat submissions," or task submissions where workers did not make a good faith effort to complete the task conscientiously (Eickhoff and Vries 2013). We discarded the "cheat" responses that rated the quality-check message incorrectly. We then assigned additional workers to make up for the lost responses.

Fix-Grammar task

From the set of top-rated messages, if a message was marked by a minimum of two workers as "Grammar needs fixing," we requested five additional workers to improve the message by fixing any grammatical errors without changing the content of the message. Subsequently, we requested another set of five workers to choose the best amongst the set of fixed messages.

Evaluation of Quality

To evaluate the quality of the crowd-generated messages, we asked experts to blindly rate the quality of crowd-generated as well as against those created by other experts (Figure 13). Moreover, we used poorly rated messages created by the crowd and non-expert messages as controls. In the following paragraphs, we describe the evaluation process. The evaluation addressed two questions:

- How does the quality of crowd-generated messages compare with expert-generated ones as judged by experts?
- How does the cost and duration of crowd-tasks compare to that of experts?

Message Generation

We used AMT to recruit crowd-workers. The crowd pipeline consisted of three tasks as described in the previous section: Generate, Rate and Grammar-Fix. The system also used automated processes that manage the beginning and completion of each task, as well as marshalling messages between stages of the pipeline.. We constructed five personas that cover multiple demographics and a range of roles and goals. For instance, one of the personas was *Charles*, a 48-year-old African-American male (Figure 10). We describe him as married with two kids—"being a good father and a husband is really important to him"—and his goal is " to feel more confident about his appearance." The other four personas varied in their demographics and included diverse roles and goals including feeling more in control, being there for friends and family, staying independent, feel more disciplined, feel more energetic, feeling strong, and being successful at work.

For the Generate step, we received responses from 49 unique workers. They wrote a total of 300 messages (60 messages for each persona). The cost of generating 60 messages was $13.40. The average time taken for each Generate Task (to generate three messages) was 220 seconds. Confirming our assumptions about worker effort, the average hourly wage for the workers was found to be $10.10. Workers also provided their demographic information and education level. They described themselves as having at least a high school education, with 47% being college graduates.

For the Rate step, we received responses from 40 unique workers. There was no overlap with workers who wrote the messages. We received a total of 1,500 message ratings, giving us five ratings for each message (n=300). The total cost of rating all the messages was $9.24 per persona. The average hourly wage for the workers was $9.60. The average time taken for each task (to rate at most 13 messages) by each worker was 263 seconds. The distribution of the ratings is presented in Figure 15.

For the expert evaluation, we selected the top 16 highest rated messages from 60 messages generated for each persona, for a total of 16*5= 80 messages. The number was chosen to match the number of messages generated by expert writers, as described below. Five messages out of the total 80 top messages for all of the five personas required grammatical fixing. These were marked as “Needs fixing” by at least two workers. From the Grammar-Fix tasks, we received five new messages for each message at the cost of $0.05 per message. For the select-best-grammar task, we showed ten workers the “fixed” messages, from which they chose the best one. The crowd-workers were paid $0.05 for each set. Finally, the system replaced the original message with fixed message selected by the largest number of workers. The total cost for this task was $3.50 for all the personas. The average hourly wage for workers was $9.86.

To summarize, the crowd-sourced message generation system takes a persona as an input and returns a ranked list of messages. The tasks completed within a few hours. In the Generate task, crowd-workers generate many messages that are then assessed on their quality in the Rate task. The system then selects the top 26% messages and leverages additional crowd-workers to check the grammatical correctness of the messages. If a message needs fixing, additional crowd-workers generate alternatives for the given messages, out of which the system selects the best one as a replacement for the original.

In this experiment, we started with five personas and ended up with 16 messages per persona. The complete process cost us about $26 per persona. Although we did not restrict crowd-workers from completing tasks in multiple stages of the pipeline, we found that each individual stage was completed by a unique set of crowd workers. Within individual stages, workers were restricted to one task for each persona. On an average, workers wrote messages for 2.04 personas and rated messages for 2.88 personas.

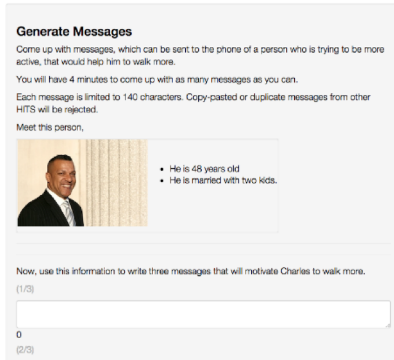

Figure 14. Expert Generation Task

Expert Message Generation

To compare crowd-generated messages with expert-generated ones, we hired three expert writers to generate messages for the same five personas that were shown to the crowd-workers. The writers were recruited from a pool of students who had completed a Master's program in Public Health and had completed a course in health communication within the last couple of years. They first attended a one-hour session, conducted by the third author, to help brush up on health communication theory relevant to writing good messages. They were then asked to write messages for the same mobile health intervention that was shown to the crowd workers and were asked to keep track of their time. They each generated five messages for each persona for a total of 15 messages per persona (total=75 messages). This resulted in 45 messages per persona, or 225 messages overall. Finally, 16 messages were randomly selected for each persona for the final set of 80 messages. The prompt for the tasks was similar to the prompt for the Generate task shown to the crowd-workers, except for the removal of tips, as shown in Figure 14. They were paid $30 per hour, as approximated from the average pay received by health communication

Writers	Evaluators
Crowd-workers	Two Experts
Experts	
Crowd-workers (Poorly rated)	
Non-experts	

Figure 13. Method of evaluation. First, we asked experts, crowd and lay people to generate messages. Then we conducted a blind evaluation with experts.

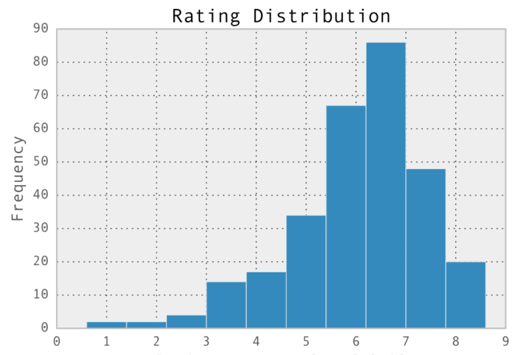

Figure 15. Overall Rating Distribution

professionals (Salary: Health Communications Specialist n.d.). The three experts spent a total of 9 hours to write the messages.

Control Messages

Although we aimed to compare high-quality lay messages to expert-generated messages, we foresaw two problems: first, if all of the messages are of high quality, a rater may tend to magnify small differences in the messages or lose sensitivity to the differences; second, if all of the messages are of similar quality, a rater may find the task to be uninteresting. Therefore, we also took a set of poorly rated messages (handpicked low-rated messages) from the crowd messages and a set of lay messages (created by students with no health communication training). Following the same task instructions as those shown to the expert writers, we asked graduate students from a university department unrelated to health or communication, and with no formal health communication training (n=5) to write three messages per persona. From this set of 15 messages per persona, we randomly selected 4 messages per persona. The prompt presented to the "lay" students did not contain health communication tips. We expected the messages created by untrained individuals who were not given health communication tips to be of poorer quality than the expert-generated messages.

To summarize, we developed four sets of messages for evaluation:

- Expert Messages: 80 expert messages (16 per persona)
- Crowd Messages: 80 high rated crowd messages (16 per persona)

- Control Messages (Crowd) Messages: 20 low-rated crowd messages (4 per persona)
- Control Messages (Non-Health Major) Messages: 20 lay messages (4 per persona)

Blind Evaluation with Experts

To compare the quality of the crowd-generated messages to that of the expert-generated messages, we conducted a blind evaluation with a different set of experts. The two expert raters were health communication practitioners, each having more than 4 years of professional experience writing behavior change messages. The expert evaluation task was divided into subtasks consisting of 20 messages each (Figure 16) for each persona contained 20 messages created for that persona, randomly selected from the sets created above. Each subtask contained 20 messages created for a single persona persona: eight crowd-generated messages, eight expert-generated messages, two control messages from the low-rated crowd-generated messages, and two control messages from the lay messages. Experts were asked to rate the overall quality of each message on a 10-point rating scale. We divided all of the messages (n=200) into 10 batches of 20 messages each. The expert evaluators were unaware of who had written the messages and how the crowd had rated the messages.

