Abstract of “Personalized Systems for Guided and Flexible Self-Experiments” by Nediyana Daskalova, Ph.D., Brown University, May 2020.

The current paradigm in health tracking research, as performed in fields such as public health, social sciences, and research initiatives like mHealth, is to find generalizable effects that can be disseminated to the public. However, by definition, there only has to be a small effect on a subset of that population for those studies to claim a positive result. In order to avoid following general advice and to find out what works specifically for them as individuals, some people perform experiments on themselves (self-experiments). However, in reality the insights people draw are often flawed because not everyone is trained to conduct scientifically valid experiments.

The aim of this work is to study how individuals perform self-experiments and to build novel systems that guide people through the steps of such experiments. This dissertation outlines tools and guidelines to make self-experimentation accessible by providing easy to understand interventions and results. The first system presented is SleepCoacher: an automated system for self-experiments in sleep that includes a sleep tracking smartphone application that collects data from sensors and user input to provide and evaluate the effect of actionable personalized recommendations for improving sleep. Next, based on findings from studies with the system, the SleepCoacher approach is modified to add more guidance, flexibility, and agency in the self-experimentation process to create a robust app called SleepBandits. In SleepBandits, Thompson Sampling, a heuristic from reinforcement learning, estimates how likely it is that the behavior change is helpful for the given user. Finally, extending this line of research, Self-E is developed as a system for broader self-experiments, which lets the user choose which behaviors to change and automatically breaks down the experiment into a series of steps, and communicates them to the user through actionable messages. Together, these systems are a step towards a vision of perpetual self-experiments, where users can continuously receive recommendations and change little snippets of their behavior to constantly improve their well-being.

Personalized Systems for Guided and Flexible Self-Experiments

by
Nediyana Daskalova
B. A., Computer Science, Grinnell College, 2014
Sc. M., Brown University, 2016

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

Providence, Rhode Island
May 2020

This dissertation by Nediyana Daskalova is accepted in its present form by
the Department of Computer Science as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date _____ _____

Jeff Huang, Director

Recommended to the Graduate Council

Date _____ _____

Nicole Nugent, Reader
(Warren Alpert Medical School)

Date _____ _____

Michael Littman, Reader

Approved by the Graduate Council

Date _____ _____

Andrew G. Campbell
Dean of the Graduate School

Vita

Nediyana Daskalova was born and raised in Sofia, Bulgaria. She received a B.A. in Computer Science from Grinnell College, Iowa. In 2014, she joined the Computer Science Department at Brown University in Providence, Rhode Island for her doctorate degree. She earned a Sc.M. in 2016 and completed her Ph.D. under the advisement of professor Jeff Huang. During her time as a Ph.D student, Nediyana completed three internships: at Yahoo in Sunnyvale, CA (2016), at Microsoft Research in Redmond, WA (2017), and at Instagram in New York, NY (2018). While at Brown, she founded a mentorship program for PhD students and participated in other initiatives to promote community-building in the computer science department.

Acknowledgements

First and foremost, I would like to thank my advisor, Jeff Huang, for the endless support and guidance throughout my whole PhD and every aspect of my life around it. His care and commitment to his research, classes, and students has been my inspiration for striving to be better not only as a scientist but as a person. He has always encouraged us to work with undergraduate and masters students, and has created a supportive and friendly environment in our lab.

I am also incredibly grateful to Nicole Nugent, who has not only served on my committee since my first year, but has also been instrumental in the development of all of our systems and has been a co-author on our papers. Her advice and eternal enthusiasm for the work kept me motivated through the ups and downs of the PhD. I would also like to thank Michael Littman and Ellie Pavlick, who offered invaluable guidance and advice to make this a better dissertation.

This dissertation would not have been possible without all the amazing students, mentors, and collaborators I have been lucky to work with over the years: Cintia Araujo, Guillermo Beltran Jr, Frank Bentley, Julie Boergers, Joseph Jay Williams, Niharika Jhingan, Yusuf Karim, Eindra Kyi, Bongshin Lee, Jessica Lundin, John McGeary, Danae Metaxa-Kakavouli, Chester Ni, Kevin Ouyang, Andrew Park, Adrienne Tran, Lisa Wang, and Jina Yoon.

Our department has been very supportive overall, but big thanks especially to tstaff and astaff, particularly Dawn Reed, Donald Johwa, Eugenia DeGouveia, and Lauren Clarke, for caring for us with all their hearts.

I am also grateful to the community of graduate students, whose moral and social support has been even more powerful than their feedback on numerous presentations and research questions. The list is endless, but I would like to especially thank Alexandra and Mira. And most importantly, thank you to my family for the encouragement and support: Alejandro, Assya, Yagodinka, and Vesselin.

This dissertation was supported by the Brown University Seed Award, National Science Foundation IIS-1656763, the Brown University Data Science Institute, and the Brown University Catalyst Grant.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Thesis Statement: Self-experiments are difficult for people to conduct themselves, due to tracking fatigue, lack of randomization in the experimental conditions, and flawed data analysis. Guidance from automated mobile apps can improve understanding and introduce randomization in their experiments by balancing agency with scientific validity. Several iterations of this framework lead to systems that guide the user through the steps of collecting data and evaluating the effectiveness of personalized recommendations, resulting in a lowered barrier to the wider adoption of self-experimentation.

1.1 Motivation

"To find out what happens when you change something, it is necessary to change it." —- George Box

Data derived from male subjects for diagnoses and treatments has long been considered one-size-fits-all in biomedical research [91]. However, generalized health advice based on that data might not be applicable to each individual's specific outcomes and could lead to negative reactions. In contrast, self-experiments, a form of single-case studies, focus on finding outcomes that work specifically for a given person, rather than on finding generalizable results that can be disseminated to the public. Self-experimentation is a form of scientific experiment in which the experimenter is the subject under study [186]. Sometimes called an N-of-1 study, it has been applied in a variety of research fields, such as medicine, psychology, and research initiatives like mHealth [62, 141]. This dissertation aims to investigate how people perform self-experiments and how we can build tools to better support them when they need guidance. The goal of this work is not to discover knowledge about a broad population, nor to find correlations, but to investigate behaviors' causality: to help people learn about what affects them, specifically for the parts of their lives that matter to them the most.

Personal informatics incorporates the collection, analysis, and reflection on various facets of personal data and experiences, primarily with the aid of technology. Recent research shows that 69% of U.S. adults already

engage in self-tracking practices [70]. Studies have also identified how people perform self-tracking, what reasons motivate them to do so, and what data they track most commonly [116, 39]. The findings from these investigations have shown that people generally perform descriptive analyses and that tracking one's own behavior is a beneficial process. People perform self-tracking for various reasons, including to be mindful of their behavior or to improve their lives by solving a current problem they are experiencing. Although self-tracking has other benefits and can reveal interesting insights or correlations, we focus on determining a causal relationship, which requires a different approach: self-experimentation. Our findings are aimed at empowering people to run effective self-experiments.

Beyond passive monitoring, the next step towards better understanding one's self is to perform self-experiments: to create and test hypotheses on the effect of small behavior changes [68]. Self-experimentation, however, can be challenging for people as it entails collecting and analyzing data in a systematic way. Even "extreme users" as defined by Choe et al. with experience in self-tracking encounter difficulties in rigorous self-experimentation [39]. Many individuals who perform self-tracking do not have the capability to conduct such analyses or run rigorous experiments, and may create "under-specified goals that [are] not actionable" [115].

As a first step towards making self-experimentation accessible, we focused our approach on a single domain: sleep. Sleep problems are estimated to impact over 70 million people in the United States alone, resulting in $50 billion of lost productivity annually [130]. However, people might not know exactly how to improve their sleep habits, and prior studies have shown that they are receptive to sleep-related suggestions [51]. The most common suggestion for improving sleep is to follow the generic sleep hygiene guidelines, such as "sleep 7–9 hours" or "avoid caffeine close to bedtime" [3, 5]. While these guidelines may be helpful for the overall population, they fail to acknowledge individual differences and thus might be inappropriate or even detrimental to an individual's sleep. For example, chronotypes [156], the characterization of a person's tendency to wake up early or feel refreshed when they go to sleep late and wake up late, are not typically incorporated in these recommendations.

Individually-tailored methods for improving sleep require patients to be observed in a sleep clinic by a physician using costly and obtrusive sensor technology such as polysomnography. In contrast, prior research has shown that people are most interested in unobtrusive sleep monitoring technology that does not require additional devices [36], making the smartphone an ideal form factor for sleep monitoring. Indeed, widespread use of tools to track aspects of our daily lives are on the rise, meaning that while the bulk of our work has been completed in the domain of sleep, our findings can be applied to other areas of well-being and lifestyle.

1.2 Overview of Contributions

Each chapter in this dissertation builds towards our goal of making self-experimentation accessible to novices. We develop guidelines and systems (including smartphone apps) that guide the user through the steps of the self-experiment. A central theme of our approach is making all our apps publicly available. To this end, we open source our system code and list the apps publicly on the Android and iOS app stores. The bulk of this thesis has been previously published in journal and conference papers.

SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations

***Daskalova, N.**, Metaxa-Kakavouli, D., Tran, A., Nugent, N., Boergers, J., McGeary, J., & Huang, J. Sleep-Coacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST 2016).*

As mentioned earlier, sleep is an ideal domain to start exploring solutions for systems for self-experiments. Previous work has shown that even when people are aware of the general sleep hygiene guidelines, they often do not adhere to them [126]. We set out to study how we can automate the process of providing personalized health recommendations. We developed SleepCoacher: an integrated system, which combines automated data collection using smartphones with input from professional clinicians to collect user data. The system guides users through iterative experiments to test the effect of recommendations on their sleep. Unlike most other sleep tracking systems, SleepCoacher was developed through a novel approach of collaborating with sleep experts, as described in Chapter 3. The design choices behind how the system selected and phrased the recommendations, as well as how it presented the results, were made with their guidance and feedback. Chapter 3 also provides a detailed explanation of the system and the studies we conducted to evaluate its effectiveness.

Cohorts of Self-Experimenters: Lessons Learned from Personal Informatics Self-Experiments

***Daskalova, N.**, Desingh, K., Papoutsaki, A., Schulze, D., Sha, H., & Huang, J. Lessons learned from two cohorts of personal informatics self-experiments. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (Ubicomp 2017).*

Chapter 4 describes the guidelines for self-experiments we developed based on our observations of how two cohorts of novices designed and developed their own self-experiments [47] without the help of any specific guiding system. The study showed that people often design experiments that are heavily affected by confounding variables or ones that require too much manual tracking, which in turns leads to tracking fatigue. Based on these findings and on a review of existing literature, we have developed a set of guidelines that novices can use to run successful self-experiments based on an iterative framework that encourages them to revise the setup until they are confident in it.

SleepBandits: Guided Flexible Self-Experiments for Sleep

***Daskalova, N.**, Yoon, J., Wang, Y., Beltran, G., Araujo, C., Nugent, N., McGeary, J., Williams, J., & Huang, J. SleepBandits: Guided Flexible Self-Experiments for Sleep. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2020).*

In the studies we completed to evaluate both SleepCoacher and the guidelines for self-experiments, we noticed a recurring theme of tracking fatigue and a lack of adherence to the intervention. In order to address those issues, we developed SleepBandits, a modified version of SleepCoacher, described in detail in Chapter 5. In addition to collecting sleep tracking data, the SleepBandits smartphone app allows people to select their own interventions and length of their experiments. To achieve that, it uses the Thompson Sampling heuristic to provide users with the condition (baseline or intervention) that is most likely to help their sleep on any given day. Every day, SleepBandits presents the likelihood that the intervention is helpful, so the user is free to move on to a new intervention at any point. To evaluate the utility of this approach, we released the app to the Google Play Store and conducted a study with a broad audience of users. We find that SleepBandits and its use of Thompson Sampling are effective at helping novices perform sleep self-experiments, while ameliorating the issues of tracking fatigue and lack of adherence to the intervention.

Self-E: Guided Self-Experimentation Beyond Sleep

***Daskalova, N.**, Kyi, E., Ouyang, K., Park, A., Nugent, N., & Huang, J. "Self-E: Practical Self-Experiments." (in submission).*

While the first self-experimentation systems we built were focused on the domain of sleep, we wanted to explore whether the approach for flexible self-experiments is applicable to other aspects of our well-being. Thus, we developed **Self-E**, a novel system which guides users through the steps of self-experiments beyond sleep. Chapter 6 describes the system implementation and study in more detail. Similarly to SleepBandits, Self-E lets the user choose which behavior to change and then it automatically guides them through the whole experiment. Self-E tracks common variables like productivity, diet, and exercise, and provides instruments for reporting them: for example, it uses experience sampling to collect responses related to mood tracking.

Investigating the Effectiveness of Cohort-Based Sleep Recommendations

***Daskalova, N.**, Lee, B., Huang, J., Ni, C., & Lundin, J. Investigating the effectiveness of cohort-based sleep recommendations. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (Ubicomp 2018).*

Self-experimentation systems that provide personalized recommendations are limited by the amount of data collected for each user. The first time an individual uses SleepBandits or Self-E, they have to either pick

a recommendation from a general list or come up with a hypothesis on their own. However, these systems can be further modified to provide suggested experiments based on what cohorts of other similar to the individual users have undertaken in the past. To investigate this claim, Chapter 7 describes a qualitative study with which we explored how to leverage other users' data to provide meaningful recommendations and what users believe their cohort should be based on.

Chapter 2

Related work

This chapter presents an overview of the previous work related to self-experiments, personal informatics, single-case designs, and current methods for sleep tracking.

2.1 Behavior Change and Persuasive Technology

The transtheoretical model of health behavior change claims that there are six stages of change: precontemplation, contemplation, preparation, action, maintenance, and termination [143]. Most self-technologies are focused on helping people monitor their behavior and health progress over time.

Persuasive technology aims to promote changes in users' behaviors or attitudes [67]. Researchers often try to change behavior based on a set of generic guidelines, for example to prompt smoking cessation [66]. One such behavior change system, ShutEye, focuses on displaying sleep hygiene guidelines on a user's mobile phone home screen [18]. On the other hand, Fogg's "Tiny Habits" approach implies that knowing what change to make is not always enough to actually change behavior. However, it might be easier for users to create a new habit if they are already seeing positive changes from their self-experiment.

Such technologies, however, assume that there is a generalized set of advice that works for everyone, and may neglect the reality of individual differences. Prior work indicates that an individually-focused closed-loop system consisting of self-monitoring and suggestions can improve sleep [48]. The systems outlined in this dissertation aim to address the lack of personalized tools providing actionable feedback.

Some non-clinical studies evaluate their systems based on the behavior change of users, but Klasnja et al. [106] argue a better focus for early stage human-computer interaction technologies would be on the users' experiences with the system. Thus, in our work, we analyze qualitative data from participant interviews and questionnaires, as well as some quantitative metrics about the use of our systems. This method is in line with previous studies, which use similar feedback from participants to evaluate their systems.

2.2 Self-Tracking and Personal Informatics

People have been tracking their own behavior, health, and feelings for a long time. Diaries are an example of such record-keeping as they provide the means to look back and reflect on one's experiences, or simply because "we forget all too soon the things we thought we would never forget" [54]. Recently, self-tracking devices have become ubiquitous, allowing passive tracking of a wide range of variables. People can now record not only aspects of their health such as calories and amount of time slept, but also how they spend their time and money.

The most common tools for self-tracking are smartphones and other portable and wearable devices, such as FitBits [1]. Previous studies of self-tracking have addressed areas such as food intake [44, 45, 40], personal fitness [81], multiple sclerosis [11], mindfulness [12], migraines [164, 140], menstrual tracking [60], personal finance [98], mental wellness [99], and productivity [187, 87, 103, 157, 102]. Prototypes for manual and automatic self-tracking of general factors in one's life have been developed by Kim et al. [104].

However, the amount of data people collect about themselves is so overwhelming that certain innovations focus on synthesizing the information from multiple platforms and presenting it in a simpler, more understandable form [19]. Interpreting the collected data is challenging, so people often turn to health providers for help [39, 116, 164]. It has been observed that even experienced users fail to make the most of their personal data even if they desire to do so [38]. The "Quantified Self" community comprises individuals who use and design tools for personal informatics [112]. Quantified Self participants hold Meetups around the world, during which they present what they tracked, how they tracked it, and what they learned from it. Choe et al. studied this community by analyzing videos from the Meetups and extracting valuable lessons from the self-tracking practices of this extreme user group [39]. They found that Quantified Self enthusiasts compromised the validity of their results due to three common pitfalls: 1) tracking too many things, 2) not tracking triggers and context, and 3) lacking scientific rigor, such as not including control conditions. In our research, we look at how guidance can support more scientifically rigorous N-of-1 style experiments.

Hekler et al. point out that many current self-tracking technologies do not provide the tools to self-experiment, as knowledge on its own is not enough for behavior change [84]. While self-tracking tools can help people make their own interpretations about their data [11], that alone does not lead to actionable changes [115]. However, such tools can be useful in gathering the appropriate data to then be used to determine causal relationships for effective lifestyle interventions [47, 94]. These tools can also be complemented by previous research in the persuasive technology field to develop encouraging and trustworthy software [67, 31].

2.3 Self-Experimentation and Existing Sytems

Sanctorius of Padua conducted one of the earliest documented examples of scientific self-experiments. Since then, self-experimentation has been applied in a variety of research fields such as medicine and psychology.

One notable example of such experiments performed as a classroom assignment is that of Allen Neuringer in an introductory psychology class [133]. Neuringer asked his students to perform a self-experiment for 2 weeks to illustrate the possibilities of self-experiments outside the laboratory. Some of his students designed their experiments in phases and mainly looked at the difference of means between the conditions. In this dissertation, we build on that model by providing more structure to the steps of the self-experiment, making it accessible to a broad range of novices.

The value of personal informatics comes from the process of discovery and reflection on one's data. Anyone can start self-tracking, but only people who know what to study and how to interpret the results will gain useful insights [154]. Li et al. derived a stage-based model of personal informatics composed of five stages (preparation, collection, integration, reflection, and action) and identified barriers that current systems pose in each stage [116]. They argue that personal informatics tools should allow users to iterate on their experimental stages to find the optimal procedure; this supports the iterative model of self-experiment design that we present in Chapter 4. Epstein et al. [61] build on Li et al.'s model by expanding it to the processes of preparing and selecting tools for behavior change, as well as maintaining the new behavior. They emphasize that some trackers either wanted to or had to change tools during their experiments [61].

Some forms of personal informatics do not require self-experimentation. For instance, people track simple things such as daily steps or number of push-ups to motivate themselves toward specific goals, or to archive various aspects of their lives as a new way of journaling and reflection [39]. However, to determine if there are causal relationships between variables in their lives, people must perform self-experimentation to get scientifically valid results. Self-experiments can be further motivated by a user trying to solve a problem by finding the right behavior change protocol to address the issue [115]. Lee et al.'s work investigate how to help people (mostly students) develop a behavior change protocol using habits developed based on triggers and SMART (Specific, Measurable, Actionable, Realistic, and Timely) goals.

Roberts has played a pioneering role in introducing self-experimentation to the self-trackers that are new to the Quantified Self community [152, 153]. He ran numerous experiments over a period of twelve years, identifying several novel causal relationships which he later found to be related to conventional research findings. He also popularized a method for weight reduction based on his experiments, which was anecdotally reported to be effective. He argues that self-experimentation has several benefits over conventional research, including strong self-motivation, no limit to experiment duration, and easier idea generation and validation.

Karkar et al. present an initial framework of how to run self-experimentation, specifically with a focus on problems with irritable bowel syndrome [95]. While the framework works well for the proposed cases with no carryover effect, it would need to be customized further to adapt for other domains. Furthermore, another limitation of the main premise of the framework is that the self-experimenter would re-run the study if they are not satisfied with the results after the end of the study. We build on this work by accounting for these issues in our proposed guidelines for self-experimens, discussed in Chapter 4.