One expert rated all 200 messages and another expert rated 120 messages (for 3 out of 5 personas). These numbers were decided by the experts given the amount of time they had available. As a result, we received two expert ratings for 120 messages and one expert rating for 80 messages.

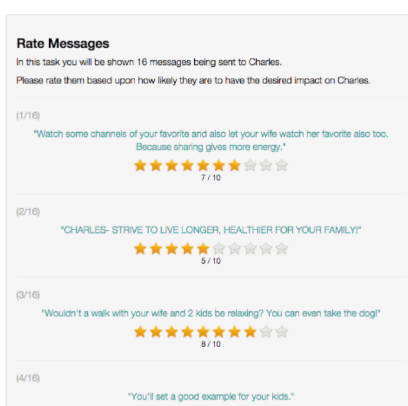

Figure 16. Expert Evaluation Task

Table 8 Agreement between Experts. The percentages represent the number of messages that are within the respective percentage of the median

METRIC	AGREEMENT
Expert Agreement (within 10%)	25%
Experts Agreement (within 20%)	43%

Results: Quality and Cost of messages

Quality of Messages

Can crowd-workers generate a sufficient number of high-quality physical activity promotion messages for a behavior change intervention program? To answer this question, we conducted two analyses: first, we compared the mean scores of expert-generated messages and crowd-generated messages as rated by the two experts; next, we compared the likelihood of producing a "good" crowd-generated message to the likelihood of producing a "good" expert-generated message at different thresholds. Table 7 provides the mean ratings of two experts for the four types of messages. The 1.06 difference in the mean scores of expert messages and crowd messages is significant (p=.001) suggesting that overall, expert-generated messages are significantly better than crowd-generated messages. We also found that both expert and crowd messages were significantly better than the control messages (p<.001). There was no significant difference between the two types of control messages.

We found that the two experts had low agreement (Table 8). The Intra-class Correlation Coefficient (ICC) (2,1) coefficient between them was 0.35. We found that only 25% of the

Table 7. Comparison between Expert, Crowd, Control-Crowd, and Control-Lay messages on the scores given by Expert-1, Expert-2 and their means

	Expert Messages	*Crowd High-Quality Messages*	*Control Messages (Crowd)*	*Control Messages (Non-Expert)*
Expert 1	8.27	7.41	4.67	4.72
Expert 2	5.71	4.44	1.83	1.45
Expert (mean)	6.99	5.93	3.25	3.09

messages had z-scores within 10% agreement and 43% within 20%. Figure 17 highlights the differences in how the four message types were rated by each expert. Such disagreements have been observed in other creative crowdsourcing literature where experts disagree over their judgment on artifact quality (Noronha, Hysen, Zhang, and Gajos, 2011). In order to deal with disagreement, researchers have followed different strategies, such as taking the union of expert critiques, evaluating items where experts do agree, or even having only one expert. Rather than avoid disagreement, we opted to analyze each expert's evaluation separately.

Given a lack of an objective measure of the quality of messages, we defined a criterion based on the assumption that at least some expert-generated messages must be of high quality and the expert judgment of expert-generated messages must be credible. Intuitively, we would like to set a threshold such that messages rated above that threshold are considered "good." We can then compare the number of "good" messages generated by the crowd-workers to those generated by the experts by setting the threshold at a given expert score on expert-generated messages. Conservatively, we could assume that only the top ten percent of expert-generated messages are good. Less conservatively, we could assume that most expert-generated messages (scored above or equal to the median) are good. In our analysis, we assess multiple candidate thresholds.

In Figure 18, we calculate the likelihood of crowd-generated messages being considered good at different thresholds. In particular, we examine the percentage of messages that are good at the

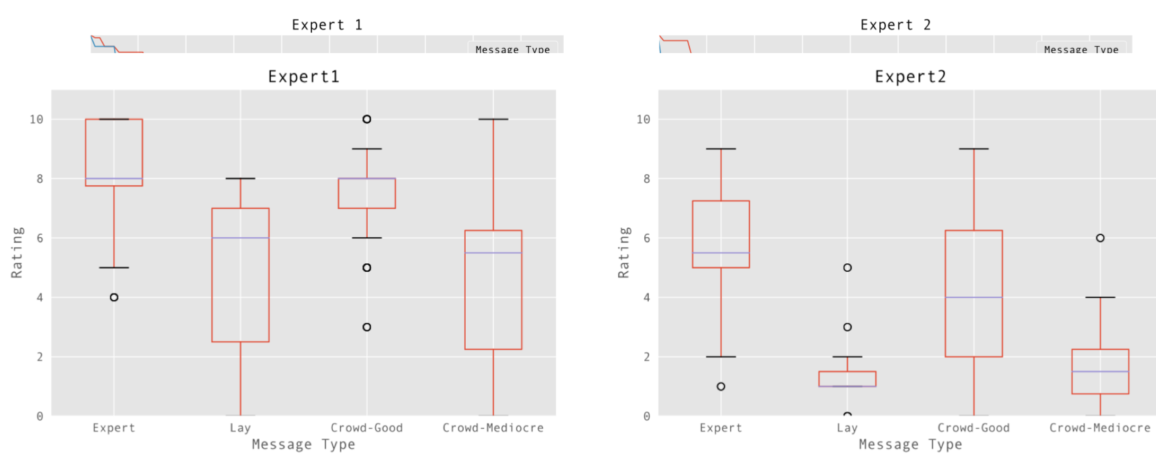

Figure 17. Box plot of the message ratings by each expert for each message type. Low ratings on some expert-generated messages suggests that not all expert messages are useful.

Table 9. Number of good messages at the median threshold and the calculated cost for writing those messages for experts and crowd in USD.

	Number of Good Expert Messages	Expert Message Cost	Number of Good Crowd Messages	Crowd Message Cost
E1	58	$3.53	35	$3.33
E2	24	$8.53	17	$6.85
Mean Cost		$6.03		$5.09

quantile score of the respective expert on expert messages. At the median threshold, when the threshold rating for being a good message is set as the median rating of expert messages, we observe that for the Expert 1, the likelihood of good crowd-generated messages and the likelihood of good expert-generated messages are .44 and .73 respectively. That is to say, according to Expert 1, 44% of all crow-generated messages and 73% of all expert-generated messages were rated above or equal to 8 (i.e. median score of Expert 1 on expert-generated messages). For Expert 2, these likelihoods are 0.35 and 0.5 respectively (for a 5.5 (using linear interpolation) median score of Expert 2 on expert-generated messages). These numbers indicate that at a relatively conservative threshold, the likelihood of generating good messages using the crowdsourcing system is between 60% and 70% of the likelihood of generating good messages using experts. Following this result, we can calculate and compare the cost of generating good messages.

Cost Analysis of Crowd vs. Experts

Cost of experts: three expert writers spent nine total hours writing their messages. We calculate the cost of experts based on a yearly salary of $52,203 (Salary: Health Communications Specialist, 2017). Adding a 22% fringe to this amount, the loaded yearly salary of an expert is $62,643. Calculated for a total of 2080 work hours in a year, the average hourly wage for an expert is then $30.15. For 9 hours, the cost of an expert would then be $271.35. Experts generated 106 messages out of which 80 messages were selected randomly. The total cost of having experts write 80 messages was (80/106)*(9*30.15) = $204.79.

Cost of crowd: For the crowd, the cost of the pipeline is equal to the sum of the cost of individual steps of the pipeline: Generate, Rate and Grammar-Fix. For the Generate step and Rate step, 300 messages were generated (220 were discarded) at a total cost of $113.20. The total cost of fixing the grammar was $3.50.

To summarize, the total cost of generating 80 messages using expert writers was $204.79, and the total cost using the crowdsourcing system was $116.5. Next, we calculate the cost of writing "good" messages, where the threshold for these messages is set at a conservative threshold of the median score for each expert. Table 9 shows the number of good messages generated by experts and by crowd-workers, as rated by Expert 1 and Expert 2 at their respective median thresholds. Table 9 also compares the per-good-message cost of expert- and crowd-generated messages, as rated by individual experts, as well as their consensus.

We observe that in all cases, the average cost of generating a good message using the crowd is lower than the cost of generating a good message using experts. The per-message cost of expert- and crowd-generated messages were $3.53 and $3.33 respectively for Expert 1, and $8.53 and $6.85 for Expert 2. The difference in the cost of messages was $0.2 and $1.68 for Expert 1 and Expert 2 respectively. The mean difference between the cost of crowd-generated messages and expert-generated messages was $0.94, which comes to 15.67% of the mean expert-generated message cost ($6).

To reiterate, the goal of this study was to leverage crowd-workers to generate quality messages for a hypothetical intervention. The results demonstrate that a crowdsourcing system can generate messages at a 15% lower per-good-message cost than a system that uses expert writers (Table 9). Thus, even though the crowdsourcing pipeline generates many more "bad" messages, it can successfully generate more "good" messages than experts could at the same cost.

Discussion

We designed and evaluated a crowdsourcing system with the goal of producing quality messages for a hypothetical behavior change intervention. The system used a process that consisted of three main steps: a Generate Task, a Rate Task, and a Fix-Grammar task. Through this system, we were able to direct online crowd-workers to generate high-quality messages for physical activity promotion at a lower cost than the estimated cost for experts.

We conducted a blind evaluation with two experts to assess the quality and cost of crowdsourced physical activity promotion messages compared to the quality and cost of expert-generated messages. Overall, we found that crowd-generated messages were as good as expert-generated messages in 65% of the cases. Moreover, we found that the per-message cost for the messages generated by the crowd was about 15% less than the per-message cost of the messages generated by experts. Moreover, experts reported feeling fatigued when asked to write many messages while the crowd does not get fatigued because it includes a large pool of writers. The results of the evaluation indicate that this system may be a promising option for the scalable generation of messages for long-running and highly tailored physical activity interventions.

Although we demonstrated the effectiveness of using crowd-workers to generate high-quality messages at a lower cost than experts, the crowdsourcing system that we designed still relies on experts' input to filter good messages from the bad ones. It is worth noting that expert-generated messages appear to require curation as well. Our findings indicate that the current system does not replace experts but reduces the burden on experts for writing quality health-promotion messages.

Further improvements could make the system even more efficient at generating high-quality messages. First, we need to better understand how to integrate crowd-workers into an expert-driven process. A possible approach would be to ask experts to leverage crowd messages and rewrite the messages for a real-world intervention. Insights from this could lead to improving the steps of the pipeline. Secondly, although our evaluation gives us an idea of the overall quality of the messages, we do not know the rationale that was used in the observed expert judgments. Moreover, we observe low agreement between the experts, indicating that individual perspectives come into play during the evaluation. We believe that different experts highlight different ways in which the messages might be deficient. In ongoing work, we hope to unpack this complexity in message evaluation. A rubric-based evaluation might allow us to disaggregate the evaluations into dimensions such as the appropriateness of tailoring for a given persona and appropriateness of tone (e.g. judgmental, overselling). Evaluating messages in terms of these dimensions may provide us with a better understanding of the characteristics of quality messages.

In addition to the quality of messages, we see other dimensions according to which messages can be evaluated. We conjecture that crowd-workers may be better at generating a diverse set of messages because of the greater number of ideas generated. Existing techniques that increase the generation of creative ideas from crowd workers may further improve the diversity (Siangliulue et al. 2015). In the future, we would like to use existing approaches to systematically evaluate the diversity of messages generated by the crowdsourcing system, varying the parameters of the system.

Finally, we have yet to compare crowd-generated messages to expert-generated messages in terms of their effectiveness at changing behavior. In ongoing research (Smith et al. 2017), we use the crowd-sourcing system designed in this study to generate messages for a mobile intervention that encourages users to walk more. The results of that research may provide further validation of our approach.

Conclusion

In this work, we presented a crowd-sourcing system that leverages domain knowledge to guide online crowd-workers to write quality messages for physical activity promotion. In a blind evaluation, two professionals evaluated the quality of messages written by experts and online crowd-workers on a 10-point scale. We found that the per-message cost of using crowd-workers was about 15% less than the per-message cost of using experts, showing that the proposed crowdsourcing system is more efficient than experts in generating physical activity promotion messages.

Chapter 6 Conclusion

In Chapter 1, I introduced a high-level architecture for a context-aware system that comprised of three steps- sensor data aggregation, context-management, and intervention generation. This dissertation contributes towards improving the latter two stages. In the following paragraphs, I will discuss the future work from each of the studies and outline some next steps for researchers to further improve our understanding of context-aware systems for physical activity promotion.

Sweet spots

The study presented in Chapter 3 provided a phenomenologically grounded account of how people make and execute plans for physical activity. The qualitative evidence supports the notion of sweet spots to be amenable to computational reasoning. Although we did not build or evaluate a computational model of sweet spots, the phenomenological grounded nature of the model increases its likelihood to support the creation and execution of planning physical activities. We believe that our approach embodies the principles of human-centered design and takes the first step towards building better computational models that depend upon an accurate representation of human behavior.

Chapter 3 presented three design implications for a system that supports creation and execution of plans for physical activity using the notion of sweet spots: a) suggesting plans based on past user behavior, b) monitor the effects of changing contextual conditions to support the execution of plans, and c) support reflection on plans that might succeed or fail. As a next step to develop these applications, we imagine future researchers to take an iterative, user-centered design and research process combining infrastructure development, application development, quantitative analysis and modeling, qualitative research, and repeated user studies. Future research may consist of two overlapping streams of work. In the technical stream, software components may predict sweet spots by estimating the follow-through likelihood of candidate plans for carrying out activities. Prototype applications may be built that use sweet spot prediction to provide planning assistance, just-in-time re-planning, and reflection on plan failures to help people better incorporate physical activity into their schedules. In the second stream, user studies can be conducted to collect data that will be needed to build and validate models, and to evaluate the prototype application. In this work, sweet spot prediction would provide benefits for certain individuals, but also identify for whom it provides benefits and why.

While our focus is on physical activity promotion, we believe that the systems that are developed could aim to support a range of different regular, time-consuming behaviors, including preparing healthy meals, learning a new skill or area of knowledge, spending time with kids or other loved ones, or performing self-care activities such as meditation or journaling.

Heed

Chapter 4 articulated the design space for Situated Self-Reporting devices and presented the design of Heed, an instantiation of SSR devices. Our findings show that Heed devices complement smartphones in the coverage of activities, locations and interaction preferences. The single-purpose and dedicated location of Heed devices were advantageous in many situations. In the future, researchers may explore other dimensions of SSR devices.

We saw a noticeably higher use of Heed devices among participants who made an effort to decrease their smartphone use in their daily life.

Our findings strengthen our claim that SSR devices are more suitable for people who are inclined to reduce their smartphone, use such as parents on behalf of their children, or elders. However, we cannot be conclusive about such claims without providing empirical evidence. Future studies exploring the use of SSR devices may be used to achieve better self-reporting behavior from specific populations.

We also note that Heed devices only hint at the potential of creating personalized self-reporting devices. We can imagine many possibilities that are worth exploring; for instance, a plate may be augmented with low-cost touch sensors (Zhang, Laput, and Harrison 2017) to help the user track their food. We imagine that creating a wide range of self-reporting tools would make it easier for ESM researchers to choose the appropriate tools based on their study needs. Such a toolkit will continue to evolve with new technologies and new opportunities to embed sensors in customized objects.

Most studies exploring the use of self-reporting tools focus on quantitative results. We recommend that researchers further delve into the qualitative data to extract nuanced user perceptions of self-reporting devices.

Crowd-generated messages

Chapter 5 shows that crowd-workers can write to write high-quality messages for a physical activity promotion intervention. At the end of the chapter, we outlined a few ways in which the quality of messages maybe further improved, in addition to improving the messages along other dimensions such as diversity of messages. We also noted that a real-world evaluation is necessary for validating a crowdsourcing approach to generate messages for applications.