The goal of self-experiments in the context of personal informatics is finding knowledge about oneself

that is individually meaningful [115, 39]. Previous systems have explored self-experimentation in specific domains, such as TummyTrials for IBS management [94], or Trialist for chronic pain management [17]. QuantifyMe, another self-experimentation app, asked users to follow a rigid experimental schedule that only 1 of the 13 participants was able to adhere to and to finish a self-experiment [162]. These studies contributed to self-experimentation literature through collecting and analyzing qualitative data [57], providing rationale for single-case design experiments [94] and improving quantitative data evaluation methods [47, 94, 95]. Two systems, Paco [63] and Galileo [180], exist to help people conduct experiments outside of a study setting. However, they are not optimized to help novices conduct self-experiments with only their own data.

2.4 Single-Case Research Design

The traditional method that clinical studies employ to determine causality involves numerous subjects and is called a randomized control trial (RCT)[42]. Single-case study designs (SCDs), on the other hand, can help a specific individual determine the efficacy of different interventions. In SCDs, hypotheses on the effect of different changes in lifestyle or behavior are created and tested.

Due to inherent limitations, RCTs are not ideal study designs for self-experimentation. For example, most findings from RCTs are based on the responses of average persons in a study; therefore, people who are not at the center of this bell curve may not respond as expected to a particular intervention, as has been reported in reactions to drugs [131, 146]. The average can also be skewed by significant exclusion criteria or lack of diversity in recruitment for such studies [125]. RCTs would not inform an individual about their specific case, and a controlled environment is unfeasible for apps and burdensome for people.

In SCDs, on the other hand, individuals serve as their own control, and baselines are established based on whether an intervention was carried out or not, which can reduce the inferential errors of group analysis in RCTs [100]. Thus, SCDs allow the empirical testing of whether an intervention is effective for an individual, which makes them more suitable for self-experiments because they provide more personalized interventions and flexibility than RCTs [94, 119].

Self-experiments in the form of SCDs have been conducted by academics from medicine [101] and psychology [155], as well as by non-academics in areas involving well-being in Quantified Self and its practitioners [39]. The standard rigorous single-case designs, specifically Kratochwill et al.'s standards for single-case designs [111], aim to reduce the confounding effects of time-based events and other the commonly cited limitations of SCDs such as internal validity [94, 83].

These standards, created by a group of quantitative and single-case design methods experts, are based on the AB phase design. In the AB experimental design, the participants change their behavior at a predefined time in order to analyze the effects of the independent variable on a dependent variable [176]. This change in behavior signifies a new phase of the experiment. The A phases are one pattern of behavior, usually

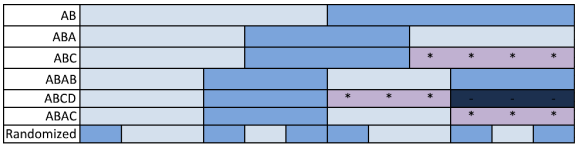


Figure 2.1: Experimental design methodologies can have different time series patterns. For the randomized design, each phase is just one measurement. For AB designs, each phase has a randomized length with multiple measurements. There can also be more than just two types of phases: in ABC, for example, there are two levels of treatment in addition to the control phase.

the baseline, and the B phases are another pattern of behavior, usually the intervention. Figure 2.1 shows the different design methods and their phases in a time series experiment [15]. For example, the ABA experimental method divides the experimental period into three phases: the experimenter starts with behavior pattern A, followed by behavior pattern B, and then A again.

According to Kratochwill et al., the ABAB phase design variation is the most appropriate one for experiments with carryover effect, as the basic AB design is susceptible to confounding variables, thus affecting the quality of the conclusions [111]. The ABAB design, and even more sophisticated designs such as the ABABAB, allow for the suggested minimum of three attempts of change to demonstrate the intervention effect. The ABAB phase design, which combines the least bias with enough time for users to acclimate to the recommendations, was used in SleepCoacher [51] (Chapter 3), which studied the success of a sleep tracking system that provided participants with personalized sleep recommendations. However, as discussed in our findings in Chapter 4, the time period required to complete an ABAB phase design might deter novices from conducting self-experiments. Thus, for the development of the SleepBandits and Self-E systems (Chapters 5 & 6), we used a randomized phase design in which each day is a new phase. Chapter 5 expands on how we further evolved our self-experimentation systems by using Bayesian techniques in the data analysis, which aim to increase rigor while maintaining practicality [165].

2.5 Sleep Tracking and Studies

Sleep is one of the most commonly tracked personal informatics variables [117]. Yet, in the domain of sleep monitoring, the existing professional solutions use specialized equipment to improve detection of some sleep events. These methods are costly and require professional oversight.

Polysomnography (PSG) is the traditional method of sleep monitoring used to detect sleep disorders [169]. PSG is an overnight study performed in a hospital or sleep clinic. It can cost patients hundreds to thousands of dollars, and requires the placement of medical equipment including electrodes on the scalp, eyelids, and chin, heart rate monitors, and other devices [22, 173]. Although this is a noninvasive procedure, PSG is obtrusive, costly, and cannot detect occasional problems as sleep in clinics may not be representative of *in situ*.

Low-cost alternatives to professional sleep tracking include smartphone applications and wrist-worn devices, such as Fitbit, which leverage a built-in accelerometer to employ actigraphy [89, 142, 161], a technique which infers sleep and wake states based on the person's movement patterns. While these trackers have shortcomings such as limited battery life, non-standardized accuracy, and discomfort [108], research shows that one in ten Americans owns one [114]. However, such devices mainly gather data and show summary statistics and general sleep tips.

Some non-clinical sleep studies have also focused on building systems that use various sensors to detect sleep events or predict sleep quality [128, 96, 78, 51]. While the accelerometer is the best feature to use when predicting sleep duration [34], it may be less accurate when placed further away from the body.

2.6 Conclusion

Self-tracking systems are becoming ubiquitous and are helping people reflect on their experiences. However, they do not provide the necessary tools to perform self-experiments. This dissertation builds on existing personal informatics research to create automated and personalized behavior-powered systems that guide users through a complete cycle of a self-experiment. Thus, the intervention is constantly evolving over time, so the user is always being asked to make actionable changes and is notified of ongoing results. The work poses a new paradigm for societal improvement – flexible and computationally guided self-experiments.

Chapter 3

SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations

This chapter presents SleepCoacher, a sleep tracking system that collects data using smartphones and provides clinician-approved personalized recommendations to improve sleep. This chapter is a summarized version of [51], where I was the first author and was responsible for the implementation, the running and analysis of the user studies and a majority of the writing.

Millions of people have downloaded sleep monitoring apps, which sense noise using the phone's microphone and movement using the accelerometer, to show users their sleep patterns [4, 134]. Users of such apps are receptive to recommendations about behaviors preceding sleep to improve their sleep hygiene [2, 18].

While offering an improvement over traditional methods, current app-based solutions lack many of the features of successful clinical methods, including personalized analysis and professional guidance. Our system, SleepCoacher, addresses this deficit by implementing a self-experimentation framework based on clinician-generated sleep recommendations. SleepCoacher goes beyond the description and visualization of sleep patterns to automatically generate tailored recommendations for improving sleep based on sensing data.

Our contribution is twofold. We present: (1) a framework for guiding users through *personalized micro-experiments* in cycles, observing the impact of data-driven recommendations over time and improving iteratively; and (2) SleepCoacher, an open-source system implementing this framework for the purpose of improving sleep.

3.1 Related Work

While personal informatics and persuasive technology tools have advantages and disadvantages, neither is sufficient for troubleshooting complex individual phenomena. Personal informatics researchers collect user data, but generally do not take the next step of using the data to generate recommendations, and test the efficacy of such recommendations. On the other hand, while single-case experiments may involve a baseline and intervention period, these experiments are often small-scale anecdotes and are not rigorous enough as they do not incorporate enough data to allow for the development of a predictive model. This work aims to combine these methodologies into an integrated closed-loop model by tracking the effects of personalized feedback over time.

3.2 SleepCoacher System

As mentioned before, sleep is one of the most commonly tracked variables of our health [117], and yet current technologies provide mainly summary statistics. If a novice wanted to better understand the effect of an intervention on their sleep, they would have to learn how to appropriately collect data, design a study setup, and analyze the data. In order to lower the burden on the user, we set out to develop SleepCoacher: a system for self-experiments in sleep. SleepCoacher combines automated data collection using smartphones with input from professional clinicians to collect user data and, in return, send daily feedback and participant-tailored recommendations to improve sleep. Participants follow each recommendation for a number of days in a predefined experimental design. The system then determines whether the intervention had a positive effect on sleep and sends the user a message with the conclusion of the experiment. It also generates a correlations profile for each user, mapping the different factors of their sleep to key metrics, and then the feedback loop repeats (Figure 3.1). Basically, SleepCoacher iteratively learns which recommendations are effective, informs the user what they should continue doing, and over time gradually improves the user's sleep in the long term.

The SleepCoacher system uses a novel recommendation testing methodology consisting of four key components: (1) gather baseline data for 5–6 days, (2) calculate personal correlations between independent and dependent variables, (3) generate and deliver relevant recommendations based on the highest correlation, and (4) test whether following this recommendation improved the target sleep variable, thus suggesting causality, by measuring the impact of the intervention over 10–11 days. This framework allows for the exploration of possible causal relationships since impact is tracked over time, as well as the cyclical structure to allow a user iteratively improve over time. The complete open-source SleepCoacher system is available online at http://sleep.cs.brown.edu.

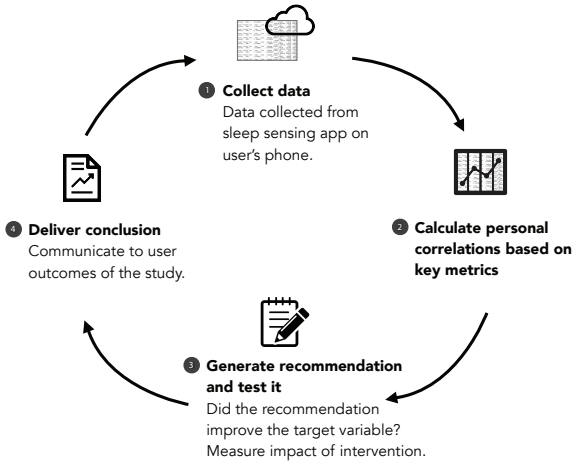


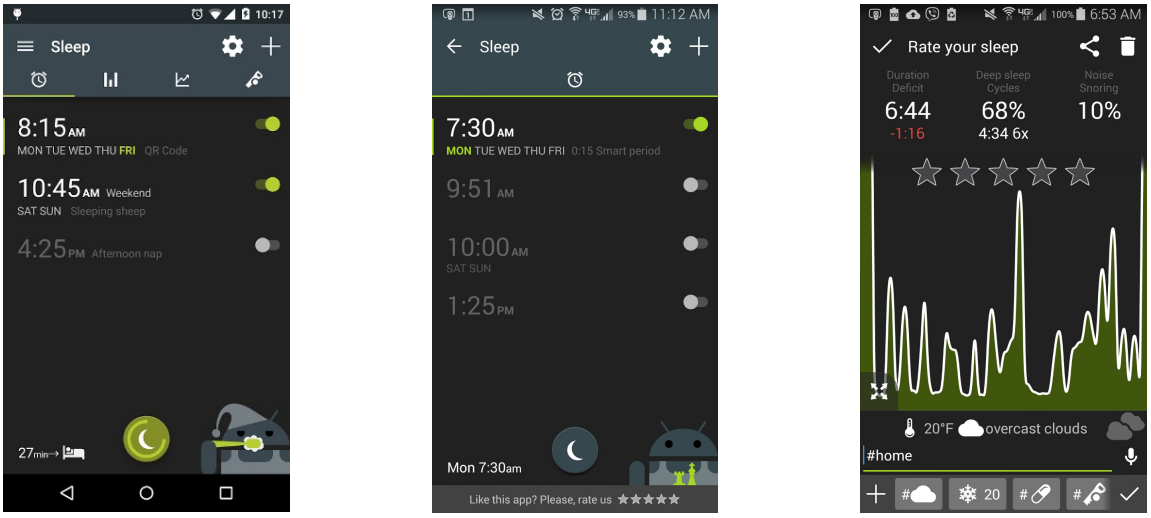
Figure 3.1: SleepCoacher employs a closed feedback loop: a user's data is uploaded to the cloud and a profile with correlations is created for each user. Next, recommendations are generated and sent back to users, who adjust their sleep habits accordingly.

3.2.1 Sensing and Data Processing

SleepCoacher's underlying framework can be applied to sleep improvement on top of any app which collects motion and noise data. For this study, we worked with developers of an Android sleep self-tracking app, Sleep as Android, which has over 10 million downloads (1.5 million of whom are active users) [4]. Sleep As Android provided us with a modified version of their publicly available app, which captures higher resolution movement data. We made further modifications to simplify the interface for our study, removing visualizations and extra options that could confuse users or influence their usage of the app and perception of recommendations. Figure 3.2(a) shows the home screen of the original Sleep As Android app with its four tabs, whereas Figure 3.2(b) shows our modified version, with just a single tab.

The application collects bed and wake times, accelerometer-based movement data at 10-second intervals, microphone noise levels at approximately 5–10 minute intervals and times of any alarms set and snoozed. Upon waking up, users stop tracking by manually indicating they are awake, and as shown in Figure 3.2(c), the app also collects the user's self-reported rating of how refreshed they felt upon waking up (by picking a 1-to-5-star rating), and user-associated tags for each night's sleep (e.g. #earplugs, #alcohol). From these features, SleepCoacher computes the sleep onset latency and awakenings throughout the night using heuristics common in sleep actigraphy literature [9, 138]. Details can be found in [51]. The app uploads the night's data to our servers under an anonymous identifier.

Our system then downloads the users' sleep data, computes statistics such as hours slept and sleep onset latency, and sends daily feedback based on these details to each user. Next, we compute Pearson correlations to determine which intervention suggestion to send to each user from a collection of recommendations provided



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

Figure 3.2: (a) The original Sleep As Android home screen contained multiple tabs with extra graphs and features that were deemed unnecessary for the purposes of our study. (b) We simplified the interface in order to minimize distractions in the app. (c) Once the user stops tracking in the morning, the app asks them to rate their sleep between one and five stars, and also lets them add tags with anything that was relevant for them, for example #home if they slept at home.

by sleep clinicians based on each user's raw data. Finally, we determine whether the recommendation had a positive effect on the target sleep variable.

3.2.2 Sleep Clinician Input

Two clinical researchers from the Bradley Hasbro Research Center and a psychiatry and sleep researcher from the Providence VA Medical Center provided input in the design of SleepCoacher's analyses. One of the clinicians is a nationally-recognized expert in behavioral sleep medicine, studying the effects of sleep disruption on family and academic functioning. The second investigates health behaviors in trauma-exposed populations and has clinical and research experience in the assessment of behavior change. The third researcher investigates individual differences and relates them to behavioral and mental health outcomes. His prior work includes measuring the impact of sleep quality on neurocognition and depressed mood.

We conducted two studies, a Preliminary and a Final one, as described in Section 3.3. In the Preliminary study, clinicians provided recommendations for each user on a per correlation basis. This process taught us that clinical insights could be pattern-matched into a collection of recommendations. In the Final study, having worked with the clinicians to generate personalized recommendations based on user data, we aimed to expand and integrate this expert feedback at scale into a more highly automated and scalable SleepCoacher system. To

do so, we surveyed clinicians and sleep literature for common independent and dependent variables affecting sleep, creating a library of 114 recommendation templates with each independent-dependent combination mapped to a recommendation. The clinicians then edited and ranked these recommendations, and each rank was integrated into the system as a weight on the likelihood of a user receiving the corresponding recommendation. More details on this process can be found in [51].

3.2.3 Collection of Recommendation Templates and Selection Algorithm

In the Final study, we focused on three measurable dependent variables: sleep rating, onset latency, and number of awakenings per hour. We created a list of all possible independent and dependent variable combinations, both positive and negative correlations. We selected recommendations from three key lifestyle dimensions: environment (specifically factors affecting sleep such as light and noise), physical state (including diet, and exercise), and mental state (for instance stress level before bed). We augmented the three lifestyle dimensions with a fourth for the special case of sleep: chronotype, an individual's natural sleep rhythm.

Each template recommendation aligned with three key criteria, ordered from most to least important: recommendations had to be measurable (easy to observe and tag), easy for users to comply with, and empirically supported by prior research. Notably, support from prior research was the least important criterion since this work focuses on identifying individual sleep responses that may or may not match existing literature.

On the day before a recommendation was due to be delivered, SleepCoacher calculated the Pearson correlation coefficients for every independent-dependent variable combination for each participant. Then, the recommendation selection algorithm identified the combination with the highest correlation and returned the recommendations mapped to the combination.

Once the recommendation template is selected, SleepCoacher tailors it according to the user's sleep statistics and the system sends the tailored recommendation to the user. The recommendation templates included average values for certain sleep factors (noisiness, sleep onset latency, frequency of awakenings) and average and optimal values for others (bed/wake time, hours slept).

3.3 User Studies

We performed two studies: the Preliminary Study (an exploratory study of 28 continuous nights), and a Final Study for 42 continuous nights. The purpose of the former was to work with clinicians to learn how they develop recommendations based on a user's data, as well as to test the mechanics of running such a study. Next, we focus on the Final Study, but details about both can be found in [51].

The ideal participants for our studies have three attributes in common: (1) their schedules are not rigorous and thus they have opportunities to enact the interventions in their sleep habits; (2) they do not have severe sleep problems that would interfere with our study; and (3) to meet logistical constraints, they have Android

smartphones in order to run our system. We chose to recruit undergraduate students for both studies, since individuals in this group are particularly at risk for poor sleep and are also early adopters of many technologies. As such, this population has much to gain from sleep tracking personal informatics technologies. Also, relative to the rigid schedule required of most full-time working adults, undergraduates have a flexible schedule that allows opportunity for intervention.

Participants were instructed to use the sleep app nightly, placing the phone on their bed near shoulder-level. To begin tracking, participants pressed a button upon getting into bed and stopped the app upon waking up. In the morning, each participant provided a rating of how refreshed they felt (1 star: very tired; 2 stars: somewhat tired; 3 stars: refreshed; 4 stars: very refreshed; 5 stars: super refreshed). They could also add personalized tags (e.g. #whitenoise, #latecaffeine).

Following the culmination of each study, each participant was given an exit survey asking, for each recommendation, whether they followed it, found it helpful, or had other comments about the experience. They were also asked whether and (if so) how participating affected their sleep habits.

3.3.1 Final Study

The participants, 11 women and 8 men, were all undergraduate students between 18 and 23 years of age. Of our 19 participants, 17 recorded their sleep for at least 80% of the duration of the study, and the remaining two were excluded from data analysis.

Each participant received a total of 2 recommendations during this study, one every 21 days. Figure 3.3 shows the study setup based on the single-case design (SCD) standards format of the *ABAB* phase design [111], where the *A* phases are the no-intervention days, and the the *B* phases are the days with the intervention (following the recommendation). The SCD standards further state that each phase should have a minimum of 3–5 measurements, and since one measurement for sleep tracking is one night, that meant a minimum of 3–5 nights. We chose 5 nights since in the Preliminary study we saw that 3 nights were not enough to show effect on sleep. Thus, one *ABAB* cycle would be complete in 20 days. We tracked participants for a final day to

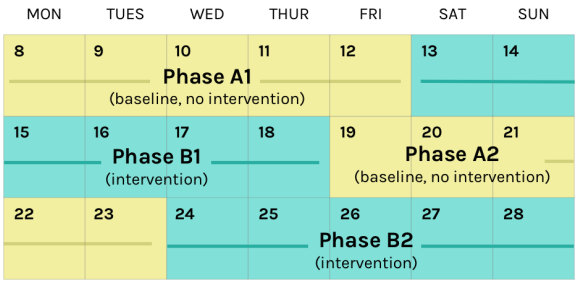


Figure 3.3: In the *ABAB* phase design of our Final Study, *A* phases (yellow) were non-intervention days, and *B* phases (blue) were intervention days.

round to a full 3 weeks, assigning that extra day to one of the previous five-day phases at random. We repeated this *ABAB* design twice in order to better evaluate the system, so each participant received 1 recommendation every 21 days, for a total of 2 unique recommendations throughout the 6-week study duration.

To pick which recommendation to send, the correlations were calculated right before the recommendation was due and were based on all previous data. The first recommendation was given on Day 5 or 6, and the second was given on Day 25 or 26.