A context-aware system may leverage improvements of a crowd-sourcing system. However, for the system to scale to many users and many situations, it will have to rely on automated approaches. A useful stream of research that may use NLP to generate messages in real-time for the user. For instance, a system may do this by automatically generating message templates from a corpus of crowdsourced messages. The templates may further map to the various situations that a user may find herself in. It may also leverage crowd-workers to generate more messages when necessary. For instance, the new situations may arise when the user's context changes due to factors such as changing seasons or travels.

With enough training data, advanced machine learning approaches may calculate the likelihood of a message to cause physical activity in the dynamically changing context of the user. Such opportunities also highlight the need for research that examines negative consequences from the use of such tools. For instance, personalized messages may be used to affect behaviors that may be bad for the user, either in the long-term or unknowingly in the short-term.

The goal of an intervention generator varies based on the intended application. While messages are a widely used type of intervention, some applications for physical activity promotion may require other forms of interventions. For instance, a personal informatics application may present visualizations for reflection. Or a motivational intervention may present the user with photos or videos. Future research may look at exploring the potential of using crowdsourcing to improve other forms of interventions as well.

Conclusion

The primary contribution of this dissertation is in providing guidelines for researchers and designers of systems that promote physical activity. We first present the notion of sweet spots that account for the converging contextual factors, providing a novel and useful perspective bridge the gap between phenomenological and positivist perspectives of context to provide computational support to the planning and execution of physical activities. Secondly, we present our contribution towards the design of data collection tools for gathering behavioral data from participants. This can then be used to design better context-aware applications for physical activity promotion. We designed the Heed system comprising of portable light-weight single purpose self-report devices that lower the burden of self-reporting in specific contexts. Finally, we explore the use of crowdsourcing systems to generate personalized content for context-aware applications in a scalable way. We designed a crowdsourcing system that leverages domain knowledge to write personalized behavior change messages for large-scale context-aware applications. The system can generate physical activity promotion messages in a more cost-efficient manner than experts.

Building real-world applications for physical activity promotion is challenging because of the extreme diversity in contextual conditions and user characteristics. This thesis contributes towards three important but unexplored areas: 1) building behavior models amenable to computational reasoning, 2) designing better tools to improve our understanding of human behavior, and 3) developing new applications that scale existing effective ways of achieving behavior change. We believe that improvements in these areas will support researchers and designers to build context-aware systems that are effective in promoting physical activity in their users.

References

Abowd, Gregory D., Anind K. Dey, Peter J. Brown, et al.
1999 Towards a Better Understanding of Context and Context-Awareness. *In* Handheld and Ubiquitous Computing. Hans-W. Gellersen, ed. Pp. 304–307. Lecture Notes in Computer Science, 1707. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/3-540-48157-5_29, accessed October 6, 2013.

Ackerman, Mark
2000 The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. Human-Computer Interaction 15(2): 179–203.

Agapie, Elena, Lucas Colusso, Sean A. Munson, and Gary Hsieh
2016 PlanSourcing: Generating Behavior Change Plans with Friends and Crowds. *In* Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing Pp. 119–133. CSCW '16. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2818048.2819943, accessed May 9, 2017.

Ainsworth, Barbara E., William L. Haskell, Melicia C. Whitt, et al.
2000 Compendium of Physical Activities: An Update of Activity Codes and MET Intensities: Medicine & Science in Sports & Exercise 32(Supplement): S498–S516.

Ajzen, Icek
1991 The Theory of Planned Behavior. Organizational Behavior and Human Decision Processes 50(2): 179–211.

op den Akker, Harm, Miriam Cabrita, Rieks op den Akker, Valerie M. Jones, and Hermie J. Hermens
2015 Tailored Motivational Message Generation: A Model and Practical Framework for Real-Time Physical Activity Coaching. Journal of Biomedical Informatics 55: 104–115.

Akker, Harm op den, Valerie M. Jones, and Hermie J. Hermens

2014 Tailoring Real-Time Physical Activity Coaching Systems: A Literature Survey and Model. User Modeling and User-Adapted Interaction 24(5): 351–392.

Apache Cordova
N.d. https://cordova.apache.org/, accessed November 14, 2017.

Bandura, Albert
2002 Social Foundations of Thought and Action. The Health Psychology Reader: 94–106.

Banovic, Nikola, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K. Dey
2016 Modeling and Understanding Human Routine Behavior. *In* Pp. 248–260. ACM Press. http://dl.acm.org/citation.cfm?doid=2858036.2858557, accessed September 4, 2016.

Barkhuus, Louise, and Anind Dey
2003 Is Context-Aware Computing Taking Control Away from the User? Three Levels of Interactivity Examined. *In* UbiComp Pp. 149–156. Springer.

Bauer, Jared S., Mark W. Newman, and Julie A. Kientz
2014 What Designers Talk about When They Talk about Context. Human–Computer Interaction 29(5–6): 420–450.

Bauman, Adrian E., Rodrigo S. Reis, James F. Sallis, et al.
2012 Correlates of Physical Activity: Why Are Some People Physically Active and Others Not? The Lancet 380(9838): 258–271.

Beal, Daniel J.
2015 ESM 2.0: State of the Art and Future Potential of Experience Sampling Methods in Organizational Research. Annu. Rev. Organ. Psychol. Organ. Behav. 2(1): 383–407.

Bentley, Frank, Konrad Tollmar, Peter Stephenson, et al.
2013 Health Mashups: Presenting Statistical Patterns Between Wellbeing Data and Context in Natural Language to Promote Behavior Change. ACM Trans. Comput.-Hum. Interact. 20(5): 30:1–30:27.

Berkman, Elliot T., Nicole R. Giuliani, and Alicia K. Pruitt

2014 Comparison of Text Messaging and Paper-and-Pencil for Ecological Momentary Assessment of Food Craving and Intake. Appetite 81: 131–137.

Berkman, Lisa F.
1995 The Role of Social Relations in Health Promotion. Psychosomatic Medicine 57(3): 245–254.

Bernardos, A. M., P. Tarrio, and J. R. Casar
2008 A Data Fusion Framework for Context-Aware Mobile Services. *In* 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems Pp. 606–613.

Bernstein, M. S., G. Little, R. C. Miller, et al.
2010 Soylent: A Word Processor with a Crowd Inside. *In* Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology Pp. 313–322. http://dl.acm.org/citation.cfm?id=1866078, accessed December 1, 2012.

BLE Nano
N.d. RedBear. http://redbearlab.com/blenano/, accessed November 13, 2017.

Bull, Fiona C., Matthew W. Kreuter, and Darcell P. Scharff
1999 Effects of Tailored, Personalized and General Health Messages on Physical Activity. Patient Education and Counseling 36(2): 181–192.

Burgin, Chris J., Paul J. Silvia, Kari M. Eddington, and Thomas R. Kwapil
2013 Palm or Cell? Comparing Personal Digital Assistants and Cell Phones for Experience Sampling Research. Social Science Computer Review 31(2): 244–251.

Carraro, Natasha, and Patrick Gaudreau
2013 Spontaneous and Experimentally Induced Action Planning and Coping Planning for Physical Activity: A Meta-Analysis. Psychology of Sport and Exercise 14(2): 228–248.

Chang, Yung-Ju, Gaurav Paruthi, and Mark W. Newman
2015 A Field Study Comparing Approaches to Collecting Annotated Activity Data in Real-World Settings. *In* Proceedings of the 2015 ACM International Joint Conference on Pervasive

and Ubiquitous Computing Pp. 671–682. UbiComp ’15. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2750858.2807524, accessed May 14, 2016.

Cheng, Justin, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein
2015 Break It down: A Comparison of Macro-and Microtasks. *In* Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems Pp. 4061–4064. ACM. http://dl.acm.org/citation.cfm?id=2702146, accessed April 24, 2016.

Chilton, Lydia B., Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay
2013 Cascade: Crowdsourcing Taxonomy Creation. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pp. 1999–2008. CHI ’13. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2470654.2466265, accessed July 25, 2014.

Choe, Eun Kyoung, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz
2014 Understanding Quantified-Selfers’ Practices in Collecting and Exploring Personal Data. *In* Pp. 1143–1152. ACM Press. http://dl.acm.org/citation.cfm?doid=2556288.2557372, accessed April 29, 2014.

Choi, Jaesung, Miyoung Lee, Jong-koo Lee, Daehee Kang, and Ji-Yeob Choi
2017 Correlates Associated with Participation in Physical Activity among Adults: A Systematic Review of Reviews and Update. BMC Public Health 17. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5404309/, accessed January 14, 2018.

Conner, Tamlin S., Howard Tennen, William Fleeson, and Lisa Feldman Barrett
2009 Experience Sampling Methods: A Modern Idiographic Approach to Personality Research. Social and Personality Psychology Compass 3(3): 292–313.

Consolvo, Sunny, Katherine Everitt, Ian Smith, and James A. Landay
2006 Design Requirements for Technologies That Encourage Physical Activity. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pp. 457–466. CHI ’06. New York, NY, USA: ACM. http://doi.acm.org/10.1145/1124772.1124840, accessed September 30, 2013.

Creswell, John W.
2013 Qualitative Inquiry and Research Design: Choosing among Five Approaches. Sage.

Csikszentmihalyi, Mihaly, and Reed Larson
2014 Validity and Reliability of the Experience-Sampling Method. *In* Flow and the Foundations of Positive Psychology Pp. 35–54. Springer. http://link.springer.com/10.1007/978-94-017-9088-8_3, accessed May 14, 2017.

Darker, C.D., D.P. French, F.F. Eves, and F.F. Sniehotta
2010 An Intervention to Promote Walking amongst the General Population Based on an 'Extended' Theory of Planned Behaviour: A Waiting List Randomised Controlled Trial. Psychology & Health 25(1): 71–88.

Davidoff, Scott, Brian D. Ziebart, John Zimmerman, and Anind K. Dey
2011 Learning Patterns of Pick-Ups and Drop-Offs to Support Busy Family Coordination. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pp. 1175–1184. CHI '11. New York, NY, USA: ACM. http://doi.acm.org/10.1145/1978942.1979119, accessed May 15, 2016.

Davidoff, Scott, John Zimmerman, and Anind K. Dey
2010 How Routine Learners Can Support Family Coordination. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pp. 2461–2470. ACM. http://dl.acm.org/citation.cfm?id=1753699, accessed September 15, 2016.

Deng, Jia, Olga Russakovsky, Jonathan Krause, et al.
2014 Scalable Multi-Label Annotation. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pp. 3099–3102. CHI '14. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2556288.2557011, accessed April 23, 2016.

Dey, Anind, Gregory Abowd, and Daniel Salber
2001 A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. Human-Computer Interaction 16(2): 97–166.

Dey, Anind K., Katarzyna Wac, Denzil Ferreira, et al.
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. UbiComp '11. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2030112.2030135, accessed March 23, 2017.

Difallah, Djellel Eddine, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux
2015 The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon Mturk. *In* Proceedings of the 24th International Conference on World Wide Web Pp. 238–247. International World Wide Web Conferences Steering Committee. http://dl.acm.org/citation.cfm?id=2741685.

Dombrowski, Stephan U., Falko F. Sniehotta, Alison Avenell, et al.
2012 Identifying Active Ingredients in Complex Behavioural Interventions for Obese Adults with Obesity-Related Co-Morbidities or Additional Risk Factors for Co-Morbidities: A Systematic Review. Health Psychology Review 6(1): 7–32.

Dourish, Paul
2004 What We Talk about When We Talk about Context. Personal Technologies 8(1): 19–30.

Dunton, Genevieve F., Jimi Huh, Adam M. Leventhal, et al.
2014 Momentary Assessment of Affect, Physical Feeling States, and Physical Activity in Children. Health Psychology 33(3): 255–263.

Dunton, Genevieve Fridlund, and Audie A. Atienza
2009 The Need for Time-Intensive Information in Healthful Eating and Physical Activity Research: A Timely Topic. Journal of the American Dietetic Association 109(1): 30–35.

Dunton, Genevieve Fridlund, Audie A. Atienza, Cynthia M. Castro, and Abby C. King
2009 Using Ecological Momentary Assessment to Examine Antecedents and Correlates of Physical Activity Bouts in Adults Age 50+ Years: A Pilot Study. Annals of Behavioral Medicine 38(3): 249–255.

Dunton, Genevieve Fridlund, Yue Liao, Stephen Intille, Jimi Huh, and Adam Leventhal
2015 Momentary Assessment of Contextual Influences on Affective Response during Physical Activity. Health Psychology 34(12): 1145–1153.

Dunton, Genevieve Fridlund, Carol K. Whalen, Larry D. Jamner, and Josh N. Floro
2007 Mapping the Social and Physical Contexts of Physical Activity across Adolescence Using Ecological Momentary Assessment. Annals of Behavioral Medicine 34(2): 144–153.

Eickhoff, Carsten, and Arjen P. de Vries
2013 Increasing Cheat Robustness of Crowdsourcing Tasks. Information Retrieval 16(2): 121–137.

Epstein, Daniel A., An Ping, James Fogarty, and Sean A. Munson
2015 A Lived Informatics Model of Personal Informatics. *In* Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing Pp. 731–742. UbiComp '15. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2750858.2804250, accessed May 16, 2017.

Estellés-Arolas, Enrique, and Fernando González-Ladrón-de-Guevara
2012 Towards an Integrated Crowdsourcing Definition. Journal of Information Science 38(2): 189–200.

Ferscha, Alois, Simon Vogl, and Wolfgang Beer
2001 Context Sensing, Aggregation, Representation and Exploitation in Wireless Networks. Scalable Computing: Practice and Experience 6(2).

Gartner Survey Shows Wearable Devices Need to Be More Useful
N.d. https://www.gartner.com/newsroom/id/3537117, accessed November 7, 2017.

Gauvin, Lise, W. Jack Rejeski, and Beth A. Reboussin
2000 Contributions of Acute Bouts of Vigorous Physical Activity to Explaining Diurnal Variations in Feeling States in Active, Middle-Aged Women. Health Psychology 19(4): 365.

Goasduff, Laurence, and Christy Pettey
2012 Gartner Says Worldwide Smartphone Sales Soared in Fourth Quarter of 2011 with 47 Percent Growth. Gartner, Inc 15. http://www.pressebox.com/attachment/447933/Mobile+Devices+4Q11+FINAL+EMEA+version.pdf, accessed April 24, 2016.

Gollwitzer, Peter M.
1999 Implementation Intentions: Strong Effects of Simple Plans. American Psychologist 54(7): 493.

Gollwitzer, Peter M., and Paschal Sheeran
2006 Implementation Intentions and Goal Achievement: A Meta-Analysis of Effects and Processes. Advances in Experimental Social Psychology 38: 69–119.

Google
2016 Find Time for Your Goals with Google Calendar. http://blog.google:443/products/calendar/find-time-goals-google-calendar/, accessed May 15, 2017.

Grubert, Jens, Matthias Kranz, and Aaron Quigley
2016 Challenges in Mobile Multi-Device Ecosystems. MUX: The Journal of Mobile User Experience 5(1). http://arxiv.org/abs/1605.07760, accessed November 14, 2017.

Guba, E. G., and Y. S. Lincoln
1985 Naturalistic Inquiry (Vol. 75). Beverly Hills, CA: Sage.

Hagger, Martin S., Nikos L.D. Chatzisarantis, and Stuart J.H. Biddle
2002 A Meta-Analytic Review of the Theories of Reasoned Action and Planned Behavior in Physical Activity: Predictive Validity and the Contribution of Additional Variables. Journal of Sport and Exercise Psychology 24(1): 3–32.

Hara, Kotaro, Shiri Azenkot, Megan Campbell, et al.
2015 Improving Public Transit Accessibility for Blind Riders by Crowdsourcing Bus Stop Landmark Locations with Google Street View: An Extended Analysis. ACM Transactions on Accessible Computing (TACCESS) 6(2): 5.

Hara, Kotaro, Vicki Le, and Jon Froehlich
2013 Combining Crowdsourcing and Google Street View to Identify Street-Level Accessibility Problems. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pp. 631–640. ACM. http://dl.acm.org/citation.cfm?id=2470744, accessed April 23, 2016.