3.3.2 Recommendations and Daily Feedback

In both studies, participants were asked to track their sleep every night and enter a rating and tags in the morning. In the Final Study, users received a text message with some statistics about their sleep every day at 10pm (called "daily feedback"). In the event that a user did not track the previous night's sleep, this was communicated to the user in lieu of a daily feedback message. Otherwise, one of four other daily feedback option was sent at random, giving statistics about the individual's hours slept, onset latency, or awakenings for the previous night. Table 3.1 includes two of those options.

3.3.3 Example Final Study Scenario

In phase A1, a participant tracks her sleep with comments and ratings. On Day 5, SleepCoacher computes correlations and finds the highest one of 0.7 between bedtime and onset latency. The system finds the recommendation templates mapped to the given combination and sends one as a text message: "On average, you go to bed at 11 pm. We've noticed that when your bedtime is consistent you tend to fall asleep faster. For the next 6 days, try going to bed at a consistent bedtime, around 11 pm." She then follows the recommendation for phase B2. Then, SleepCoacher prompts her to stop following it for another 5 days (A2), and then prompts her to follow the same recommendation again (B2). At the end of B2, SleepCoacher evaluates the effect of the recommendation and sends her a text message with the outcome: "Based on your data for the last 3 weeks, following the recommendation to go to bed consistently at 11pm helps you fall asleep 23% faster."

3.4 Findings

Based on lessons learned from the Preliminary Study, SleepCoacher sent a greater diversity of suggestions in the Final Study. It also focused on more actionable recommendations, avoiding the ones that had a low compliance rate in the Preliminary Study.

Study, Phase	Example message sent to user
Preliminary study	"When your bedtime is variable, you have more trouble falling asleep. Try to go to bed around the same time every night." "The longer you slept, the better you rated your sleep quality. You might need more sleep. On average, you slept {N} hours. Experts recommend 7–9 hours of sleep." "You wake up more often when it's noisy: consider using earplugs or a white noise generator (from an app on your phone, website on your computer)"
Final study, $A1$	**Daily feedback**: "Last night, it took you {N} minutes to fall asleep, and you slept for a total of {N} hours." **Daily feedback**: "Last night, you slept for a total of {N} hours and woke up about {N} times per hour. Usually we experience 3–5 awakening arousals every 90 minutes."
Final study, $B1$	**Recommendation**: "On average, you go to bed at {N}am/pm. We've noticed that when your bedtime is consistent you tend to fall asleep faster. For the next {N} days, try going to bed at a consistent bedtime, around {N}am/pm" **Recommendation**: "On average, you sleep for {N} hours. We've noticed that when you get {N} hours of sleep, you are on average more refreshed. For the next {N} days, try getting {N} hours of sleep. That might mean that you have to go to bed earlier than usual, so plan ahead to get {N} hours of sleep every night" **Recommendation**: "On average, the noise level of your bedroom is {N}. We've noticed that when your room is noisy during the night, you tend to take wake up more during the night. On average you wake up {N} times per hour. For the next {N} days, listen to light soft music or white noise or wear earplugs. Please tag #earplugs afterwards." "Please remember to follow your recommendation today and add a rating and a comment in the morning. Your rec was: {Recommendation}" + {Daily feedback}
Final study, $A2$	"Starting tonight, for the next {N} days, you do not need to follow the recommendation" "No need to follow the rec tonight" + {Daily feedback}
Final study, $B2$	"Starting tonight, please follow the same rec again for the next {N} days. Your rec was: {Recommendation}" "Please remember to follow your recommendation today and add a rating and a comment in the morning. Your rec was: {Recommendation}" + {Daily feedback}
Final study, End	"Based on your data for the last 3 weeks, following the recommendation to {Recommendation} did not improve your sleep or we just don't have enough data to make a conclusion" "Based on your data for the last 3 weeks, following the recommendation to {Recommendation} helps you [feel {N} more refreshed] OR [wake up about {N}% less] OR [fall asleep {N}% faster]"

Table 3.1: Examples of templates used to send messages to users depending on which study they participated in, and the phase in the ABAB experiment cycle. Messages in the Final study were automatically generated using a collection of recommendation templates.

3.4.1 Greater Adherence, Greater Improvement

In the Preliminary Study, we sent each user three recommendations, one per three days, for a total of 66 recommendations. Participants were free to choose whether to follow the recommendations or not. When surveyed, users reported following 32 of 66 recommendation cases. For some recommendations, such as wearing earplugs, we could not tell from the raw data whether the user followed them. In the Final study, we addressed this challenge by only sending recommendations which could be verified from the data and we did not need to rely on self-reported compliance rate.

While there was sleep improvement in the results of the Preliminary Study, it was not enough to show causation. We address this in the Final Study by conducting more rigorous experiments through an *ABAB* phase design. In the Final Study, we sent two recommendations to each participant over the course of 6 weeks. For each recommendation, we guided the participant to follow an *ABAB* phase design by telling them what to do each day via a text message (Table 3.1). Since each of the 17 participants received two recommendations, we had 34 cases to observe the effect on their sleep. Overall, the target variables improved in 22 of the 34 cases. A closer analysis shows that the more a user adhered to our *ABAB* study design, the greater the change in improvement. Figure 3.4 shows the improvement rate of the target dependent variable for the respective adherence rate for each of the 34 cases in this study. There is improvement in 13 of the 16 cases when adherence rate is higher than 60%, but only 9 of the 18 cases with rate lower than 60% improved. Target variables were improved in all 7 of the cases when adherence was higher than 80%.

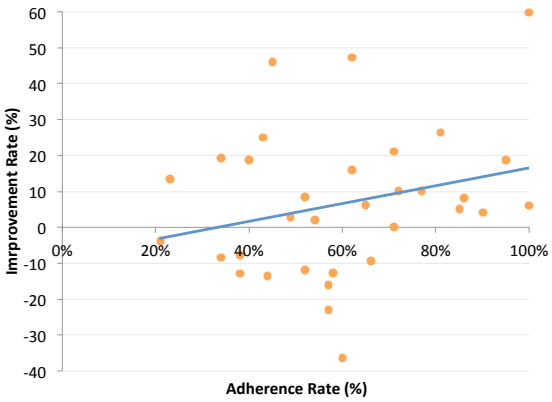


Figure 3.4: The more a participant adhered to the experimental outline in the Final Study, the more their target sleep variable improved. All participants with adherence rate higher than 80% improved their sleep.

3.4.2 Reasons for Non-Adherence

In the Final Study's exit survey, there were only two instances when users said they did not follow their given recommendation. Reasons for non-compliance fell into two main groups: participants were often not intrinsically motivated, or they found it difficult to follow concrete suggestions due to lifestyle constraints. When users found the effort or time-cost of following a recommendation to be low, many were happy to follow recommendations. In other cases, however, users were deterred by the effort needed to adjust to a new sleep behavior. Many users reported following recommendations *"as much as possible."* Overall, participants report their busy schedules and overwhelming amount of work as reasons for not being able to adhere to recommendations. This suggests that a future system needs to be more flexible and potentially suggest recommendations that do not necessarily concern exact and drastic changes, but rather start with incremental improvements. Alternatively, it could let participants select their own interventions.

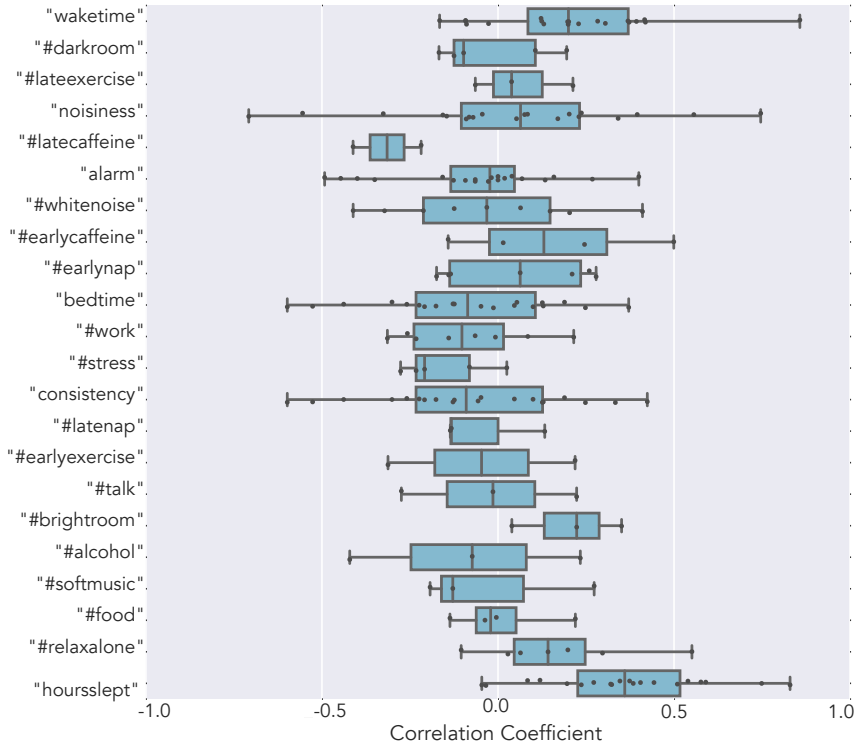


Figure 3.5: Aggregate rating correlations across all participants show large individual variation for some variables, but not for others. Every dot is a user in our study. Each bar represents the lower bound, first quartile, second quartile (median), third quartile, and upper bound, respectively. The variables with "#" are either pre-defined or personal tags.

3.4.3 Individual Differences in Correlations

Research has shown that individuals show great variation in which key factors influence sleep and other aspects of life quality [19]. Figure 3.5 shows the aggregate correlations between rating and all available independent variables across all participants. The size of the bars suggests this large degree of variation. For example, while all participants had a positive correlation with hours slept (the more hours they slept, the higher their rating), the correlation between bedtime and rating varied. This is expanded in Figure 3.6, which shows the correlations for just two participants. One of them has a high negative correlation between rating and bedtime (later bedtime leaves this participant less refreshed). The other, in contrast, has a high positive correlation between bedtime and rating (this user feels better with a later bedtime).

The range (and sign) of correlations between the independent variables and awakenings per hour or sleep onset latency are similarly varied, further strengthening the claim that recommendations must be tailored to each user's data. This data suggests that before accumulating sufficient personal data for a user, a future system can start by providing a base recommendation that works for a majority or plurality of people, such as increasing hours slept, and later tailor the recommendation algorithm parameters as more data is collected.

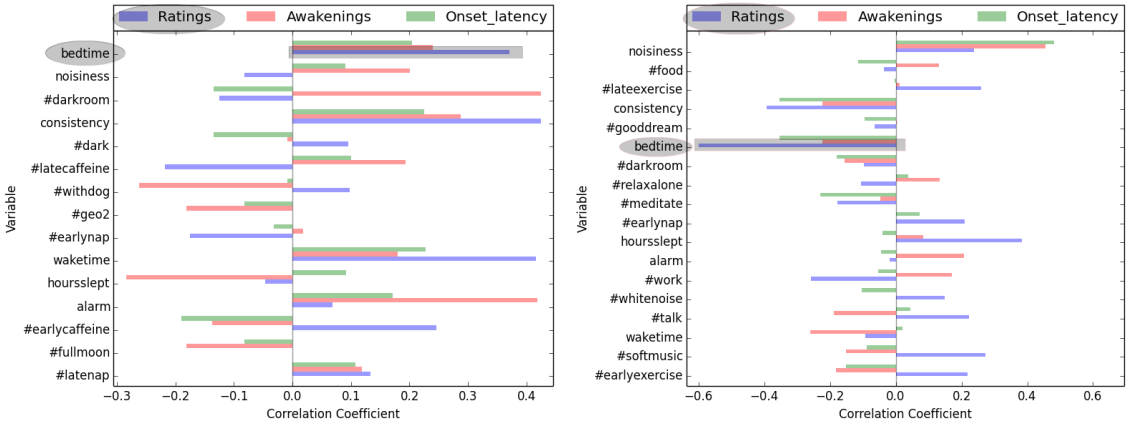


Figure 3.6: There is large individual variation across correlations between independent and dependent variables. Here, the sleeper on the left has a strong positive correlation between bedtime and rating, whereas the one on the right has a strong negative correlation for the same variables.

3.4.4 Areas for Improvement

We conducted exit surveys following each study. Overall, users felt their sleep habits were positively influenced by SleepCoacher, which is consistent with previous research on self-monitoring and suggestions [48]. Furthermore, participants in the Final Study rated the second recommendation as more personalized, which strengthens the intuition that as we collect more data for users, they receive recommendations that they are

increasingly able to recognize as personalized.

Feedback from participants about possible improvements across studies revolved around three points: (1) make the recommendations more flexible (for example by focusing on something easy to change in the sleep environment or providing multiple options and allowing the user to choose); (2) take into account more aspects such as whether the recommendation would affect a partner or roommate; (3) add explanations or references to justify each recommendation.

In the Preliminary Study, participants asked for more specific, personalized, and more frequent recommendations. The lack of concrete metrics drawn from their data made some participants less convinced that recommendations were indeed based on person-specific patterns. Thus, for the Final Study, we tried making them more personalized by adding some actual statistics for the user's sleep as can be seen in Table 3.1. We made them more frequent by sending a sleep feedback text every day, which included one piece of information about the person's sleep last night, as can be seen in Table 3.1. In the Final Study, 4 of the 17 participants said the daily feedback would be better if it combined all the information we had for the previous night. Further suggestions for improving the daily feedback were related to phrasing, adding more diverse or detailed daily feedback, and adding immediate feedback on whether the person followed the recommendation last night.

3.5 Discussion

3.5.1 Helping Users Help Themselves through Computation

At its core, the framework behind SleepCoacher provides guidance and scaffolding for users to make targeted behavior changes, and evaluates the results of those adjustments. In the Final Study, participants conducted small-scale personal experiments, altering a sleep attribute and tracking the results of that change over time. Each person has different needs, constraints, and responses to health interventions, so experimentation at an individual level is particularly valuable. Additionally, by tracking these mini-experiments and their outcomes, SleepCoacher can give better recommendations to similar users in the future through a rapid feedback cycle.

As with any automated system, attempts to force changes in user behavior may quickly be perceived as annoying and thus fall into disuse. Instead of using a prescriptive model of feedback, recommendation systems should aim to empower users by informing them about the effects of following a given recommendation.

To enable users to reliably troubleshoot through complex sleep problems, we take inspiration from control systems engineering. A closed-loop system requires four components: first, a forward path for input; second, error reduction by adjusting the system input; third, a feedback path for system output that either increases or reduces the next input; fourth, reliable and repeatable performance. We investigate how this structured cycle of repeated self-experiments could enable people to sleep more successfully and improve their quality of life.

3.5.2 Limitations

One limitation of this study is that actigraphy's degree of sensitivity does not allow it to distinguish between a user awake in bed but not moving, a user in deep sleep, or an empty bed, hindering accuracy in measuring sleep onset latency. Additionally, the commercially available sleep tracking app we used to evaluate our system does not require users to rate their sleep, though this was necessary for adherence to our study. Thus we were only able to analyze data for 81% of the nights.

Imperfect tailoring of recommendations occasionally had unintended consequences. One user, whose suggestion was to wear earplugs, gently informed us, *"I am hearing impaired and take out my hearing aids when I sleep,"* so this advice was inappropriate. To better support a diversity of users, these systems must be developed conscientiously, with the flexibility to accommodate differences.

3.6 Conclusion

This work presents a framework for guiding users through personalized, cyclical micro-experiments, combining the benefits of convenient technologies with the efficacy of ongoing observation and individually-tailored treatments. We develop and evaluate SleepCoacher, a self-tracking system for sleep improvement that automates single-case experiments through actionable recommendations.

SleepCoacher's recommendations are generated by identifying correlations between sleep behaviors and outcomes; the recommendation text comes from a collection of templates generated with the help of clinicians. We evaluate this system and the framework underlying it by conducting two user studies with a total of 43 participants. Our results demonstrate that as users adhere more to the system, they derive greater benefits, including improvements in perceived restfulness, sleep onset latency, and frequency of awakenings. We also note that correlations between aspects of sleep differ dramatically between users, validating the need for personalization, as well as the need to conduct micro-experiments targeting causality. Flexible systems that let users select their own intervention might further lower the barrier for self-experimentation.

Clinicians seek to tailor general health guidelines to their individual patients, but are limited by reliance on the person's self-report and infrequent patient interactions. Rather than attempt to recreate polysomnography and expert counseling sessions, computationally-enhanced interventions suggests a vision for healthcare that includes but also goes beyond face-to-face communication. SleepCoacher is the first step towards a personalized sleep coach for every user, with the capabilities of an automated data-driven learning algorithm.

Chapter 4

Cohorts of Self-Experimenters: Lessons Learned from Personal Informatics Self-Experiments

This chapter presents a set of guidelines for running successful self-experiments that address the pitfalls novices encounter, such as inadequate study design and analysis methods. This chapter is substantially similar to [47], where I was the first author and was responsible for the initial draft of the framework, the analysis of the studies, and a majority of the writing.

While SleepCoacher was a first step towards building an automated system for self-experiments, we wanted to learn more about how people run such experiments on their own across different domains, without specific guidance from professional tools. Previous research has shown that novices encounter pitfalls such as inadequate study design and analysis methods. We conducted a study to determine whether a set list of guidelines for self-experiments can help people run successful ones with whatever tools and analysis methods they have available. Thus, in this chapter we seek to answer: (1) what lessons for future self-experiments can we extract from observing novices run such experiments, and (2) how do guidelines on self-experiment design affect the way people run self-experiments and analyze their data?

We present the findings on self-experimentation from of two cohorts of participants (a total of 34 students in a Human-Computer Interaction seminar) performed an experiment of their choice on themselves as part of a class assignment. The students designed hypotheses, tracked the appropriate variables, and submitted reports comprising their procedures, a day-by-day journal, visualizations, and analyses. The first cohort was given minimal guidance, and the lessons we learned from how they conducted their self-experiments

informed a structured set of self-experiment guidelines. In order to evaluate their effectiveness, we then asked the second cohort of participants to follow these guidelines. Finally, we further iterated on the guidelines based on what we learned from the second cohort's self-experiments.

There are two main contributions of this work: (1) a series of lessons about self-experimentation from an exploratory study with two cohorts performed in a classroom setting, and (2) a proposed set of guidelines which aims to help non-scientists run N-of-1 style self-experiments and discover positive effects of behavior change. This is the first systematic analysis of multiple self-experiments conducted in a structured and guided environment where participants are given the freedom to choose their experiment. We combine the findings from both cohorts to present these guidelines, which can help both future self-experimenters and designers of self-experimentation tools, as one way of conducting an iterative self-experiment.

4.1 Self-Experiment Study

4.1.1 Study Method

We performed an exploratory study with two cohorts of participants: Cohort 1 and Cohort 2. The study in both cohorts was distributed as an assignment in two offerings of a seminar at a university.

Based on the findings from Choe et al.'s study [39], self-experimenters need some background in analysis and visualizations in order to effectively design their own experiments and learn from the results. Students in our cohorts developed the essential background by reading about analysis methods and visualizations and discussing them in class. They were also introduced to topics in experimental methods and behavior analysis. Furthermore, building on Choe et al.'s advice, we learned that it is not enough to have a basic understanding of analysis methods; rather, self-experimenters need to be cognizant of specific methods for analyzing results from single-subject experiments. However, real-world novice self-experimenters might not have any prior knowledge in these areas. Thus, future self-experimentation systems need to guide them through all the steps of data collection and evaluation. In this chapter, we are focusing on general guidelines that novices can follow even before the proper self-experimentation tool is created.

4.1.2 Students' Self-Experimentation Methods

4.1.2.1 Self-Experiments Method: Cohort 1

In the first part of the study, a cohort of 20 computer science students ran a month-long self-experiment as an assignment in Human-Computer Interaction seminar. In total, the assignment lasted 5 weeks, with a combined 1 week for planning and analysis and 4 weeks for tracking. The students were instructed to design and conduct a self-experiment by forming two hypotheses based on at least one independent and two dependent variables. However, they were not given any guidance on exactly what type of analysis to perform.

Table 4.1: Demographics and experience of students in the two cohorts.

	Cohort 1	Cohort 2
Total Cohort Size (N)	20	14
Male	12	4
Female	8	10
Undergraduate	8	10
Graduate	12	4
Computer Science Concentrators	19	12
Statistics Experience	13	8
Self-Tracking Experience	5	7

Stage	Goal	Length
Stage 1	Exploration	1 week
Stage 2	Preliminary Hypothesis Testing	2 weeks
Stage 3	Real Experiment	6 weeks

Figure 4.1: Stages of the study design in Cohort 2: students start with an Exploration period, followed by a Preliminary Hypothesis Testing, and finally they run the Real Experiment for 6 weeks.

4.1.2.2 Self-Experiments Method: Cohort 2

In the second part of the study, a cohort of 15 students ran a semester-long self-experiment in a later offering of a Human-Computer Interaction seminar. They were asked to follow specific guidelines about their experiment design and analysis. The assignment was run in three stages, as shown in Figure 4.1. More details about the stages can be found in [47]. It is important to note that based on the pitfalls encountered by Cohort 1 students, we made sure the self-experimentation had a built-in time for trying out the whole study design and analysis from end to end and revise it if necessary before the official data collection began in Stage 3. In total, Cohort 2 students had 9 weeks for the assignment.

Students were instructed to track any combination of independent and dependent variables, as long as there was a testable hypothesis and the data could be analyzed. We observed commonalities across all students in the various aspects of the process, including variables, confounding factors, and statistical results.

Data from self-experiments is autocorrelated and probably not normally distributed, which means that not all regular statistical analysis methods are appropriate. Participants in Cohort 2 were asked to perform specific analysis on their collected data: a t-test and an effect size calculation with a confidence interval for the standardized mean difference. According to single-case design literature, standardized mean differences and effect sizes are appropriate for self-experimenters as they are simple to perform [171].

4.1.3 Participants' Expertise with Statistics and Personal Informatics

Based on what we learned from the Cohort 1 self-experiments, we created a set of guidelines that novices could use if they were looking for guidance on running such experiments, including how to analyze the results. We presented these guidelines to Cohort 2 participants so that we could evaluate their effectiveness and improve them further. Unlike Cohort 1 students who were introduced to a wide variety of statistical methods, Cohort 2 students focused on using a difference of means test and looked at the size of the effect.

We turned to literature about single-case designs and self-experiments for advice on the best way to perform analysis on self-experiment data since this data is autocorrelated and might not be normally distributed. These stipulations are important as they violate the assumptions of most common analysis methods. Despite the abundance of techniques that attempt to address these pitfalls, there exists no single completely agreed-upon method for analyzing data from self-experiments. However, Smith's review of current methods and standards for analysis suggests using standardized mean difference approaches because the effect sizes calculated with these were the least affected by autocorrelation [171]. The advantage of such methods is that they are relatively simple to perform, which makes them appropriate for novice self-experimenters.

4.2 Study Findings

This section focuses on two of the most interesting and relevant to this dissertation findings. Each subsection is focused on a single issue uncovered in the Cohort 1 study and then addressed in the Cohort 2 study. More detailed findings can be found in [47].

4.2.1 Randomization in the Self-Experiment

Randomized single-subject experiments can be helpful for individualized treatments of patients, and systematic replication can lead to insights about a larger population [59]. The students in Cohort 1 were introduced to randomization as a good practice for an experiment. However, none of the ones who performed some form of AB phase design randomized the start of each phase.

This lack of randomization decreased the validity of their experiments. We believe that students did not realize that randomization could and should be applied to self-experiments specifically, or perhaps they did not know how to apply it to their design. To address this problem in the next study, we provided Cohort 2 students with specific guidelines on how to introduce randomization in their experiments by randomizing the moment of phase change [86]. All students' experiments in the Cohort 2 study involved some randomization.

We find that providing students with a simple script to randomize the start of the phases helped decrease the effect of confounding variables [86]. Future self-experimentation tools could help people perform more rigorous experiments by automatically introducing randomization in the experiment.

4.2.2 Self-Experiment Analysis Method

In Cohort 1, four of the students did not perform any tests to analyze the data—they relied solely on data visualizations for identifying differences between the phases. However, visual analysis is known to be inconsistent and affected by autocorrelation, meaning that it is not a reliable way to reach a valid scientific conclusion [171]. It might be more useful for generating a hypothesis, so we suggested that Cohort 2 students incorporate visualizations in Stage 2, Preliminary Hypothesis Testing.

Cohort 2 students were confused how to interpret the p-values after they performed the required analysis methods. However, the same students reported that the confidence intervals and effect sizes were easier to calculate and understand. Therefore, as previous findings suggest [73], we recommend those analysis methods to novice self-experimenters.

4.2.3 Tracking Fatigue

Possibly because of the increased study duration, tracking fatigue became a prominent challenge in the Cohort 2 study. All students in the class expressed that they felt decreasing motivation to continue tracking their data. That these self-experiments were part of a class assignment was likely the motivation for them to complete the study. It is interesting to note that students who tracked a variable more than once a day expressed that they wished they had less to track, but they had been too optimistic when they first designed their experiments. We find that self-experiment methods and technologies must address the trade-off between extending the length of the study to allow for more conclusive results and preventing tracking fatigue.

4.3 Proposed Self-Experiment Guidelines

The cohorts of self-experimenters helped expose the challenges and tensions we described in Section 4.2. Here, we revise our initial guidelines, and offer them as suggestions for one proposed way of running a self-experiment, meant to empower novices who are looking for guidance. Below is a summarized version of the guidelines, with the full details in [47]. We propose that the self-experiment be separated in 3 stages:

(1) Stage 1: Exploration – try out any devices and variables you think you might be interested in tracking.

(2) Stage 2: Preliminary Hypotheses Testing – formulate hypotheses and perform a two week test to assess the data collection, measurement, and analysis methods, and to operationalize the variables.

(3) Stage 3: Actual Experiment – either a completely randomized ABAB phase design with a set length, or a Bayesian ABAB phase design without a set length (as discussed below).

Table 4.2: Summary of the challenges identified by both cohorts and their suggested mediation.

Challenge	Suggested Mediation
collect reliable self-tracking data	explore various tools
generate a testable hypotheses	explore variables; iterate on hypotheses
support iteration	conduct preliminary hypotheses testing
control for carryover effects	randomize phase order or each measurement
conduct and interpret statistical analyses	calculate mean differences and size of the effect
avoid tracking fatigue	automate tracking; conduct Bayesian analysis

4.3.1 Choose Testable Hypotheses

As summarized in Table 4.2, one of the tensions we identified is between the need for a testable hypothesis and the lack of clarity on how to come up with one. Based on our findings, we suggest that after the initial Exploration stage, the experimenter picks what she thinks her variables should be, and then designs and conducts a mini self-experiment that runs through the basic structure of the Real Experiment. The goal of this second stage is to operationalize the intervention, variables, and measurements in order to make sure that the intervention is significant enough to make a difference. We discuss how tools can help self-experimenters choose appropriate variables by combining Karkar et al.'s framework with Lee et al.'s use of SMART goals for behavior change in the Discussion section [115, 95]. Furthermore, self-experimentation systems can provide instruments for reliable and rigorous data collection.

4.3.2 Conduct and Interpret Statistical Analyses

A major challenge that self-experimenters face is how to analyze data in the best way and how to interpret the results. We turn to literature about the appropriate statistical analysis methods for single case experiments, such as Smith and Duan et al. [171, 58]. In the two cohorts of our study, students were asked to run a t-test, calculate the standardized difference of means, and compute the confidence interval and size of the effect for their data. However, the p-value from the t-test was challenging for novices to interpret. Cohort 2 students claimed in their reports that the difference of means and size of the effects made more sense when they were analyzing their results. Furthermore, if at the end of the long self-experiment, the p-value revealed an inconclusive result, it would be uncertain whether it is due to an insufficient number of measurements or to the actual lack of evidence against the null hypothesis.

Therefore, in order to mediate this issue, we follow the recommendation of Smith [171], who emphasizes that a novice should take a simple approach and look at the difference of means. Then in order to interpret the data, she would look at whether the effect is large in the expected direction. Alternatively, one could analyze the data with a two-sample t-test, with the assumption that each measurement is an independent data point even though the data is autocorrelated since it is from the same person. Duan et al. summarize the most

common models used for dealing with autocorrelation in the data, but those models might be too sophisticated for novices to apply on their own [58].

4.3.3 Bayesian Analysis as a Way to Reduce Tracking Fatigue

One tension that we identified, and was more pronounced in Cohort 2, was between the necessary minimum length of the study and the experience of tracking fatigue. As a possible mediation, we propose the use of Bayesian analysis, as it could be particularly well-suited for self-experimentation. Bayesian methods have been discussed before in relation to single-case designs, but the focus has been on meta-analysis across participants [168]. Alternatively, Jones has focused on a Bayesian analysis using p-values as likelihoods [93].

Bayesian-based experiments can use either the ABAB phase design or the completely randomized design. Schmid and Duan further highlight the usefulness of Bayesian methods when the study design is not fixed, as it provides the opportunity to adapt the design as the study is going on [58]. This makes the Bayesian approach unique by allowing self-experimenters to stop the experiment at any moment and see the probabilistic likelihood of their interventions being effective, and simply having this option could reduce the effects of tracking fatigue. Kay et al. explore the benefits of Bayesian statistics, emphasizing that they lead to more reasonable conclusions for small-n studies, and shift the question towards the strength of the intervention effect rather than a binary "does it work" [97]. Kay et al. also discuss that confidence intervals are often misinterpreted, which makes them less reliable for use with self-experiments. The previously mentioned PREEMPT study also suggests using Bayesian analysis on the N-of-1 trials [17].

A Bayesian multi-armed bandit approach like Thompson sampling [7, 32] could also reduce the common problem of tracking fatigue by having the participant do more of the condition that is more likely to be beneficial. Specifically, the random probability that they are assigned a condition is equal to the posterior probability of that condition being beneficial. This method has been used in applications from website testing [167] to education [190] and increases the amount of benefit to users beyond traditional A/B testing.

However, it is important to note that this kind of analysis might be more challenging than a simple difference-of-means test for novice self-experimenters and can require more initial data. Thus, this method could be implemented in a system for guiding self-experiments that would analyze the collected data and provide a probabilistic result on any day. We present a possible implementation of this method in our SleepBandits systems in Chapter 5. One issue that arises, however, is that because of the nature of the self-experiment, if the experimenter looks at the current result of the experiment, she will become biased and affect the following measurements. Therefore, while it is possible to check the current probability on any day, it poses the danger of affecting the later actions and their effects. Bayesian analysis uses prior probabilities to calculate the posterior probability. In the case of self-experiments, prior probabilities might be helpful as they bring in information from previous self-experiments that might be relevant. Duan et al. discuss in greater detail the issues of the analysis of N-of-1 trials, which are important considerations for self-experiments [58].

4.4 Discussion

4.4.1 Designing Tools for Self-Experiments

In the Cohort 1 study, we attempted to address the pitfalls that Choe et al. [39] point out. However, participants still faced challenges with every step of the experimental process, including designing the experiment, collecting data, and analyzing data. We identified some of the main tensions when conducting self-experiments, and we summarized our suggestions on how to mediate them in a list of self-experimentation guidelines which we then tested with Cohort 2. We provide these guidelines as a response to each of the tensions shown in Table 4.2 so future developers of self-experimentation tools can use them as one way to provide guidance if experimenters were looking for help.

To address the challenge of finding the most reliable and convenient way to track data, it would be helpful if self-experimentation tools could recommend what variables to track and what the most common methods of tracking are. Thus, in order for designers and developers to create effective self-experimentation tools, they need to find a way to help the user explore various tools for collecting data. This might be especially important for populations with lower scientific or technological literacy, who might need more guidance from the moment they decide to start self-tracking. Complete novices might not know what devices and applications to even look at or what variables they might want to track, so a possible first step could be to prepare a guide of most commonly tracked variables and what people used to track them. This process would be a part of the Exploration Stage in our model.

In order to address the challenge of finding an appropriate testable hypothesis, we recommended that self-experimenters explore different variables and iterate on their hypothesis formulation. Similarly to showing a list of common variables and their tracking methods, self-experimentation tools might show a list of hypotheses that were commonly tested for the variables and tracking tools they picked. Furthermore, the self-experiments tool could guide the novice through setting up the self-experiment by asking a series of simple questions. Table 4.3 shows these sample questions, based on Karkar et al.'s framework [95]. They can be further combined with Lee et al.'s use of SMART (specific, measurable, actionable, realistic, timely) goals to make sure the hypotheses naturally lead to a more successful behavior change [115, 113]. For example, if a user picks "sleep quality" as her dependent variable, and "exercise" as her independent variable, the tool can then ask more specifically what she wants to track and give further options such as "time to fall asleep" and "frequency of exercise."

Furthermore, we need personal informatics tools that, by design, emphasize the iteration on the hypothesis and thus encourage and enable users to perform rigorous self-experiments. The Preliminary Hypothesis Testing stage of our model would be the perfect time to do so. Such tools need to guide users through the initial stages of the self-experiment to operationalize their variables (as in Table 4.3), visualize their preliminary data, and generate hypotheses. By conducting a more scientifically rigorous experiment, users

Table 4.3: Suggested questions for novices to match Karkar et al.'s framework [95]. The answers to all yes/no questions should be "yes." (DV – dependent variable, IV – independent variable).

Karkar et al. Absolute Requirements	**Our Suggested Questions**
DV: Well-specified (not part of the original Karkar et al. requirements)	These are the most common things people have tried to improve. Pick something you want to improve: [list of most common DVs], or something else. Now, pick a more specific aspect of your DV: [list of specific aspects]
DV: Recurrent episodes or flare-ups	Is your DV something that happens or that you do more than once in a lifetime?
DV: Quantifiable and measurable	Is your DV something that you cannot change just because you want to? Can you measure your (DV) either with a device or by hand?
IV: Controllable and actionable	Can you change your IV just because you want to? Can you measure your IV with a device or manually by hand?
IV: Well-specified	These are the most common things people have tracked. Pick something you think might influence your DV: [list of most common IVs], or something else. Now, pick a more specific aspect of your IV: [list of specific aspects]
DV: Follow the application of the independent variable within a defined period	Does your DV happen after your IV (within a reasonable time)?
IV and DV must not result in any serious health risks (immediate and/or long-term)	Is it safe for you to change your IV and DV (no health risks)?
People must be uncertain about the effect of the independent variable on the dependent variable(s)	Do you want to find out how your IV affects your DV?

are less likely to be affected by confounding variables and are more likely to reach a conclusive result in a shorter period of time.

A crucial part of any experiment is selecting the appropriate study design. However, as we saw in our study, this was challenging for the novices of Cohort 1, who received no guidance, as their designs were flawed from the beginning. Therefore, this would be an important piece that self-experimentation tools can help with—the designers and developers of such tools could lead the user through a series of questions in order to find the best study design. We summarize those questions in Table 4.4.

For example, one of the main tensions in our study was between a fully randomized experiment and the possibility of a carryover effect, e.g., if the person chooses to experiment with sleep, the application can easily point that that anything affecting sleep might involve a carryover effect, therefore a randomized phase design might be better than a completely randomized one.

Table 4.4: Suggested tasks to further guide the self-experiment beyond the choice of variables.

Tool side	**User side**
Choose variables to track	Use questions from Table 4.3.
Confirm hypothesis question	"Will I fall asleep faster if I exercise for 30 minutes that day?"
Guide user towards most appropriate study design: either completely randomized design or randomized AB phrase design	"Looks like you are tracking your sleep, and it might take a few days for a change to show its effect on sleep. So it's best to follow a "randomized phase design" which means that you will be doing the same thing a few days in a row. We will remind you every day about whether you should exercise tonight."
Use Bayesian statistics to analyze the results on the backend of the tool.	Advanced self-experimenters can perform further analyses
Present the results in a manner that novices might be most comfortable with	"If you exercise tonight, there is a 30% chance that you will fall asleep faster."

The statistical analysis methods of participants in both cohorts delineated a clear tension between conducting sound analysis of the data and reaching easy to understand findings. This is another a piece of the self-experiment that would benefit greatly from the help of a personal informatics tool. Designers could create template tools for commonly tracked activities. For example, 13 students tracked similar activities, such as sleep quality. Although the students were introduced to various experimental methods, many were still not confident in their skills and their ability to choose the appropriate kind of analysis. Self-experimentation tools

could be designed in a way that provides both basic and advanced means of computing statistics after running an experiment: the analysis method could be selected based on the individual's interest and the variables they want to track. One of the simpler methods we suggested based on Smith was to use mean differences and look at the size of the effect [171]. However, there are some more sophisticated common methods in the literature for analyzing data of N-of-1 trials [58]. The tool could present such sophisticated modules and others like intervention analysis [77, 88] and provide further guidance on when to use each tool.

One final tension identified by the study was between the length of the experiment and the needed quantity of data. If a variable can be measured only once a day, the suggested minimum study length, according to the single-case design standards, is twenty days [111]. In the Cohort 2 study, the duration was extended to six weeks as it was a semester-long project. The novice self-experimenters expressed high levels of tracking fatigue by the end of the study. In our suggestions, we recommended, similar to Duan et al., that Bayesian analyses are used in order to mitigate some of those effects, but further work is needed on developing tools that support such methods [58].

Thus, researchers and designers of self-experimentation tools need to further investigate how to strike a balance between this tracking fatigue and the participants' desire to have tracked more variables. We can turn to behavior change literature about possible solutions focused on helping users stay motivated to keep tracking throughout the duration of the experiment. At the same time, technologies need to allow for the passive tracking of a wider range of variables. One suggested way to mitigate this tension is also to build tools that include methods for calculating the power of the experiment, which would help determine how many data points are needed to find an effect.

4.4.2 Limitations

One of the main limitations of this study is that it was conducted in a class setting in a university. While we did not specifically answer questions about how to conduct self-experiments beyond what has already been discussed in this chapter, it is important to note that students were free to talk amongst themselves and this could have affected how they chose their variables and the rest of the methods. The most important effect of the classroom setting was perhaps that despite the immense tracking fatigue, students continued with the experiment, even though, as they pointed out, they would have given up if they were tracking on their own.

Another limitation of the study is that we conducted this study with two cohorts of university students, who had relatively high statistical and experimental literacy. We have suggested some ways to make self-experiments more accessible to people who lack this kind of background, but we have not yet tested them on such a population.

However, both of these limitations were necessary to help us create a controlled environment where we could present participants with the exact methods we wanted to. Future work could focus on developing a self-contained self-experimentation tool that can allow studies with broader populations outside the class

environment while still preserving the ability to control what is being suggested to users.

While we suggest questions in Table 6 and tasks in Table 7 to guide users to perform self-experiments, there might be further limitations that we have not yet addressed. Our goal was to make the questions as straightforward as possible. However, the language and format might need to be altered for specific populations. Veterans, for example, a high percentage of whom suffer from post-traumatic stress disorder (PTSD), might need to specifically focus on self-experiments related to sleep, which is particularly affected by PTSD. Therefore, a self-tracking tool for this vulnerable population would have to be more sensitive towards the kinds of variables they can track related to sleep and interventions that might be most effective for them (such as cognitive behavioral therapy [74]), rather than for the general population. Similarly, the questions and tasks might need to be fine-tuned to address the needs of other specific populations, but the overall framework would remain the same.

4.5 Conclusion

We described a systematic analysis of self-experiments conducted by two cohorts of 34 novices. The first cohort was given minimal guidance, and the lessons we learned from how they conducted and analyzed their experiments were turned into guidelines for the second cohort. We further iterated on these guidelines based on what we observed in the second cohort. We present the guidelines as one way of conducting self-experiments, aimed at novices who want to self-experiment. Based on our study, our guidelines offer an iterative structure for designing self-experiments, and propose a Bayesian approach to making statistical conclusions as better suited to self-experiments so they can be shortened or extended while ongoing.

Our work contributes to the broader understanding of personal informatics, extending prior work that emphasized the importance of self-experimentation and showed that self-trackers often lack the background to run a rigorous experiment. We learned how people conduct self-experiments when they are given guidance and a basic understanding of experimental design. This allows us to move from broad population studies, which are often cast too widely, to single-case studies which are immediately relevant and targeted to oneself.