Hekler, Eric B., Matthew P. Buman, David Ahn, et al.
2012 Are Daily Fluctuations in Perceived Environment Associated with Walking? Psychology & Health 27(9): 1009–1020.

Hekler, Eric B., Predrag Klasnja, William T. Riley, et al.
2016 Agile Science: Creating Useful Products for Behavior Change in the Real World. Translational Behavioral Medicine 6(2): 317–328.

Hekler, Eric B., Susan Michie, Misha Pavel, et al.
2016 Advancing Models and Theories for Digital Behavior Change Interventions. American Journal of Preventive Medicine 51(5): 825–832.

Hektner, Joel M., Jennifer A. Schmidt, and Mihaly Csikszentmihalyi
2007 Experience Sampling Method: Measuring the Quality of Everyday Life. Sage.

Hernandez, Javier, Daniel McDuff, Christian Infante, et al.
2016 Wearable ESM: Differences in the Experience Sampling Method Across Wearable Devices. *In* Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Pp. 195–205. MobileHCI '16. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2935334.2935340, accessed December 18, 2016.

Herrmann, Lynn Katherine, and Jinsook Kim
2017 The Fitness of Apps: A Theory-Based Examination of Mobile Fitness App Usage over 5 Months. MHealth 3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5344171/, accessed January 4, 2018.

Hoffman, C., D. Rice, and H. Y. Sung
1996 Persons with Chronic Conditions. Their Prevalence and Costs. JAMA 276(18): 1473–1479.

Hong, Hwajung, Svetlana Yarosh, Jennifer G. Kim, Gregory D. Abowd, and Rosa I. Arriaga
2013 Investigating the Use of Circles in Social Networks to Support Independence of Individuals with Autism. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pp. 3207–3216. ACM. http://dl.acm.org/citation.cfm?id=2466439, accessed April 23, 2016.

Huberty, Jennifer L., Diane Ehlers, Jason Coleman, Yong Gao, and Steriani Elavsky
2013 Women Bound to Be Active: Differences in Long-Term Physical Activity between Completers and Noncompleters of a Book Club Intervention. Journal of Physical Activity and

Health. http://scholarworks.boisestate.edu/kinesiology_facpubs/111/, accessed September 14, 2016.

Humpel, Nancy, Neville Owen, and Eva Leslie
2002 Environmental Factors Associated with Adults' Participation in Physical Activity: A Review. American Journal of Preventive Medicine 22(3): 188–199.

Hurling, Robert, Michael Catt, Marco De Boni, et al.
2007 Using Internet and Mobile Phone Technology to Deliver an Automated Physical Activity Program: Randomized Controlled Trial. Journal of Medical Internet Research 9(2): e7.

Internet of Things Consumer Devices: Ownership and Preferences
N.d. I-SCOOP. https://www.i-scoop.eu/internet-of-things-guide/iot-consumer-device-ownership-preferences/, accessed November 6, 2017.

Intille, Stephen, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides
2016 μEMA: Microinteraction-Based Ecological Momentary Assessment (EMA) Using a Smartwatch. *In* Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Pp. 1124–1128. UbiComp '16. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2971648.2971717.

Intille, Stephen S.
2004 Ubiquitous Computing Technology for Just-in-Time Motivation of Behavior Change. Stud Health Technol Inform 107(Pt 2): 1434–1437.

Kim, Joy, Justin Cheng, and Michael S. Bernstein
2014 Ensemble: Exploring Complementary Strengths of Leaders and Crowds in Creative Collaboration. *In* Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing Pp. 745–755. ACM. http://dl.acm.org/citation.cfm?id=2531638, accessed April 24, 2016.

Kittur, Aniket
2010 Crowdsourcing, Collaboration and Creativity. XRDS: Crossroads, The ACM Magazine for Students 17(2): 22.

Kittur, Aniket, Boris Smus, Susheel Khamkar, and Robert E. Kraut

2011 Crowdforge: Crowdsourcing Complex Work. *In* Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology Pp. 43–52. ACM. http://dl.acm.org/citation.cfm?id=2047202, accessed July 6, 2014.

Klasnja, Predrag, and Wanda Pratt
2012 Healthcare in the Pocket: Mapping the Space of Mobile-Phone Health Interventions. Journal of Biomedical Informatics 45(1): 184–198.

Kohavi, Ron, Roger Longbotham, Dan Sommerfield, and Randal M. Henne
2009 Controlled Experiments on the Web: Survey and Practical Guide. Data Mining and Knowledge Discovery 18(1): 140–181.

Kreuter, Matthew W., David W. Farrell, Laura R. Olevitch, and Laura K. Brennan
2013 Tailoring Health Messages: Customizing Communication With Computer Technology. Routledge.

Kroeze, Willemieke, Andrea Werkman, and Johannes Brug
2006 A Systematic Review of Randomized Trials on the Effectiveness of Computer-Tailored Education on Physical Activity and Dietary Behaviors. Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine 31(3): 205–223.

Kummerfeld, Bob, Lie Ming Tang, Judy Kay, and Farahnaz Yekeh
2015 SAL: A Small, Simple, Situated, Ambient Logger. *In* Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers Pp. 403–406. UbiComp/ISWC'15 Adjunct. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2800835.2800929.

Kwasnicka, Dominika, Justin Presseau, Martin White, and Falko F. Sniehotta
2013 Does Planning How to Cope with Anticipated Barriers Facilitate Health-Related Behaviour Change? A Systematic Review. Health Psychology Review 7(2): 129–145.

Lamiell, James T.
2003 Beyond Individual and Group Differences: Human Individuality, Scientific Psychology, and William Stern's Critical Personalism. Sage.

Latham, Gary P, and Edwin A Locke
1991 Self-Regulation through Goal Setting. Organizational Behavior and Human Decision Processes 50(2). Theories of Cognitive Self-Regulation: 212–247.

Lathia, Neal, Kiran K. Rachuri, Cecilia Mascolo, and Peter J. Rentfrow
2013 Contextual Dissonance: Design Bias in Sensor-Based Experience Sampling Methods. *In* Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing Pp. 183–192. UbiComp '13. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2493432.2493452, accessed April 10, 2017.

Ledford, Christy J. W.
2012 Exploring the Interaction of Patient Activation and Message Design Variables: Message Frame and Presentation Mode Influence on the Walking Behavior of Patients with Type 2 Diabetes. Journal of Health Psychology 17(7): 989–1000.

Ledger, Dan, and Daniel McCaffrey
2014 Inside Wearables: How the Science of Human Behavior Change Offers the Secret to Long-Term Engagement. Endeavour Partners.

Lewis, Sophie, Samantha L. Thomas, R. Warwick Blood, et al.
2011 "I'm Searching for Solutions": Why Are Obese Individuals Turning to the Internet for Help and Support with "Being Fat"? Health Expectations: An International Journal of Public Participation in Health Care and Health Policy 14(4): 339–350.

Li, Ian, Anind K. Dey, and Jodi Forlizzi
2012 Using Context to Reveal Factors That Affect Physical Activity. ACM Trans. Comput.-Hum. Interact. 19(1): 7:1–7:21.

Liao, Yue, Eleanor T. Shonkoff, and Genevieve F. Dunton
2015 The Acute Relationships Between Affect, Physical Feeling States, and Physical Activity in Daily Life: A Review of Current Evidence. Frontiers in Psychology 6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4688389/, accessed April 19, 2016.

Little, Greg, Lydia B. Chilton, Max Goldman, and Robert C. Miller
2010 Turkit: Human Computation Algorithms on Mechanical Turk. *In* Proceedings of the 23nd

Annual ACM Symposium on User Interface Software and Technology Pp. 57–66. ACM. http://dl.acm.org.proxy.lib.umich.edu/citation.cfm?id=1866040, accessed July 7, 2014.

Lopez, Kay A., and Danny G. Willis
2004 Descriptive versus Interpretive Phenomenology: Their Contributions to Nursing Knowledge. Qualitative Health Research 14(5): 726–735.