Chapter 5

SleepBandits: Guided Flexible Self-Experiments for Sleep

This chapter presents SleepBandits, a system that helps people run self-experiments on their sleep. It incorporates a sleep tracking app and a list of suggested experiments, developed with the help of clinicians. This chapter is a substantially similar to [52], where I was the first author was responsible for the overall system design, back-end server implementation, the running and analysis of the user studies, and a majority of the writing.

5.1 Introduction

By integrating the findings from the first version of SleepCoacher (Chapter 3) and what we learned about how people conduct experiments on their own (Chapter 4), we developed a set of design principles for self-experiments that focus on maximizing user agency by identifying interventions that work specifically for the self-experimenter. We implemented these principles in the domain of sleep since, as discussed in Chapter 3, it is a focus of commonly conducted self-experiments, and allows for objective measures such as time to fall asleep and awakenings per hour [48, 162].

While self-tracking apps are popular among the general public with 10 million+ downloads on the Google Play and App Store, user compliance to continued tracking and behavior change is highly variable. Thus, it is crucial to evaluate our principles with a real-world implementation in the wild. This requires combining the natural environment of a consumer app with the statistical analysis of an empirical research study. We designed and developed a mobile app called SleepBandits. However, when we deployed it to the Google

Play Store, we published it under the name "SleepCoacher," as it was in line with our previous work and better reflected the app's purpose to the user. However, throughout this chapter, we will refer to the app as "SleepBandits." Users voluntarily downloaded and used it without any direct interaction with the authors. Like other consumer apps, we used online marketing strategies such as paid advertising campaigns and social media to recruit users.

Participants in previous studies [39, 47] experienced *tracking fatigue* if the experiment was too long or burdensome: a loss of interest in tracking because of the time and effort required to achieve a meaningful outcome. Thus, we explored an experimental design that alleviates this issue while nudging towards higher validity. Rather than a classical experimental approach or a randomized controlled trial, our user-centric method focuses on incorporating the flexibility people need to conduct an experiment in their daily lives.

As suggested in previous chapters, our implementation uses Thompson Sampling, a Bayesian approach, to analyze the data so that users receive results relatively early, which helps avoid tracking fatigue. Users receive a probabilistic outcome of what affects their sleep after only a few nights of tracking rather than several weeks. This study compares two designs: in one, users were shown the calculated result of their experiment after 2 nights in each condition (total of 4 nights minimum). In the other, they had to spend 5 nights in each condition before seeing the result summary (10 nights minimum). We find that although a 10-night study period is more rigorous, it may be too long for users as only 7% of those in the 10-night group reached a result compared to 17% in the 4-night group.

Our contribution is twofold. We present a set of proposed design principles for guided flexible self-experiments and an implementation of an open source system, SleepBandits, that embodies the proposed principles in the form of a robust app available on the Google Play Store. We discuss how this self-experimentation system maximizes user agency, and investigate how 365 active users chose an experiment, how long they conducted it for, whether the flexibility of the approach made self-experimentation appealing to novices, and what can be further improved in the design principles.

5.2 Related Work

5.2.1 Existing Systems and Frameworks for Self-Experiments

Overall, existing consumer apps for sleep tracking in general provide mainly descriptive statistics and do not guide users through self-experiments. The system presented in this chapter, SleepBandits, is the first system to implement a Bayesian approach to guided self-experiments. Paco [63] and Galileo [180] are two systems that help people conduct self-experiments in a non-lab setting. However, neither system is optimized for novice users to design their experiments: users either have to share their data with the creators of an existing experiment, or get overwhelmed with the multitude of forms to fill out when creating their own experiments. While these approaches might be ideal for a more advanced self-experimenter, it is unclear whether they are

simple and straightforward enough for a broad audience.

QuantifyMe [162] and SleepCoacher [51] are two systems aimed specifically at guiding novices through the steps of the self-experiment in a simpler manner. However, they both lack the flexibility in experiment choice and study length that people need. QuantifyMe, for example, allowed users to choose one of only four preset experiments, and its six-week study approach was too strict (only one of the 13 participants completed an experiment) [162]. SleepCoacher, focused on self-experiments in sleep, assigned people an experiment rather than letting them select one. It was also not tailored to be a robust system for self-experiments and it required a 3-week experiment length, which led to tracking fatigue and loss of interest in experimentation [51]. Furthermore, both systems were evaluated with participants recruited through campus mailing lists which do not represent the general population.

TummyTrials [94], another self-experimentation app with a focus on irritable bowel syndrome, applied a framework for self-experimentation in personalized health [95]. It allowed users to set the length of their experiment beforehand (default was 12 days, 6 per condition), but they were not able to change it once the experiment began. The participants were also not completely autonomous in setting up their experiments: they received guidance from the researchers as to what hypotheses to test and how to interpret the experiment results. The study identified areas for future improvement such as: (1) using domain experts to design a list of valid experiments and dependent variables that people can choose from, (2) incorporating "flexibility in the design to have tolerance for missing or corrupted data and ensuring common failure points are accounted for in the design," and (3) seeking a balance between scientific rigor and the reality of everyday life [94].

Our design principles build on the findings from existing self-experimentation frameworks and systems and introduce the flexibility to account for conducting such experiments independently in the wild. With these principles, users can select what interventions to try, which variable to focus on, as well as for how long to conduct an experiment. Furthermore, while our principles still guide users towards a specific condition each day, they tolerate actual user compliance in the interest of flexibility and user agency.

5.2.2 Comprehensible Results

An important design consideration identified by existing systems is how to display the experiment results to users. Previous studies have used difference of means or p-values [51, 94, 162]; however, the statistics surrounding null-hypothesis testing can be confusing for the lay audience [47, 165]. Probabilities, on the other hand, have been shown to be easier to understand if reported reliably [136]. However, it is important to acknowledge that probabilities still require a level of numeracy that not all potential users possess.

5.2.3 Dynamic Experimentation and Thompson Sampling

Multi-armed bandit algorithms have been used to ensure that data from experiments yields practical improvements. They have been applied to testing in educational games [123], identifying effective explanations and

feedback messages [190, 191], activities for mental health [139], and interventions for behavior change [110]. While there are many algorithms for solving these problems, Bayesian approaches like Thompson Sampling [7, 32] may be more easily interpreted by users [191, 165]. Our current work investigates whether such an algorithm provides users with a clear way to understand their self-experiment results.

In contrast to null-hypothesis testing, Thompson Sampling provides easily understandable numeric results that update rapidly with the user's progress (e.g., 64% chance that "earplugs" is better than "no earplugs"). In health, it offers an advantage over traditional A/B testing because the user is instructed earlier and more frequently to follow the condition that is more likely to improve their sleep. This helps users achieve their health goals sooner and may also reduce tracking fatigue.

5.3 Design Principles for Guided Self-Experiments

Previous studies have shown the need for a self-experimentation system that both maximizes user agency and introduces scientific rigor to how people run such experiments in their daily lives. Building upon recent related work [51, 162, 94, 95], we chose to focus on two main questions while developing a set of design principles for guided self-experiments: (1) How can we create a system that grants user agency in the self-experiments to address the tension between scientific rigor and the demands of everyday life?, and (2) How do we calculate results from these experiments and present them to the users in an intuitive and ongoing manner?

The four principles listed below aim to aid in addressing the needs of novices and in designing systems that support flexible self-experimentation. While there are other principles that could play a role in the effectiveness of such systems, we chose these four to focus on based on prior research [51, 94, 145, 162]:

- **Guided Agency** refers to the need not only to provide flexibility to users to select their self-experiment hypothesis and length, but also to give them guidance by nudging their choices towards the best practices (as illustrated by the findings in [47, 51, 94]). This can be accomplished through providing experiment length suggestions, a shortlist of first-time experiments, or a recommended, auto-generated experiment schedule.

- **Scientific Rigor** needs to be introduced in the experiment, for example by incorporating randomization to help account for confounding variables, since novices often do not account for them in their own designs (as shown in [47]). Randomizing the experimental condition is one way to accomplish this, and our approach uses Thompson Sampling to display one condition more frequently but still at random.

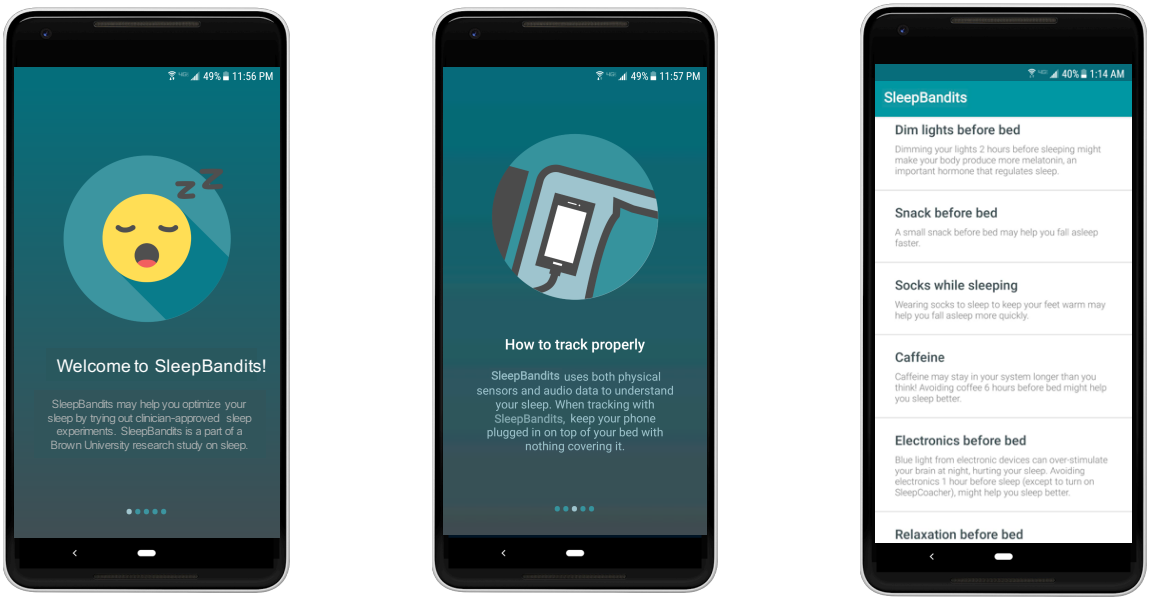
- **Tolerance** refers to the need to accommodate real-life circumstances such as missing data and lack of compliance to the experimental condition because if the experimental design is too rigid, novices will not be able to follow it (as only 1 of the 13 participants in [162] managed to finish an experiment). An 'as-treated' analysis can be applied to calculate an experiment result despite the user not following

the randomized study schedule perfectly. However, the effect of the experiment can also be calculated with an 'as-instructed' approach, and both results can be shown to the user to emphasize how much deviation from the study schedule has lowered the scientific rigor of the results.

- **Comprehensibility** refers to presenting the experiment results in an easy-to-interpret way, rather than the p-values that can be challenging for novices (as shown in [51, 94]). One way to do that is to present probabilities generated from Bayesian analysis [115, 165], such as Thompson Sampling.

5.4 SleepBandits System

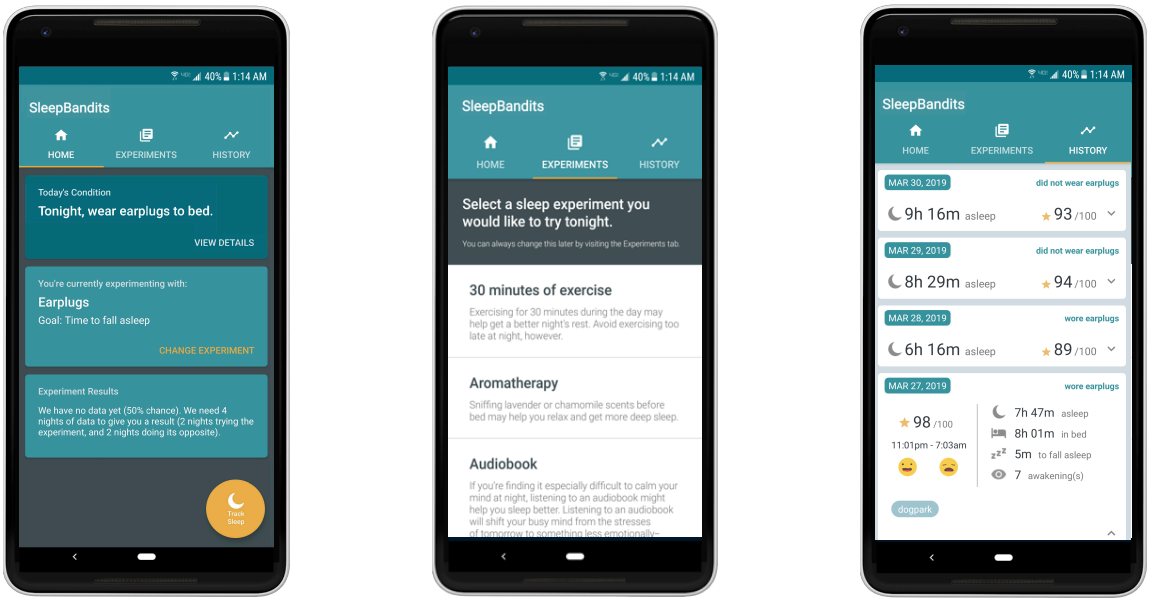
To demonstrate the value of our design principles, we implement them in SleepBandits, a system for self-experiments for sleep. SleepBandits is comprised of two components: an interactive Android smartphone application and a backend server that stores the data and performs the analysis. The complete open-source SleepBandits system is available online at http://sleep.cs.brown.edu.



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

Figure 5.1: SleepBandits onboarding screens for new users. (a) Welcome screen, explaining that this is part of a research study. (b) Screen explaining what to expect from the app and to keep the phone on the bed while sleeping. (c) The user initially has six curated experiments to choose from.

The SleepBandits mobile application was designed to work without any interaction with the researchers and run on various Android OS versions and Android smartphone models. The application collects sleep



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

Figure 5.2: SleepBandits screens. (a) Home: tonight's condition on top, then the current experiment with the option to change it, and current results below. (b) Experiment selection: users first select an experiment during onboarding, but then are free to change it at any point. (c) History of sleep outcomes: users receive an update with summary statistics for every night they track.

data by using the device's built-in microphone and accelerometer to track sound amplitude and the user's movements during the night, so the phone must be placed on the bed overnight (Figure 5.1(b)). Unlike traditional sleep tracking studies, users do not have to keep a diary of manual entries with their sleep statistics.

When users go to bed, they open the SleepBandits application and tap the "Track Sleep" button (Figure 5.2(a)) to begin collecting data. When they wake up in the morning, they tap the "Wake up" button to stop tracking, and the application compresses then uploads the encrypted data to the server. The server decodes the received data, calculates time to fall asleep and awakenings per hour, and sends back an encrypted version of this data in a few seconds. The app then decrypts it and sends a push notification to the user. Clicking into the notification, the user sees a summary of their sleep factors (Figure 5.2(c)): time to fall asleep, number of awakenings, and hours slept. This gives the user immediate feedback on their previous night's sleep quality. Keeping track of the data is done automatically, minimizing the burden of self-tracking.

5.4.1 List of Self-Experiments

According to our **Guided Agency** principle, the system must provide users with the ability to select their own experimental hypothesis, while limiting their options in order to guide complete novices towards more

scientifically based experiments. Thus, SleepBandits contains a list of 26 possible interventions, some of which are shown in Figure 5.2(b). We developed this list by using general sleep hygiene guidelines [69] as a starting point. We then surveyed medical literature and sleep research journals for habit recommendations. Finally, we recruited three experts to refine the list: a clinical psychologist with experience in behavior change, an expert in behavioral sleep medicine, and a psychologist and geneticist who focuses on how individual differences relate to health outcomes. All of the experiments on the list were purposefully selected as interventions that someone can try on a given day (e.g., earplugs, chamomile tea, room temperature) and immediately see same-night effects, minimizing carryover effect.

Following the **Guided Agency** principle, our expert collaborators selected the most appropriate first-time experiments (i.e., those that were both most likely to be helpful and required the least effort to implement). These six are the only ones that users see when first selecting an experiment during the onboarding process, which helps nudge them towards selecting a valid initial experiment without being overwhelmed by choice (Figure 5.1(c)). However, once in the app, users can see all 26 in the "Experiments" tab and change to a new one at any time.

5.4.2 Self-Experiment Variables

In accordance with the **Guided Agency** principle, SleepBandits also lets users to select one of three common sleep variables: (1) time to fall asleep, (2) number of awakenings during the night, and (3) the user-reported rating of how tired they feel upon awakening. The first two are common aspects of sleep that are tracked with actigraphy sensors in sleep studies, while subjective sleep quality is often reported via paper diaries [161, 26]. We chose not to include sleep duration or timing since people's schedules, not the interventions on our list, predominantly determine those factors. While sleep quality is complex, we chose to start with the simplest experiments, so users are asked to select only one variable to focus on for each experiment, with the default being "time to fall asleep" since it is the most common sleep complaint in US adults [149].

To determine how long a user takes to fall asleep, SleepBandits employs a heuristic from the sleep literature that was previously used in SleepCoacher [51]. The limitation of this heuristic is that it uses a static threshold to determine whether someone is awake or asleep. If a user places their phone closer to their body, the data would show more awakenings than if they kept the phone further away. There is a trade-off between static and dynamic thresholds: personalizing the threshold would require at least a week of sleep data for calibration before analysis can begin, so we chose the static one in order to show results as early as possible.

5.4.3 Interface and User Flow Design Choices

The "Home" tab contains three sections, organized in a hierarchical manner: "Tonight's Condition" is at the top, followed by the current experiment and current results (Figure 5.2(a)). "Tonight's Condition" is critical as it incorporates randomization in the experiment and guides the users on what to do each day which is at

the heart of the experiment. For example, for the "Earplugs" experiment, the condition would be to "wear earplugs" on some days and to "not wear earplugs" on others. This design was based on our **Scientific Rigor** principle, as randomization helps account for confounding variables.

The "Home" tab also contains the floating button to "Track Sleep," a common Android UI element that calls the user to the main action. Once users tap on "Track Sleep," a pop-up (Figure 5.3(a)) asks them to rate how tired they feel using a visual scale with five emojis that we designed to match the states between "very sleepy" and "very awake." Here, the user is also able to tag anything else that they did during the day that might have affected their sleep.

The pop-up also (Figure 5.3(b)) asks users whether or not they had adhered to "Today's Condition." For example, if the user was required to perform an activity during the day, such as "exercise for 30 minutes," the pop-up would ask them if they had actually completed the task. However, for overnight instructions such as "listen to an audiobook," we chose to ask the adherence question the following morning, having a pop-up appear when users tap "Wake up" (Figure 5.3(b)). This implementation, related to the **Tolerance** principle, was inspired by early informal iterations of the app in which users complained that they forgot to actually listen to an audiobook even though they said that they would, but there was no way to edit their adherence for the night. While the app could have automatically tracked the adherence to some interventions, we chose to keep the design consistent and ask for the manual input of the adherence to all experiments.

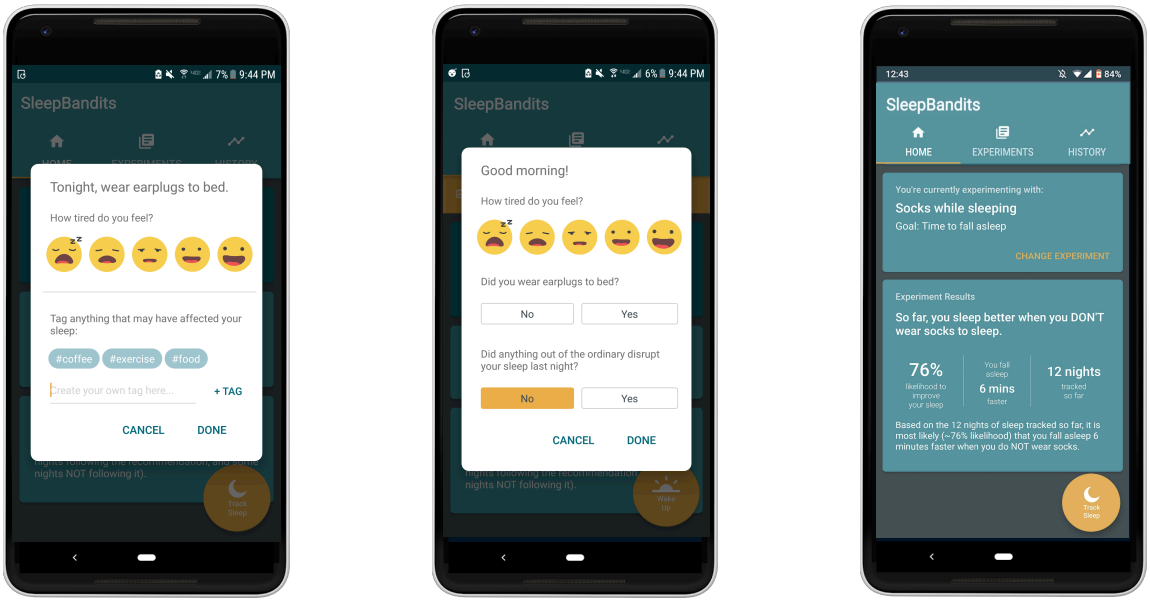
According to the **Tolerance** principle, the system needs to be able to accommodate real-life experiment compliance. To address that in SleepBandits, we applied an "as-treated" analysis [85], meaning that the difference of means was calculated according to the way users actually behaved rather than what condition they were assigned for each day.

5.4.4 Presentation of the Self-Experiment Result

SleepBandits collects data about the user and, after a few nights, uses Thompson Sampling to determine which experimental condition is more likely to improve the user's sleep. To incorporate our **Comprehensibility** principle, we had to consider how to display these results to users, some of whom might be inexperienced with statistics.

Before these results are calculated, the displayed in-app result states that there is a 50% chance that either condition will be better for sleep. After enough nights of data are collected (2 or 5 nights per condition, depending on the study group), the result changes to what is shown in Figure 5.3(c). This design is based on multiple informal iterations with users and feedback from the clinicians. In larger font is the conclusion of the experiment: "So far, you sleep better when you DON'T wear socks to sleep." This sentence was added because users noted that the text and percentage were confusing without it.

Below that, three numbers display the likelihood that the condition ("not wearing socks") is helping (76%), the size of the effect (6 minutes), and the duration of the experiment so far (12 nights)(Fig. 5.3(c)).



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

Figure 5.3: (a) User prompt before sleeping: subjective rating of how tired they feel and adherence to the condition, if applicable. (b) User prompt after waking up: subjective rating of how tired they feel and adherence to the condition (if applicable, e.g. earplugs at night). (c) Home screen explaining that wearing socks leads to falling asleep 6 minutes sooner, with 76% likelihood estimated from 12 nights of sleep.

The likelihood percent is based on the Thompson Sampling algorithm, and we discuss how it is calculated in the "Thompson Sampling" subsection below. Research shows that the difference of means is one of the best analysis methods for self-experiments due to its simplicity [47, 171]. Lastly, the sentence at the bottom places the numbers in context and summarizes the conclusion of the experiment. This way, the system provides users with information they can understand, in order to help them draw informed conclusions. Following the **Guided Agency** principle, we set a minimal length of the experiment (4 nights or 10 nights depending on the study condition), but users are free to continue their experiment for as long as they would like beyond that.

5.5 Method

We conducted a user study reviewed by our institution's Human Subjects Office between June 1, 2018 and September 1, 2019. After many iterations, the app was published to the Google Play Store in May 2018 for anyone to download. SleepBandits appeared as a regular sleep application with the appropriate affiliations and informed consent built into the app and description. All participants were people who downloaded the application voluntarily. They agreed to participate in the study and were given the agency to start or stop

using the application whenever they liked. Users were not paid monetary compensation for their participation.

Users were randomly assigned to one of two groups – the "4-night group" or the "10-night group" – to determine how many nights to require before displaying a self-experiment's result. In the 4-night group, users had to follow each condition for at least 2 nights (e.g., earplugs on 2 nights and no earplugs on another 2 nights). In the 10-night group, they had to complete at least 5 nights in each condition. We selected these lengths based on the single-case intervention research design standards [111] which require each phase of an AB phase design to have 3–5 data points. Thus, the 10-night group followed a traditional 5-day per phase AB design. The results from this group were compared to those from the 4-night group. This was designed to test the limits of this approach by being shorter than the minimum recommended standard.

SleepBandits was downloaded over 5,000 times from the Google Play store. As per our ethics protocol, agreement to the consent form and an email address are required in order to use the app. From these downloads, 1,781 resulted in registered users. We excluded 51 users from analysis due to self-identification of having a sleep disorder or taking potentially sleep influencing medication, both of which would interfere with the study results. Of the remaining users, 365 tracked their sleep for at least 1 night (39% female, 60% male, 1% other/prefer not to disclose; age range between 18 and 85 (M=33, SD=12)). This retention rate is typical for such apps because 21% of users only open an app once [121]. Overall, we collected 1,859 nights of sleep, totaling over 14,200 hours.

Finally, we conducted remote semi-structured interviews with 10 participants (5 female, 5 male) from different study groups (5 from each group) who had been using SleepBandits for various amounts of time (between 0 and 27 nights). Of these interviewees, 4 did not complete any experiments. Of the remaining 6, 4 said they improved their sleep. The goal was to get a better understanding of why they used the app for as long as they did, what challenges they faced, and what their overall impression of the flexible approach was.

5.5.1 Participant Recruiting

Understanding behavior change and self-experimentation is challenging in a lab setting, since being in a (usually paid) study leads to different behaviors than people would naturally have outside a study [120]. To recruit a natural audience, we mimicked the marketing techniques of existing sleep tracking products by posting on social news and product launch websites, advertising on sleep forums, optimizing search engine results, and creating paid online marketing campaigns. We also aimed for organic discovery in the Google Play Store. While our study does not focus on different user acquisition channels, we wanted to find users "in the wild" who would be motivated by only what the SleepBandits app itself offered.

This approach allows us to realize results that are less affected by experimenter bias which is important for personal informatics applications – especially those aimed at understanding behavior change. Specifically, while we know that tracking fatigue is one of the most common reasons people lose interest in behavior change and self-improvement through personal informatics [38, 47], we only have a qualitative understanding

of it. Experiments to identify how to overcome tracking fatigue or reduce it are nonexistent because measuring it naturally is difficult. There exists a trade-off between the challenge of tracking fatigue and the benefits of tracking. There is also a natural tension between the desire to have more days of data, a larger N, and seeing a result quickly and moving on to other experiments or other apps. The typical uninstall rate for an app is 28% [135], the one-month retention rate is 43% [121]. Therefore, behavior change and self-experiment applications in the wild must be designed with a strong focus on guided agency, scientific rigor, tolerance, and comprehensibility.

Furthermore, achieving success with voluntary users can potentially allow us to study a larger population. Most user studies tend to comprise of 10–40 users [28] as there is a natural limit to the amount of time and effort researchers can spend recruiting and engaging with participants. However, behavior studies in everyday life naturally have a lot of noise due to the variation in people's days or personalities. The format of self-experiments using traditional statistics where users follow a set study schedule for a long duration is a poor user experience. Users may get tired or frustrated with the lack of timely results, but seeing intermediate results, or "peeking," reduces the statistical validity of the experiment. Experiments can also be inconclusive even at their completion (when $p > 0.05$), leading to wasted time and uncertainty if the problem was a lack of statistical power. Thus, since self-experiments are about self-discovery, if users are not discovering anything about themselves, they will halt the experiment early.

From this, we conclude that analyzing real user data gives us a sense of what behavior change is like in the wild, since they are using an actual product rather than a research prototype. As such, rather than running a typical lab study, we deployed SleepBandits using traditional online user acquisition techniques. While this required spending more time making it compatible with many operating system versions and fixing bugs that arose from poor networks or unusual system configurations, we ended up with a system that provides benefit to any potential user. This is similar to Harvard's Lab in the Wild [148] or Citizen Science, but rather than using a survey, we offer benefits from using a sleep tracking application. We accept the challenge by Bernstein et al. [20], "to stop treating a small amount of voluntary use as a failure, and instead recognize it as success. Most systems studies in human-computer interaction have to pay participants to come in and use research prototypes. Any voluntary use is better than many HCI research systems will see."

5.5.2 Thompson Sampling

As data is collected, SleepBandits updates the parameters of a beta distribution for each experimental condition (e.g., audiobook and no audiobook), which indicates how likely an outcome is to occur. The outcome in this case is whether one experimental condition is better for the user than the other. The shape of the beta distribution is determined by the α and β shape parameters. The α is calculated as some prior probability and updated with the number of successes (number of nights when sleep is better than some threshold). The β is based on the prior of the other experimental condition and the number of failures (number of nights when

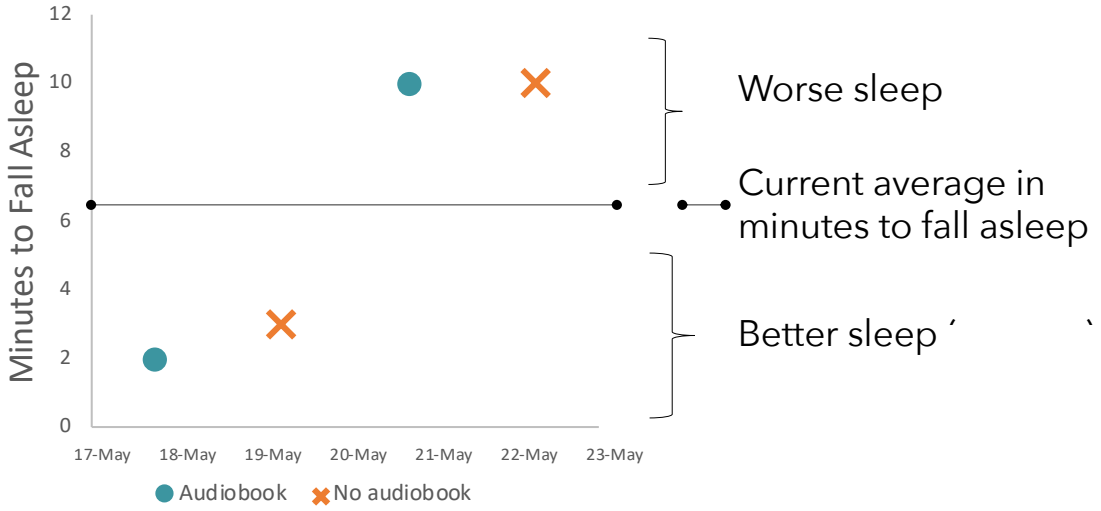


Figure 5.4: Example where the Thompson Sampling algorithm indicates a success or failure based on whether the time to fall asleep is above or below the average threshold so far.

sleep is worse than the threshold). Figure 5.4 shows an example with successes and failures.

Once we have the α and β for each condition, we can create a beta distribution for each one. Next, we sample from each distribution 1000 times, and each time we get a probability for each of the two conditions (e.g., P=0.7 for audiobook, P=0.5 for no audiobook). The one with the higher probability (most likely to be helpful, i.e., audiobook) is returned. After 1,000 times, we count how many times each condition was returned (e.g., 660 audiobook, 340 no audiobook). Thus, the likelihood the condition is helping is 66% for audiobook, and 34% for no audiobook; there is a 66% chance the user will be asked to listen to an audiobook.

Equation 5.1 shows that the goal in Thompson Sampling is to return the action x_t from a set of actions χ={1, . . . , K} that maximizes the expected value $\mathbb{E}$, where K is the number of conditions. Each observed value y_t has an associated reward r_t. Observations and rewards are modeled by conditional probabilities $q_\theta(1|\mathrm{K}) = \theta\mathrm{k}$ and $q_\theta(0|\mathrm{K}) = 1 - \theta\mathrm{K}$, where θ is the beta distribution. $q_{\hat{\theta}}$ is the expectation of θ, based on the random sample that the algorithm draws from the distribution [158].

$$x_t \leftarrow argmax_{x \in \chi} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x] \tag{5.1}$$

As shown in Figure 5.5, when a new data point arrives on the fifth day, the beta distribution for "no audiobook" is updated, and the current likelihood of audiobooks improving the target sleep outcome is estimated to be 62%. However, as a few more nights of sleep are tracked, the beta distributions keep updating and the likelihood increases to 84% (Figure 5.6).

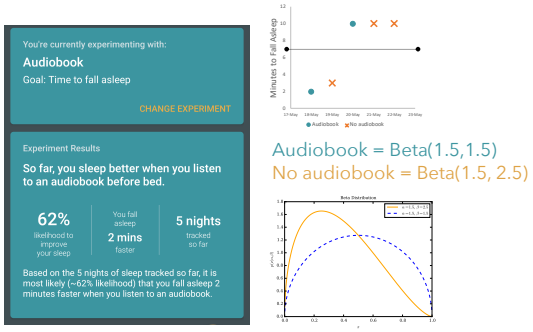


Figure 5.5: On Day 5, we have another data point for no audiobook, so the beta distributions appear as plotted, and the current likelihood of audiobooks helping is only 62%.

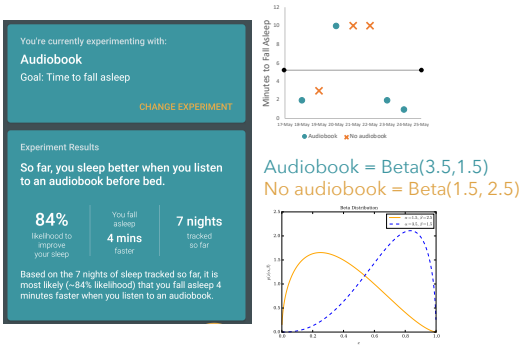


Figure 5.6: After a few more nights of data, the beta distributions changed shapes for both conditions, and the likelihood of audiobooks helping is now 84%.

In this study, we focused on just one intervention at a time as the first step to applying Thompson Sampling to self-experiments. However, it is important to note that it can be used to compare multiple interventions as well. For example, instead of the conditions being "earplugs" and "no earplugs," they could be "earplugs," "eyemask," and "socks."

5.5.3 Experiment Adherence and Duration

In a rigorous scientific setting and in previous studies [162, 51], participants are asked to use the systems in a constrained manner and to follow a specific schedule. However, our study gave participants complete freedom on how to use the application. To introduce randomization in the experiments and encourage scientific rigor, SleepBandits users were informed about which condition to follow each day (e.g., wear earplugs or do not wear earplugs). As shown in Figure 5.3(b), each day participants were asked whether they followed the app's suggestions. However, as seen in previous research [51, 94, 162], users sometimes experience unexpected life events that prevent them from adhering to the correct experimental condition. Thus, instead of excluding the nights when users do the wrong condition, we retain all the data. We recognize that this introduces a limitation in the self-experiments since the conditions in each night are not strictly randomized. People adhere to the behavioral recommendation in only about 50% of the cases [51], so this trades off some control in the randomized controlled trial configuration to accommodate natural user behavior.

5.6 Findings

The goal of this study was to implement the design principles for guided self-experiments into a robust self-experimentation system. We chose to explore these themes in the sleep domain with SleepBandits, but the principles are extensible to other domains for future research. Here, we present quantitative results along with user feedback from both the 4-night and 10-night groups. To get more context around how people were using SleepBandits, we conducted a thematic analysis on the 10 in-depth interviews. Due to the qualitative nature of this data, we did not seek measurable differences between the two groups.

5.6.1 Flexibility to Choose Self-Experiment and Target Variable

Following the **Guided Agency** principle, SleepBandits presents users with a list of experiments that have been pre-approved by experts as being beneficial for the general public. This is helpful because previous studies have shown that people often pick a behavior change they want to implement despite not necessarily knowing whether it is suitable for self-experimentation [47]. Unlike participants in the TummyTrials study [94], the ones in SleepBandits did not receive guidance from researchers on which experiment or variable to pick.

The most commonly selected first-time experiment in SleepBandits was "Relax before bed" (21% of all first-time picks). Interviewees who selected this experiment liked the low effort and preparation it required. SleepBandits also presents users with three commonly tracked sleep variables and allows the user to select one to focus on. While we expected time to fall asleep to be most commonly selected, only 39% of users selected it, whereas 46% of users selected how refreshed one felt in the morning. Most interviewees pointed out that a sign of a good night of sleep was waking up rested, which highlights the importance of letting users decide what to focus on.

Our flexible approach gives users control over what experiments to conduct, with the option to switch at any time. All interviewees thought that the ability to choose your own self-experiment was helpful, particularly for novices because, as P2 said, *"you can tailor it more closely to your life."*

5.6.2 Adherence to Instructions: Balancing Scientific Rigor and Everyday Life

Previous studies show that people do not intuitively randomize their conditions [47], even though it helps decrease the effect of confounding variables [86]. Following the **Scientific Rigor** principle, SleepBandits automatically randomizes which condition users are instructed to follow each day. While interviewees noted that the daily guidance was helpful, users only adhered to the instructions 60% of the time on average (SD=38%). The average adherence rates among previous related studies ranged from 22.5% in QuantifyMe [162], to 53% in SleepCoacher [51], and 95% in TummyTrials [94]. In comparison, the median adherence reported in randomized controlled trials was 88.4% (range: 48%–100%) [193]. SleepBandits employed an "as-treated analysis": it used all data points from a given user to calculate their result, even

if some points did not adhere to the app's instructions for that day [85]. While this reduced the effect of the randomization, a system focused solely on rigor would discard a lot of data, making experiments take substantially longer and discouraging users. Thus, SleepBandits applies the **Tolerance** principle and handles adherence rates with high variability, since this reflects the way people conduct self-experiments in the wild.

5.6.3 Effect of Minimum Experiment Length on Completion

In rigorous in-lab studies with systems like SleepCoacher [51] and QuantifyMe [162], which employed AB phase designs, participants were asked to conduct a single self-experiment over the course of weeks (16 days in [162] and 21 days in [51]), before they saw a result. TummyTrials [94] was designed to allow users to set an experiment with as few as 3 days per condition, but they only conducted a study with a set length of 12 days. The long study duration was found to lead to tracking fatigue, so SleepBandits aimed to explore whether a shorter duration led to better results.

SleepBandits employed the **Guided Agency** principle to set a required minimum number of days before a result was shown, but then let participants continue with the experiments for longer if they wanted to. All interviewees found this flexible length helpful. Nine of them thought that the required minimum number of days was fine, but most wished for a graph of their data throughout the experiment. Two of the three interviewees in the 10-night group who never reached a result discontinued self-experiment because 10 nights was too long.

As shown in Figure 5.7, the overall number of nights tracked followed a similar trend in both groups: 28% of participants in the 4-night group tracked their sleep for at least 4 nights, compared to 32% in the 10-night group. About fourteen percent of the participants in each group tracked their sleep for at least 10 nights. The presented usage rates are an important baseline for future self-experimentation systems.

This natural usage of the app could be the reason why participants in the shorter 4-night group were almost three times more likely to reach a result than those in the 10-night group: thirty-one of the participants (17%) in the 4-night group reached a result, compared to seventeen of those (7%) in the 10-night group (χ^2=19.9, $p < 0.01$). This highlights the need for systems that allow users to see results earlier since their natural inclination will likely be to conduct shorter self-experiments. That way, people will be able to learn something quantitative about themselves and make informed decisions about their behavior change choices. We also conducted a traditional t-test analysis on the users' data, which revealed that none of them would have reached a statistically significant result at the end of their experiments ($p < 0.05$).

5.6.4 Reasons for Users Ending the Experiment

Once users reached the required minimum number of days in each condition, they were shown a result (Figure 5.3(c)), representing the likelihood that the intervention was helping them. At that point, they were

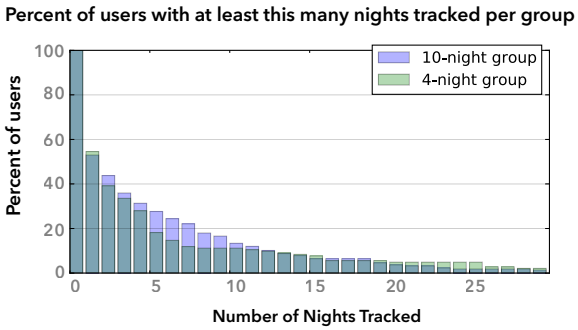


Figure 5.7: The overall number of nights tracked followed a similar trend in both groups of users. Around the fifth night, however, there is a dip in the percentage of users from the 4-night group that tracked their sleep (which is when those users saw a result in the app).

free to continue with the same experiment and collect more data points or to end this experiment by moving on to a new one or by discontinuing app use.