Luszczynska, Aleksandra
2006 An Implementation Intentions Intervention, the Use of a Planning Strategy, and Physical Activity after Myocardial Infarction. Social Science & Medicine 62(4): 900–908.

Luther, Kurt, Jari-Lee Tolentino, Wei Wu, et al.
2015 Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. *In* Pp. 473–485. ACM Press. http://dl.acm.org/citation.cfm?doid=2675133.2675283, accessed March 28, 2015.

Mailey, Emily L., and Edward McAuley
2014 Impact of a Brief Intervention on Physical Activity and Social Cognitive Determinants among Working Mothers: A Randomized Trial. Journal of Behavioral Medicine 37(2): 343–355.

Martin, C.A., D.E. Rivera, W.T. Riley, et al.
2014 A Dynamical Systems Model of Social Cognitive Theory. *In* American Control Conference (ACC), 2014 Pp. 2407–2412.

Mathur, Akhil, Nicholas D. Lane, and Fahim Kawsar
2016 Engagement-Aware Computing: Modelling User Engagement from Mobile Contexts. *In* Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Pp. 622–633. UbiComp '16. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2971648.2971760, accessed March 23, 2017.

May 2016 National Occupational Employment and Wage Estimates
N.d. https://www.bls.gov/oes/current/oes_nat.htm#00-0000, accessed May 10, 2017.

Milne, Sarah, Sheina Orbell, and Paschal Sheeran
2002 Combining Motivational and Volitional Interventions to Promote Exercise Participation:

Protection Motivation Theory and Implementation Intentions. British Journal of Health Psychology 7(2): 163–184.

Min, Chulhong, Seungwoo Kang, Chungkuk Yoo, et al.
2015 Exploring Current Practices for Battery Use and Management of Smartwatches. *In* Proceedings of the 2015 ACM International Symposium on Wearable Computers Pp. 11–18. ACM.

Mistry, Chetan D., Shane N. Sweet, Ryan E. Rhodes, and Amy E. Latimer-Cheung
2015 Text2Plan: Exploring Changes in the Quantity and Quality of Action Plans and Physical Activity in a Text Messaging Intervention. Psychology & Health 30(7): 839–856.

Molenaar, Peter CM
2004 A Manifesto on Psychology as Idiographic Science: Bringing the Person Back into Scientific Psychology, This Time Forever. Measurement 2(4): 201–218.

Morandi, Angelica, and Riccardo Serafin
2007 A Personalized Motivation Strategy for Physical Activity Promotion in Diabetic Subjects. *In* Proceedings of the 2nd Workshop on Personalisation for E-Health Pp. 51–55. http://arion.csd.uwo.ca/courses/CS9626b/papers_files/personalized_health.pdf, accessed February 8, 2016.

Morris, Robert R., and Rosalind Picard
2012 Crowdsourcing Collective Emotional Intelligence. ArXiv Preprint ArXiv:1204.3481. http://arxiv.org/abs/1204.3481, accessed July 6, 2014.

Ng, Andrew Y., and Stuart J. Russell
2000 Algorithms for Inverse Reinforcement Learning. *In* Icml Pp. 663–670.

Odom, William, James Pierce, Erik Stolterman, and Eli Blevis
2009 Understanding Why We Preserve Some Things and Discard Others in the Context of Interaction Design. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Pp. 1053–1062. ACM.

Overcoming Barriers to Physical Activity | Basics | Physical Activity | DNPAO | CDC

N.d. http://www.cdc.gov/physicalactivity/basics/adding-pa/barriers.html, accessed November 20, 2015.

Paay, Jeni, Jesper Kjeldskov, Mikael B. Skov, Lars Lichon, and Stephan Rasmussen
2015 Understanding Individual Differences for Tailored Smoking Cessation Apps. *In* Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems Pp. 1699–1708. CHI '15. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2702123.2702321, accessed May 10, 2017.

Palen, Leysia, and Marilyn Salzman
2002 Voice-Mail Diary Studies for Naturalistic Data Capture under Mobile Conditions. *In* Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work Pp. 87–95. ACM.

Patrick, Kevin, William G. Griswold, Fred Raab, and Stephen S. Intille
2008 Health and the Mobile Phone. American Journal of Preventive Medicine 35(2): 177–181.

Patrick, Kevin, Fred Raab, Marc A Adams, et al.
2009 A Text Message–Based Intervention for Weight Loss: Randomized Controlled Trial. Journal of Medical Internet Research 11(1). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2729073/.

Payne, Nicola, Fiona Jones, and Peter R. Harris
2010 A Daily Diary Investigation of the Impact of Work Stress on Exercise Intention Realisation: Can Planning Overcome the Disruptive Influence of Work? Psychology & Health 25(1): 111–129.

Perera, C., A. Zaslavsky, P. Christen, and D. Georgakopoulos
2014 Context Aware Computing for The Internet of Things: A Survey. IEEE Communications Surveys Tutorials 16(1): 414–454.

Pielot, Martin, Tilman Dingler, Jose San Pedro, and Nuria Oliver
2015 When Attention Is Not Scarce-Detecting Boredom from Mobile Phone Usage. *In* Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing Pp. 825–836. ACM.

Ponnada, Aditya, Caitlin Haynes, Dharam Maniar, Justin Manjourides, and Stephen Intille
2017 Microinteraction Ecological Momentary Assessment Response Rates: Effect of Microinteractions or the Smartwatch? Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1(3): 92:1–92:16.

Prestwich, Andrew, Rebecca Lawton, and Mark Conner
2003 The Use of Implementation Intentions and the Decision Balance Sheet in Promoting Exercise Behaviour. Psychology and Health 18(6): 707–721.

Prestwich, Andrew, Marco Perugini, and Robert Hurling
2010 Can Implementation Intentions and Text Messages Promote Brisk Walking? A Randomized Trial. Health Psychology 29(1): 40–49.

Prochaska, James O., and Wayne F. Velicer
1997 The Transtheoretical Model of Health Behavior Change. American Journal of Health Promotion 12(1): 38–48.

Rabbi, Mashfiqui, Angela Pfammatter, Mi Zhang, Bonnie Spring, and Tanzeem Choudhury
2015 Automated Personalized Feedback for Physical Activity and Dietary Behavior Change With Mobile Phones: A Randomized Controlled Trial on Adults. JMIR MHealth and UHealth 3(2): e42.

Resnicow, Ken, Rachel E. Davis, Guangyu Zhang, et al.
2008 Tailoring a Fruit and Vegetable Intervention on Novel Motivational Constructs: Results of a Randomized Study. Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine 35(2): 159–169.

Resnicow, Ken, Yan Zhou, Sarah Hawley, et al.
2014 Communication Preference Moderates the Effect of a Tailored Intervention to Increase Colorectal Cancer Screening among African Americans. Patient Education and Counseling 97(3): 370–375.

Riley, William T., Daniel E. Rivera, Audie A. Atienza, et al.
2011 Health Behavior Models in the Age of Mobile Interventions: Are Our Theories up to the Task? Translational Behavioral Medicine 1(1): 53–71.

Rosal, Milagros C, Cara B Ebbeling, Ingrid Lofgren, et al.
2001 Facilitating Dietary Change: The Patient-Centered Counseling Model. Journal of the American Dietetic Association 101(3): 332–341.

Ryan, Richard M., and Edward L. Deci
2000 Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. American Psychologist 55(1): 68.

Salary: Health Communications Specialist
N.d. Glassdoor. https://www.glassdoor.com/Salaries/health-communications-specialist-salary-SRCH_KO0,32.htm, accessed October 13, 2016.

Salary: Health Communications Specialist 2017
N.d. Glassdoor. https://www.glassdoor.com/Salaries/health-communications-specialist-salary-SRCH_KO0,32.htm, accessed September 26, 2017.

Saldaña, Johnny
2015 The Coding Manual for Qualitative Researchers. Sage.

Salvy, Sarah-Jeanne, Julie Wojslawowicz Bowker, James N. Roemmich, et al.
2007 Peer Influence on Children's Physical Activity: An Experience Sampling Study. Journal of Pediatric Psychology 33(1): 39–49.