Four of the six interviewees who reached a result (regardless of their group) explained that they stopped using SleepBandits because they did not want to keep their phone on the bed, as it was either impractical or uncomfortable. Personal issues, such as health or traveling, were another common reason for ending the experiment. Other reasons included wanting to have their sleep tracked more passively, without the need to manually start and stop tracking, as well as to know more details about their sleep such as graphs of their sleep stages during the night. Overall, the interviewees' reasons for discontinuing use of SleepBandits were centered around circumstances beyond their control, or wanting features that further visualize their data and alleviate the burden of tracking.

5.6.5 Confidence in the Thompson Sampling Likelihood Score

Overall, users who saw either a very high likelihood percent or one between 50% and 60% were more likely to continue the experiment, whereas those who saw likelihoods in the middle range (65%-85%) were more likely to stop tracking. The median likelihood on the first day of the results for those who continued tracking further was 84% (M=77%, SD=16%), whereas those who stopped tracking as soon as they reached a result saw a median likelihood of 65% (M=69%, SD=15%). The interviewees mirrored these findings: those with initial scores under 65% or above 85% talked about wanting to track their sleep longer. On average, the 48 users who reached a result saw a likelihood of 79% on the last day of their experiment. Overall, the principles behind our flexible approach increase user agency by letting users conduct the experiment until they are satisfied with the confidence level of their result.

5.6.6 Usefulness of the SleepBandits System

With this system, we are able to identify the most popular and helpful experiments in order to further iterate on the list. Four experiments were completed by at least five users: exercising for 30 minutes (where 5 of 9 users who completed it improved their target variable), wearing socks to sleep (4 of 8 users improved), listening to white noise or music (2 of 5 users improved), and relaxing before bed (1 of 5 users improved). We are also able to track which target variable improved the most. For example, users whose goal was to reduce their time to fall asleep saw an average difference of 7 minutes (SD=8.4 minutes). In comparison, people fall asleep just 12 minutes faster when they take popular prescription pills, which often have side effects [147].

In general, 17 of the 31 participants (55%) who reached a result in the 4-night group improved their target variable when implementing the intervention, compared to 11 of the 17 (65%) in the 10-night group. This could be due to certain interventions not necessarily having a positive effect in the first few days, but leading to improvements in the long term. For example, earplugs can be uncomfortable at first, but we eventually get used to them and they improve our sleep overall. Future iterations of this system can identify such interventions and suggest a longer experimental duration.

Overall, all interviewees stated that they would recommend SleepBandits to someone wanting to conduct a sleep-related self-experiment, and nine of them said they learned something about how to improve their sleep. All interviewees who saw a result said that it was presented in a clear and understandable manner and that they were able to make a behavior change decision based on it, demonstrating the application of the **Comprehensibility** principle in SleepBandits. P7 stated that *"it's easy to use and gives you immediate feedback, and if you have this information, it could also help you change your habits, and that's powerful."* While the interviewees were a self-selecting group, most of them were conducting self-experiments for the first time, so their feedback was valuable and led to multiple suggestions for optimizing the approach.

5.6.7 Suggested Improvements

During the interview, participants were also asked how the app and experimentation approach could be further improved to better fit their self-experimentation needs. Three interviewees said they would like to see other people's success rates for the experiments because it would help them pick which one to undertake. One participant also wanted to see for how long other people conducted each experiment. His intuition was that experiments that lead to subtle changes should be conducted for longer. Two interviewees specifically said that it would have been nice to be nudged to do another self-experiment after the results of their current one reached a stable point. By far, most interviewees who never finished an experiment stated that they wished the application was able to track their sleep automatically, without having to turn it on every day, and without having to keep the phone on the bed.

In summary, the aspects that participants most appreciated in SleepBandits were those that gave them agency over their experiments: the ability to see a result early, as well as the ability to pick their own

experiment and target. Overall, participants thought that the flexible approach was suitable for novice self-experimenters, but that there are some changes that can make it more effective. We consider the implications of these findings in the Discussion section below.

5.7 Discussion

5.7.1 Shortened Duration of the Self-Experiments

At their core, the design principles behind SleepBandits aim to maximize user agency and find a balance between scientific rigor and how people run self-experiments as part of their everyday lives. We built on existing systems such as QuantifyMe, TummyTrials, and SleepCoacher to explore the challenges with the guided yet flexible approach. We compared a more traditional 10-night study with a shorter 4-night one. By applying a Bayesian approach, we were able to calculate experiment results after just four data points, and our study showed that users found the presentation of the Thompson Sampling results to be clear and concise.

To check how consistent results were over time, we focused on the users who had at least 5 nights in each condition, regardless of their assigned study group (N=19). We find that there was a 17% average difference (SD=11%) between the likelihood percentages after two nights in each condition and after five nights in each condition. The average difference in time to fall asleep was 2 minutes (SD=8). Thus, the results in general were relatively consistent throughout the experiment, but there was high variability between users. Shorter experiments might not be appropriate for every user, so future work can explore ways to identify users for whom a longer study might lead to more stable results. Overall, we find that by presenting the results while users are still interested in them, the system empowers people to make educated decisions about their health.

However, it is important to consider the ethical implications of systems that deliver such prescriptive results to users. In one group, SleepBandits calculated the results after just 2 days per condition, a duration that might be too short for statistical rigor. With the language and framing of the sentence we tried to convey that the result is just an estimate of the probability that represents how likely a condition is to be helpful, but it is crucial to consider whether such results could be misleading to the novice user. Future work should take this implication into account, as we need to further explore the role of technologies like SleepBandits.

5.7.2 Challenges in the Existing Design Principles

An important takeaway from this work is the need to balance the tradeoffs between design principles. SleepBandits was not designed to directly increase user engagement, but it provides users with agency over how much they adhere to the daily instructions, as well as how long they conduct the experiment. However, this agency comes with low adherence rates and drop outs during the study: as shown in Figure 5.7, most participants only tracked for one night. This shows that the principles we have focused on might not be

enough to encourage sustained user engagement. Further work is needed to refine SleepBandits with the help of insights from previous work on user engagement [56, 30]. To increase adoption, future systems should be tolerant towards low adherence and alleviate the stress on the user by adding features such as the ability to keep their phone on a night stand, using the microphone as a secondary sensor for detecting when the user is asleep, and graphs to visualize overall progress of the target variable.

5.7.3 Nudging Users Towards Most Helpful Recommendations

Overall, our analysis shows that participants most often chose a recommendation that was towards the first on the list, so apps for self-experiments should prioritize the ones that are most likely to be helpful for the largest number of people. This is in accordance with the Nudge theory [175] which states that the healthiest choices should be most readily available. Self-experimentation systems also need to set the most appropriate defaults for each experiment. For example, we had set "time to fall asleep" as the default target variable, but users often chose a different one. Perhaps the default target variable should change for each experiment. As one interviewee pointed out, systems could even nudge people towards the optimal length for each experiment depending on the expected result. Additionally, future systems can even identify cohorts of similar users and recommend experiments that other people comparable to the given user found helpful.

5.7.4 Increasing Agency over Result Details

One trend from the participant interviews was that they often conducted a self-experiment with something that they had heard about or even tried before. Thus, participants often already had a preconceived notion of whether it was helping them sleep better or not. They then either kept using SleepBandits until the results agreed with that preconceived notion, or they stopped using the app altogether when the findings did not match their mental model. This trend has important implications: how can we design future systems in a way that both helps the user keep an open mind about the outcome and enhances the credibility of the results? If users are able to view all the details of their experimental results, they might trust the system's findings more than their initial hunches.

5.7.5 Limitations

The current implementation of SleepBandits focuses only on interventions with minimal carryover effect. For experiments that have a carryover effect, future systems can apply an AB phase design and other lessons from [51, 162, 111]. In this work we found that users prefer to conduct shorter self-experiments, but two days per condition might be too short. Thus, going further with the development of the app, we will increase the minimum number of days in the system to 3 per condition (6 nights), as recommended by the standards in

[111]. Additionally, the findings here are based on a novel system released in the wild, but further research in self-experimentation is needed to determine optimal practices and design choices.

5.8 Conclusion

This work presents a set of design principles for systems for flexible self-experiments that focus on guided agency, scientific rigor, tolerance, and comprehensibility. We implemented this approach in SleepBandits, an integrated open-source system that includes a sleep tracking app, which allows people to run self-experiments on their sleep. Our experimental results are computed using the Bayesian approach of Thompson Sampling and are continuously updated to provide a tentative outcome and its certainty.

Based on data from 365 active users, we investigated which aspects of the approach are most enticing to users and what helped them successfully conduct a self-experiment and reach a conclusion. We find that people who conducted shorter self-experiments (4 nights vs 10 nights) were almost three times more likely to reach a result. We also discovered that users who received a likelihood in the range between 65% and 85% were convinced by the results, whereas those with very high or low likelihood scores chose to continue the experiment. We can build on the lessons learned from implementing these principles for self-experiments in sleep, and apply them to other domains such as mental health and physical well-being. As self-tracking becomes easier and more common, people will be able to benefit more from new statistical approaches that provide them with personalized recommendations.

Chapter 6

Self-E: Guided Self-Experimentation Beyond Sleep

This chapter presents Self-E, a system that helps people run self-experiments beyond sleep. This chapter is substantially similar to a paper currently in submission [49].

6.1 Introduction

While sleep was the first domain in which we implemented our self-experimentation approach, other significant aspects of our lives can also be tracked, such as exercise, pain, or nutrition [94, 177, 44, 45]. They may differ for each individual, even in the case of identical twins [75]. For example, even interventions widely believed to be beneficial and harmless, such as meditation, may not be advisable for some individuals [64]. Self-experiments allow an individual to vary aspects of their lifestyle in a controlled way and discover what works for them [119, 165].

Existing studies of self-experimentation systems have focused either (1) on the *self-tracking* aspect, which helps little in determining potential causation and generating actionable goals [115], or (2) on self-experimentation systems for one *specific purpose* [51, 47, 94], or for a specific population [39, 11], and thus lacked the flexibility required in real-life contexts for a general user.

From this background, two main motivations emerged for our study. First, previous studies have surfaced the need for a *general* self-experimentation tool that people can use to investigate potential causal relationships in their daily lives [11, 95]. Second, consumer self-tracking tools tend to be designed for more technologically literate populations as compared to the average user [131]. We were motivated to build a system for a novice from the general population to increase accessibility to technological systems. To alleviate the burden on

novice self-experimenters, we developed a tool that manages data analytics while providing a guided interface with simple instructions that take most of the manual steps out of rigorous self-experimentation.

Our contribution is twofold: (1) an system, Self-E, that allows people to conduct practical self-experiments that are more applicable to real-life contexts, and (2) an empirical study of how novices perceive and instinctively design their own self-experiments, followed by insights around their two-week use of Self-E for conducting real experiments in their lives. We also present critical considerations for the design and development of future self-experimentation systems for general users, including suggestions on how to better match the novices' mental models of self-experiments.

6.2 Related Work

Self-experiments have been of interest to a diverse group of research communities. Research has shown that self-tracking as an activity can encourage *reactivity* in people where they can monitor a behavior and decide if they want to change or maintain it [132, 109, 39].

Previous work has identified challenges that people face in self-experiments, such as tracking fatigue and flawed experimental design. Even "extreme users" with above-average background experience in self-tracking face common pitfalls such as tracking too many things, not knowing what to track, not knowing how to analyze or interpret data to extract insights, and having non-actionable and under-specified goals [115]. Karkar et al.'s TummyTrials system is based on the researchers' framework that could be applied to the design of a general self-experimentation tool [95]. In this dissertation, we built on this framework to outline guidelines for self-experiments that can be helpful for novices [47], and develop a system for self-experiments in sleep [52]. Lee et al. developed a prototype for self-experiments on pen and paper [115]. However, while systems that aim to alleviate certain challenges of the self-experimentation process exist, there is currently no system to fully guide novices through general self-experiments in a practical and low-burden way.

Overall, our study extends this work through evaluating a different approach towards conducting self-experiments, one that is aimed at conducting more *practical* self-experiments. Self-E is a guided self-experimentation system, built on top of existing self-tracking efforts, to help people in scientific self-discovery so that they can make more informed decisions.

6.3 Self-E System Design

Building on our work in SleepBandits, we develop Self-E, a system for self-experiments beyond sleep. Self-E is comprised of two components: an interactive Android smartphone application and a backend server that stores the data and performs the analysis. In this section, we outline our design considerations when developing the system, as well as the way it collects and analyzes the data.

6.3.1 Design Considerations

Self-E is a guided self-experimentation tool that uses self-tracked data to deliver individualized health and behavioral insights to the user. In contrast with existing self-experimentation systems, Self-E was designed without a specific experiment or user profile in mind (e.g., TummyTrials was built for those suffering from IBS, and SleepBandits was built for sleep). Additionally, Self-E was built with learnability in mind to minimize the burden of a novice looking to explore the potential advantages of scientific self-experimentation. Lastly, Self-E was designed to encourage user adherence during experiments, which required considerations around balancing experimental rigor with practicality for the user.

6.3.1.1 Guidance vs Generalizability

Extending existing work within the domain of automated single-case experiments, we designed Self-E with a focus on generalizability. Although the list of experiments presented to the users is preset, it is designed to be as general as possible. To provide clarity and ease of understanding to the user, every experiment is structured as a pair of a "cause" intervention (often a behavioral change, such as drinking caffeine) and an "effect" variable (such as sleep). Users are only allowed to conduct one experiment at a time, but the experiments can be customized to better reflect the individual's real-life goals.

6.3.1.2 Guidance at Different Levels of Experience

Another design consideration we addressed in Self-E was to strike a balance between providing guidance while not sacrificing user agency, so that novices of different experience levels could always find and learn something in the app. Self-E minimizes user burden by employing design strategies such as notifications to schedule interventions, data abstraction, and automatic analysis of the results to simplify aspects which are otherwise challenging even for experienced users [39]. To accommodate novices at different levels of learning, we incorporate both guidance and opportunities for more fine-tuned customization.

6.3.2 Architecture

The Self-E system is comprised of a backend server built in Python and a mobile client built in Android. When a new user begins using Self-E for the first time, they are taken through an on-boarding process that briefly introduces the concept of self-experimentation and its advantages. User profiles are created upon registration with an email and are stored in the backend server, and any configurations made or data tracked are sent to a backend server throughout use of the application. Storing this data in a server rather than locally allows users to change devices without losing their experiment history. Daily check-ins are sent to users' phones via the app notifications from the backend service.

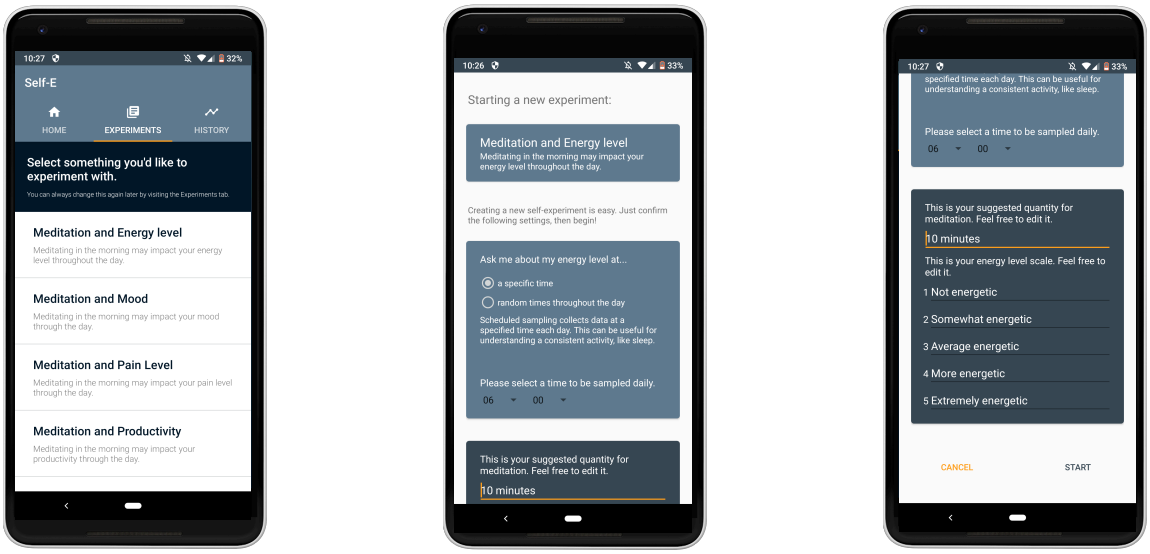
After registration, users are required to select an experiment from a list of starter experiments (shown in Figure 6.1(a)). This list was curated to highlight the advantages of self-experimentation, where the user only cares about their individual results, which can be impacted by complex factors such as genetics, lifestyle, physiology, social/cultural environment, etc. Each of the experiments was chosen in consultation with input from clinicians as well as backed by existing scientific research. The list contains 24 experiments, each of which is a combination of one of 5 interventions (meditation, physical activity, food & drink, walking, and hours slept) and one of 5 effect variables (energy level, mood, pain level, productivity, and sleep quality). The only combination that we did not include in the list was hours slept and pain level due to the lack of background research to support such experiments.

Once a user selects an experiment, they are taken to a configuration page, Figure 6.1(b). Here, they can customize the effect variable check-in time window, the check-in style (whether it will be sampled randomly in a time window or once per day at a fixed time), and the amount of the intervention they will be applying (10 minutes of meditation in Figure 6.1(b)). The flexibility that users have to change the check-in time was intended to allow them to fit self-experimentation into their schedule, to encourage a higher response rate and better adherence. Users can also customize the labels of the scale used to rate the effect variable (Figure 6.1(c)) because the ranges of experience with something like pain can vary among individuals.

Although we present many customizable features to the user, their values are set by default with the recommendations based on prevailing research. For example, an effect variable such as "mood" can be variable throughout the day, so for experiments measuring mood we recommend to users to select randomized experience sampling as the check-in style to gather sound data [72]. The scale labels for each of the effect variables are based on a commonly accepted scale for the given variable [127, 53, 29, 76], except for productivity, which does not have a widely accepted scale, so we choose to make it general (from "very productive" to "very unproductive"). These default values alleviate friction for a novice starting an experiment for the first time, while still affording the freedom for an experienced user to alter their in-app experience to fit their needs. Even though previous studies suggest varying frequencies of check-ins [181], we chose to limit the maximum number of daily check-ins to 5, as to avoid tracking fatigue [107].

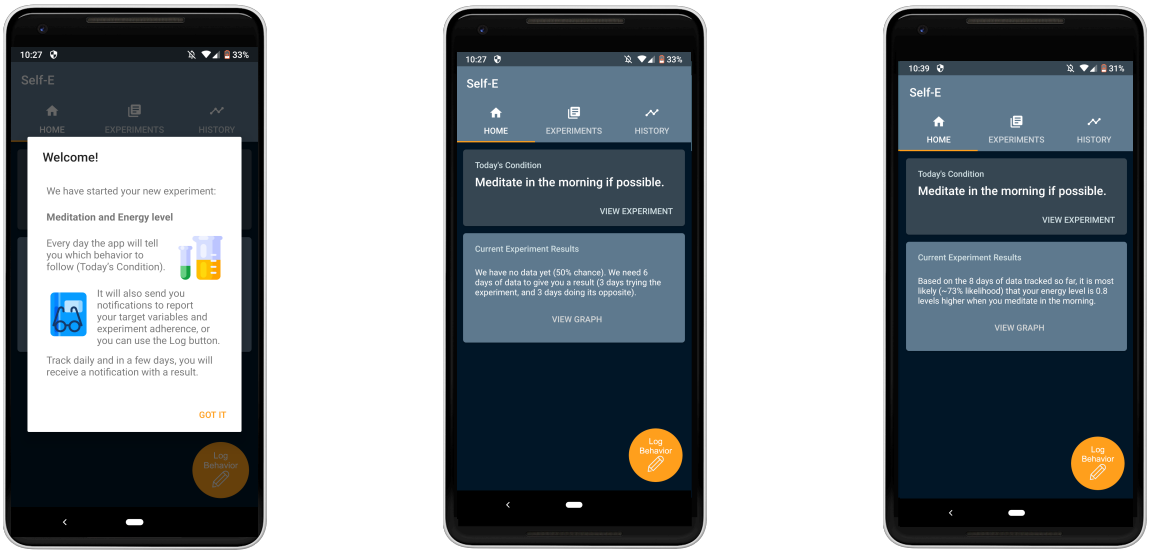
6.3.3 Experiment Flow

Once the user sets up the experiment, they are taken to the home screen of the app (Figure 6.2(b)). Users check in daily with the app at a fixed time for the intervention they are tracking (we chose 8pm as it was most appropriate for our list of experiments). Check-ins are initiated via a notification from the app, which takes the user to a pop-up dialog in the app that asks a "yes" or "no" adherence question for the intervention (Figure 6.3(a) or a rating scale for the dependent variable's effect (Figure 6.3(b)). Should a user miss the notification or decide to not answer the pop-up, they can log their data manually via the "Log Behavior" button on the home screen (Figure 6.2(b)).