Salvy, Sarah-Jeanne, James N. Roemmich, Julie C. Bowker, et al.
2008 Effect of Peers and Friends on Youth Physical Activity and Motivation to Be Physically Active. Journal of Pediatric Psychology 34(2): 217–225.

Saperstein, S. L., N. L. Atkinson, and R. S. Gold
2007 The Impact of Internet Use for Weight Loss. Obesity Reviews: An Official Journal of the International Association for the Study of Obesity 8(5): 459–465.

Schilit, Bill, Norman Adams, and Roy Want
1994 Context-Aware Computing Applications. *In* Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop On Pp. 85–90. IEEE.

Sheeran, Paschal
2002 Intention—behavior Relations: A Conceptual and Empirical Review. European Review of Social Psychology 12(1): 1–36.

Shiffman, Saul, Arthur A. Stone, and Michael R. Hufford
2008 Ecological Momentary Assessment. Annu. Rev. Clin. Psychol. 4: 1–32.

Shosha, Ghada Abu
2012 Employment of Colaizzi's Strategy in Descriptive Phenomenology: A Reflection of a Researcher. European Scientific Journal, ESJ 8(27).

Siangliulue, Pao, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow
2015 Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. *In* Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing Pp. 937–945. CSCW '15. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2675133.2675239, accessed March 29, 2015.

Smart Wearables Vendor Strategies, Opportunities, Forecasts
N.d. https://www.juniperresearch.com/researchstore/smart-devices/smart-wearables/vendor-strategies-opportunities-forecasts, accessed November 6, 2017.

Smith, Shawna N., Andy Jinseok Lee, Kelly Hall, et al.
2017 Design Lessons from a Micro-Randomized Pilot Study in Mobile Health. *In* Mobile Health Pp. 59–82. Springer, Cham. https://link-springer-com.proxy.lib.umich.edu/chapter/10.1007/978-3-319-51394-2_4, accessed August 24, 2017.

Spruijt-Metz, D., and W. Nilsen
2014 Dynamic Models of Behavior for Just-in-Time Adaptive Interventions. IEEE Pervasive Computing 13(3): 13–17.

Spruijt-Metz, Donna, Eric Hekler, Niilo Saranummi, et al.
2015 Building New Computational Models to Support Health Behavior Change and Maintenance: New Opportunities in Behavioral Research. Translational Behavioral Medicine 5(3): 335–346.

Sundmaeker, Harald, Patrick Guillemin, Peter Friess, and Sylvie Woelfflé
2010 Vision and Challenges for Realising the Internet of Things. Cluster of European Research Projects on the Internet of Things, European Commision 3(3): 34–36.

Sweet, Shane N., Lawrence R. Brawley, Alexandra Hatchell, Heather L. Gainforth, and Amy E. Latimer-Cheung
2014 Can Persuasive Messages Encourage Individuals to Create Action Plans for Physical Activity? Journal of Sport & Exercise Psychology 36(4).

Teevan, Jaime, Shamsi T. Iqbal, and Curtis Von Veh
2016 Supporting Collaborative Writing with Microtasks. *In* Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems Pp. 2657–2668. ACM.

The Michigan Tailoring System (MTS)
N.d. http://chcr.umich.edu/mts/, accessed April 23, 2016.

Troiano, Richard P., David Berrigan, Kevin W. Dodd, et al.
2008 Physical Activity in the United States Measured by Accelerometer. Medicine and Science in Sports and Exercise 40(1): 181.

Trost, Stewart G., Neville Owen, Adrian E. Bauman, James F. Sallis, and Wendy Brown
2002 Correlates of Adults' Participation in Physical Activity: Review and Update. Medicine & Science in Sports & Exercise. http://psycnet.apa.org/psycinfo/2002-11351-004, accessed April 24, 2016.

Truong, Khai N., Thariq Shihipar, and Daniel J. Wigdor
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. http://doi.acm.org/10.1145/2556288.2557044, accessed March 14, 2017.

de Vet, Emely, Anke Oenema, and Johannes Brug
2011 More or Better: Do the Number and Specificity of Implementation Intentions Matter in Increasing Physical Activity? Psychology of Sport and Exercise 12(4): 471–477.

Vet, Emely De, Anke Oenema, Paschal Sheeran, and Johannes Brug
2009 Should Implementation Intentions Interventions Be Implemented in Obesity Prevention: The Impact of If-Then Plans on Daily Physical Activity in Dutch Adults. International Journal of Behavioral Nutrition and Physical Activity 6(1): 11.

de Vries, Roelof A.J., Khiet P. Truong, Sigrid Kwint, Constance H.C. Drossaert, and Vanessa Evers
2016 Crowd-Designed Motivation: Motivational Messages for Exercise Adherence Based on Behavior Change Theory. *In* Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems Pp. 297–308. CHI ’16. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2858036.2858229, accessed May 8, 2016.

Wei, Jin, Ilene Hollin, and Stan Kachnowski
2011 A Review of the Use of Mobile Phone Text Messaging in Clinical and Healthy Behaviour Interventions. Journal of Telemedicine and Telecare 17(1): 41–48.

Weiser, M., and J. Seely Brown
N.d. The Coming Age of Calm Technology", Xerox PARC October 5, 1996.

Weiser, Mark
1991 The Computer for the 21st Century. Scientific American 265(3): 94–104.

Wen, Cheng K Fred, Stefan Schneider, Arthur A Stone, and Donna Spruijt-Metz
2017 Compliance With Mobile Ecological Momentary Assessment Protocols in Children and Adolescents: A Systematic Review and Meta-Analysis. Journal of Medical Internet Research 19(4). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5425774/, accessed November 15, 2017.

Wikipedia
2017a Affinity Diagram. https://en.wikipedia.org/w/index.php?title=Affinity_diagram&oldid=772829913, accessed November 12, 2017.

2017b Automated Planning and Scheduling. https://en.wikipedia.org/w/index.php?title=Automated_planning_and_scheduling&oldid=779693832, accessed May 13, 2017.

Wild-Hartmann, Jessica A. de, Marieke Wichers, Alex L. van Bemmel, et al.
2013 Day-to-Day Associations between Subjective Sleep and Affect in Regard to Future Depression in a Female Population-Based Sample. The British Journal of Psychiatry 202(6): 407–412.

Williamson, Vanessa
2016 On the Ethics of Crowdsourced Research. PS: Political Science & Politics 49(1): 77–81.

Xu, Anbang, Shih-Wen Huang, and Brian Bailey
2014 Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. *In* Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing Pp. 1433–1444. ACM. http://dl.acm.org/citation.cfm?id=2531604, accessed April 13, 2016.

Yardley, Lucy, Tanzeem Choudhury, Kevin Patrick, and Susan Michie
2016 Current Issues and Future Directions for Research into Digital Behavior Change Interventions. American Journal of Preventive Medicine 51(5): 814–815.

Yekeh, Farahnaz, Judy Kay, Bob Kummerfeld, Lie Ming Tang, and Margaret A. Allman-Farinelli
2015 Can SAL Support Self Reflection for Health and Nutrition? *In* Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction Pp. 134–141. OzCHI '15. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2838739.2838752.

Yuan, Alvin, Kurt Luther, Markus Krause, et al.
2016 Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. *In* Pp. 1003–1015. ACM Press. http://dl.acm.org/citation.cfm?doid=2818048.2819953, accessed April 13, 2016.

Zaslavsky, Arkady, Charith Perera, and Dimitrios Georgakopoulos
2013 Sensing as a Service and Big Data. ArXiv Preprint ArXiv:1301.0159.

Zhang, Xiaoyi, Laura R. Pina, and James Fogarty
2016 Examining Unlock Journaling with Diaries and Reminders for In Situ Self-Report in Health and Wellness. *In* Proceedings of the 2016 CHI Conference on Human Factors in

Computing Systems Pp. 5658–5664. CHI ’16. New York, NY, USA: ACM.
http://doi.acm.org/10.1145/2858036.2858360, accessed March 14, 2017.

Ziegelmann, Jochen P., Sonia Lippke, and Ralf Schwarzer
2006 Adoption and Maintenance of Physical Activity: Planning Interventions in Young, Middle-Aged, and Older Adults. Psychology & Health 21(2): 145–163.