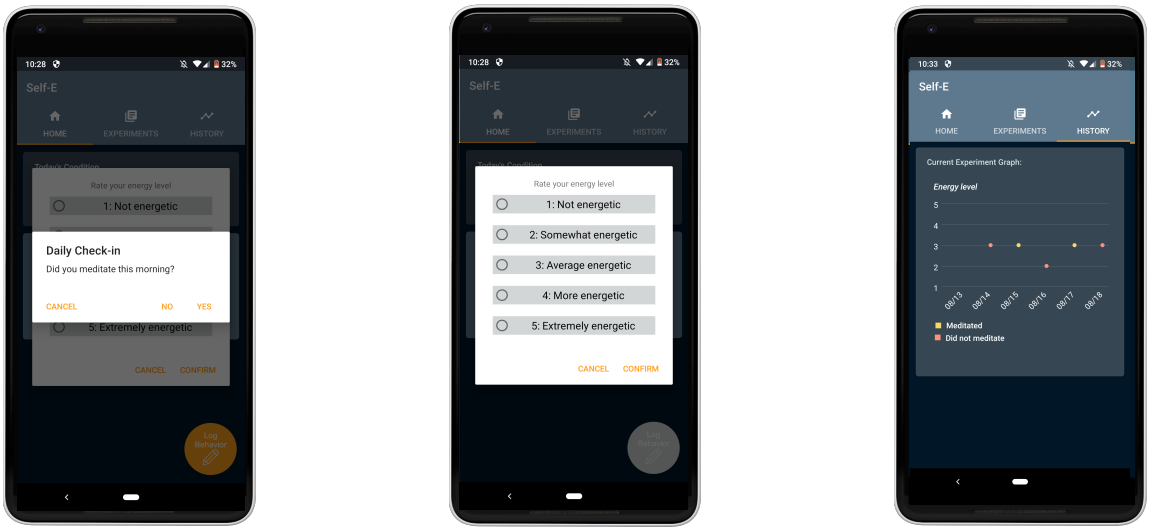
(a) **(b)** **(c)**

Figure 6.1: Self-E screens. (a) Experiment selection: users first select an experiment during onboarding, but then are free to change it at any point. (b) Experiment Setup: choose a time to be prompted and edit the goal amount. (c) Revise the labels of the scale.



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

Figure 6.2: Self-E screens continued. (a) Confirmation of the experiment setup. (b) Home: tonight's condition on top and current results below. (c) Home screen explaining that meditating in the morning leads to 0.8 increase in energy levels with a 73% likelihood based on 8 days of data.



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

Figure 6.3: (a) User prompt for the independent variable (the“cause / intervention”). (b) User prompt for the dependent variable (“the effect / the outcome”). (c) History of target outcomes: a new point appears on the graph for each day of tracking.

As a user continues to use Self-E, their data is displayed on a graph (Figure 6.3(c)), even when the user has not tracked for the minimum duration of the experiment yet. Self-E requires at least three data points in each condition (e.g. “meditate” vs “don’t meditate”) to calculate a result. This length is based on the minimum length required by the single-case intervention research design standards [111]. Similarly to SleepBandits, Self-E also applies an “as-treated” analysis [85], meaning that it only looks at which condition users actually followed on a given day, rather than whether they adhered to what they were instructed to do by the app. This type of analysis is recommended when the adherence rates are low [85]. This choice exemplifies the balance we attempted to achieve between experimental rigor and practicality.

6.3.4 Statistics

To estimate the likelihood that a given condition is helpful, Self-E implements Thompson Sampling, a Bayesian analysis approach identical to the one in SleepBandits. The approach is already described in detail in Chapter 5, section 5.5.2. Figure 6.2(c) shows an example result where the likelihood is estimated to be 73%: “Based on the 8 days of data tracked so far, it is most likely (about 73% likelihood) that your energy level is 0.8 levels higher when you meditate in the morning.

6.4 Method

Our study was comprised of three parts: (1) semi-structured initial interviews to gather the participants' background and experience with self-tracking and self-experimentation, (2) a 2-week diary study with daily voicemails, (3) and semi-structured exit interviews to discuss the participants' experiences with the Self-E app. Our goal was to understand how novices approach self-experimentation in order to improve future self-experimentation systems.

6.4.1 Participants

We recruited 16 people (4 male, 12 female) who lived in the local area by posting flyers in public spaces such as coffee shops and supermarkets and posting online in location-based communities such as Reddit and NextDoor. We selected participants who (1) owned an Android phone, (2) were over 18 years old, and (3) reflected a diversity of backgrounds in statistics and occupations. Participants' ages ranged from 20 to 70 (M=33.7, SD=15.4). Their familiarity with statistics ranged from none to expert or professional levels. The participants' occupations included 3 undergraduate students, 4 research assistants, as well as auditors, engineers, librarians, post office clerks, and baristas. Participants were compensated on a pro-rated basis: $10 for each interview, and $2 for each daily voicemail, for up to $30 for the duration of the study ($2 bonus if they completed all 14 days).

6.4.2 Study Procedure

To understand how people perceived, used, and evaluated Self-E, we first conducted semi-structured initial interviews. We then conducted a 2-week diary study, followed by an exit interview. Diary studies are high in ecological value since they allow in-situ data collection on the real experiences of users and have been used previously in human-computer interaction research for this reason [102, 172, 92, 82, 46, 137]. We followed recommendations to limit the duration of studies involving recall data collection to 2-weeks [181]. During the voicemails, participants were asked to share any challenges they encountered with the app that day, and whether they changed experiments and reasons for doing so.

In this study, we aimed to recruit a varied sample of participants in terms of backgrounds, technological and statistical literacy, and age. However, a broader sample might also tease apart what health conditions or variables such experimentation is most beneficial. While this was not the goal of our study, it is a worthy venue to explore. The main limitation of a diary study is the possibility of *demand effects*, where participants may over-observe due to their participatory status in the study [137, 166]. To minimize potential demand effects, we gave participants flexibility in terms of reporting through study design and app design. The Bayesian analysis method was able to calculate next steps and ensure validity of outcome data even if a participant failed to complete the daily intervention as guided by the app.

Figure 6.4: We conducted thematic analysis on the initial and exit participant interviews.

6.5 Findings

We conducted thematic analysis on the voicemails and interviews data. The researchers used open and axial coding techniques to come to a consensus on the emerging themes. Based on the thematic analysis, we identified the following themes: (1) previous impressions of self-experiments and instinctive way to conduct a self-experiment, (2) experiment length and continuing the experiment, (3) motivations for self-experimenting, (4) interpreting the results, and (5) self-experimental success.

6.5.1 Previous Impressions of Self-Experiments

6.5.1.1 Experience with Personal Tracking

Most interviewees had already tracked health and non-health aspects of their lives, such as physical activity, food intake, menstrual cycles, or personal productivity. Five participants reported using automated tracking tools or devices such as Fitbit wearables, and they were content with the ease of use. However, tracking fatigue and the high user burden of other tools was one of the main reasons for discontinued tracking.

Once they had some data collected, participants said they would sometimes look back at it and most often try to identify trends, a common type of insight [38]. Four participants mentioned that they looked at their past data to identify patterns to inform their future behavior (for example, if they had been eating too many sweets, they would try to temporarily reduce the amount). While these techniques were not forms of self-experimentation, they reveal that behavior change, where participants motivate themselves to increase one known “good” variable, was the dominant way of making use of self-tracked data.

6.5.1.2 Informal Experience with Self-Experiments

We investigated whether people had attempted to understand what affects their own individual well-being and how these strategies influenced them. This model of self-tracking focuses on *understanding* rather than motivating behavior change. The consensus among participants was that they should incorporate such personalized methods in their lives as much as possible. Two other themes that emerged were that (1) participants had tried informal self-experiments before to see if something was helping them personally, and (2) they saw the need to adapt to new interventions as their life and body changed.

Nine participants had carried out informal self-experiments with varying degrees of rigor. P13 tracked her skin condition for a month to identify what skin products worked for her. P1 started incorporating breakfast and found herself feeling better. However, a single day where she accidentally did not have breakfast was enough to convince her that breakfast was the habit that helped her feel better. In some cases, the participant stopped doing something by accident and found out that they felt better without it, such as P2 who found out that sugary carbonated drinks made her jittery. Additionally, three participants brought up the idea that their bodies were perpetually changing with time (both with age and on a monthly or seasonal basis), so the things that were once good for them might no longer be as beneficial. As P11 explained, *"you're always just kind of readjusting parameters."*

This mindset was in contrast with some participants' that conventional population-level advice was difficult to follow and did not apply to them. Some participants, like P9, said that *"I don't really listen to all the advice, because I don't know how to get started and keep myself accountable."* P5 felt that she *"did not feel like there's anything that needs to be fixed by changing things."* She explained how she had tried to follow a health tip, but was inconsistent in her efforts. She said, *"I tried my best to put my phone down 30 minutes before I want to be asleep ... it's not easy to implement."* People desired immediate results or would otherwise give up on the new change.

6.5.1.3 Perception of the "Self-Experiment" Concept

Most of our participants had not heard of "self-experiments" before the study, but they had an intuition about what they were (P10: *"doing an experiment except you do it on yourself to see if there's any changes based on what you changed in your life"*). Two participants mentioned that the word itself carried negative connotations: *"when I think of an experiment, I think of something that may be harmful,"* (P4), and *"I'm actually a little put off by it because it's not a super intuitive way to think about it. To me it's just an attempt to get better"* (P11).

6.5.1.4 Instinctive Self-Experiment Design

During the initial interview, we asked participants to design their own self-experiment so we could gain a better insight into how they think such experiments should be conducted. Overall, when asked to come up with a new self-experiment, novices stated a general goal like "improve sleep" and often had to be nudged to make it more specific (e.g. "wake up less") and to clarify how long they would conduct it for, how they would keep track of the variables, and how they would measure success.

Regarding the setup of the experiment, fifteen of the sixteen participants said that they would simply implement the change for a given period of time. Most commonly, that duration ranged between 2 weeks (N=6) and 1 month (N=4), but participants were split about whether they would implement it every day (e.g. drink coffee) or if they would do something a certain number of times per week (e.g. exercise). Either way, most of them explained that they would gradually build up to a certain goal amount of the intervention. P1, for example, wanted to see if meditating 10 minutes per day would help her mood. However, she said she would *"try to build my way up to 10 minutes, at two minute increments. Because if I try to do the full 10, I'm going to just stress myself out."* Similarly, P15 wanted to decrease her sugar intake, but would *"try to make it gradual rather than cold turkey—decrease per day for a while until it's down to quite a little."*

Only P11 discussed incorporating a baseline period going forward. For everyone else, the baseline was their life up until the start of the self-experiment. P11 was also the only one who mentioned multiple conditions (week 1 would involve just eating one cookie a day, week 2 was going to be about eating just honey, and week 3 would when she eats neither of those). Three participants brought up confounding variables in some way and tried to account for them in their designs, such as making sure the intervention was the only change happening during that time in their lives, or having both weekends and weekdays in the test phase.

Participants said at the end of their predetermined period they would look back and mentally compared whether there was an improvement over how they were feeling. In essence, participants would instinctively conjure an interrupted time-series for understanding the effects of a behavior change.

6.5.2 Experiment Length and Continuing the Experiment

6.5.2.1 Length of the Experiment

Once each user completes at least three days in each condition, Self-E calculates the result with the help of Thompson Sampling. Most participants thought that the default length of 6 days for the experiment was just enough, and they liked that they were free to continue the experiment for as long as they wanted after they got the result. Others, however, thought that experiments should be longer, provided that the data shows up on the graph earlier to give immediate feedback. P14 wanted a longer baseline period because *"If I had just the six days, it would have felt very skewed to me just because the last two weeks [were hectic]."* Similarly, P15 pointed out that *"two weeks was not really enough to really show"* the benefits of the experiment.

6.5.2.2 Compliance Rate

Every day Self-E instructed people which condition to follow that day. On average, the compliance rate to those instructions across all participants was 73%. People did not follow what the app recommended for several reasons: (1) they forgot to check their daily condition and unintentionally did its opposite (N=6), (2) they were stressed or too busy that day (N=3), or (3) because they did not want to (N=4). A common trend, when following the instruction was not possible, was for participants to *"do the best [they] could,"* meaning that they substituted the exact behavior that the app required with something as similar as they could afford. Often this substitution was necessary because many of the experiments required an action in the morning in order to identify its effects during the day. For example, P1 exercised at night instead because that was what her schedule allowed.

We also found that noncompliance is related to the perception of the self-experiment as a jumping off point towards a new healthy habit. P11, for example, chose to run an experiment with decreasing the amount of sugar to see if that helped her sleep better. However, she did not always comply with the app's instructions because she wanted to actually stop eating sweet things for a while: *"I'm not going to intentionally do something bad for me. I'm not going to be like nine Oreos and see how I sleep. That's a bad idea. I don't want to do that."* Two other participants brought up feeling guilty for not complying with the app's instructions. P5 elaborated that it was *"not like really guilt because that would be silly, but like a little guilt tearing me up because no, I didn't do it today."*

6.5.2.3 Reasons for Continuing or Changing the Experiment

We sought to understand why people chose to continue with the same experiment despite seeing the calculated result for it. Six of the eleven participants who reached a result said that they continued running the experiment because they were not sure if the intervention was really helping them or not yet, so they wanted to see if the result would change with more data points. Another common trend was that the app did not nudge users to change their experiments once they reached a result: it never prompted them to do so (as P12 mentioned), nor did it suggest any related and interesting experiments to move on to (as P9 would have liked). Two participants explained that they actually wanted to change experiments, but were not sure what would happen to their existing data.

Only three participants changed experiments during the course of the study. They explained that they wanted to try new experiences after getting a result for the first one. P4 and P11 both changed theirs because they were bored of the intervention. As P11 explained, it was because she was *"the kind of person who does the thing for two days, [gets bored] and then does another thing."* P13 wanted to restart his experiment because he noticed that some of his data was faulty, so he changed experiments to a different one and then immediately changed back to the one he was originally doing.

6.5.3 Motivation for Self-Experimenting

But why self-experiment in the first place? Participants brought up two reasons as to why they would conduct a self-experiment on their own: (1) a desire to identify what general advice works specifically for them, or (2) as a starting point to implement a behavior change and actually stick with it through self-monitoring first. Six participants said that they chose their specific experiment in Self-E because they had previously heard that it could lead to health benefits so they wanted to see if it helps them and by how much. P14 said that *"It's about finding what actually works for me and what I can work into my schedule versus just like overall science saying, this thing helps with these."* P9 elaborated that *"People hear about all these things they should do. And it's kind of like a weight on your shoulders. But maybe it's more about self-knowledge – that really helps people because they realize what works for them."* Others had specific goals which they hoped to achieve through self-experiments, such as improving sleep quality, and would try the most enticing experiment aimed at those goals.

6.5.4 Interpreting Results

The graph and the sentence that summarized the result of the self-experiment were mostly clear and easy to understand for users. P2 particularly liked that the app *"organized everything in my brain"* and P14 said that *"what was ultimately more helpful was the result sentence because it summed it up and the data underneath it: the likelihood and the points difference."*

Overall, some participants revealed that they trust and agree with the results that Self-E was showing them (P2, P6, P8, P10), especially when it confirmed how they already felt (P5, P10, P15). The reasons for not trusting the results revolved around three main themes, all related to confounding variables. First, they felt that **the effect of the intervention was too negligible** compared to the effects of other things in their daily life (P15). P11 elaborated that there were a lot more confounding variables in her life that were affecting her at the time, and the one cookie she was eating a day was not as impactful as the rest: *"So like reporting how my sleep was and whether I eat a cookie or not, like literally one cookie.. or what if I took a nap that day? or what if I run every other day... like there's so many variables that, I assume anecdotally, have a stronger effect on my sleep."*

Second, if **the results contradicted a previously held belief**, users always brought up potential confounding variables that could have been the reason for the unexpected result, such as the *"extreme data points"* for P14. People expressed skepticism of the app and their own feelings due to preconceived ideas of potential confounding variables (P12, P15, P5, P11). P15, for example, did not trust the results because they were too early: *"It was kind of unexpected, because I was thinking oh you know maybe it'll be lower energy for the first couple of days, and then overall higher energy. But it just said it was overall less which is little bit unexpected to me. But maybe it's just something that needed more time."*

Third, **the high likelihood percentage or its drastic fluctuation** over the course of the experiment raised

suspicion. P10 and P13 did not feel that the likelihood was convincing because it was unclear how it was calculated. Similarly, P12 and P15 were skeptical because the result either fluctuated too quickly (from 66% to 58% to 97%) or was too high on the first day (96%). P5, P9, and P11 noted that the fluctuations were likely due to unclear distinctions between the two experimental conditions. P13 thought that he was not varying his hours slept enough to lead to a difference, and P5 said that on the days she was not meditating, she was reading before bed, but she realized over time that *"it had the same effect on sleep."*

Related to that, we find that most participants pointed out that a convincing likelihood percentage would be between 65% and 85%, but it would also depend on the level of effort and time required to conduct the interventions. P14 said *"it depends on how much the time investment is. If you go running for an hour, you have a 70% likelihood of improving your mood, that would be different than if you meditate for five minutes in the morning. It has a 70% chance to improve your mood, but it's different investments."*

6.5.5 Self-Experiment Success

Overall, twelve participants (75%) found Self-E useful for their self-experimentation needs. P2, for example, explained that *"The app helped me in self-diagnosing when the doctor wasn't helping."* She *"learned that it wasn't the food that I was eating, it was more of physical activity that was causing my inflammation. Because I'd be like, Hey, I didn't eat my trigger food, I still hurt. What's going on?"* For P12, Self-E served a difference purpose than the usual self-tracking apps: *"I feel like for some goals, I try something new and I don't often stick to it. With this, there is the initial learning curve, but then it kind of pushes you to keep on with it until you kind of get past that point. And then you can figure out if you actually want to keep doing it."*

Almost all of our participants said they would recommend Self-E to a novice who is interested in trying to see if an intervention works for them. Some specific reasons they list are that it provides the motivation to stick with one thing (P7, P16), and that it *"provides a minimal amount of structure what would still give you some flexibility to play with, while making it drastically easier to track it all"* (P14). Most participants (75%) said that they would continue using Self-E after the diary study. Some reported wanting to use it potentially at a different time of their life, such as when they were less busy (P7) or *"if [they were] curious about something else"*(P11). The reasons they brought up for why they would use it were that *"it feels more scientific than saying I have a feeling that it works"* (P10), that it was a *"good way to keep data about myself while doing something new"* (P5), or that there was another experiment on the list that they were interested in trying out (P5, P13, P14). For P2, Self-E motivated her to continue self-tracking habits in her daily life and to explore other tracking apps for different purposes.

The aspects of the Self-E app that participants thought were most helpful were: (1) the preset list of experiments because *"everyone could find something"* (P12), (2) the structure that the app provided for the experiment itself, and (3) the low user burden of the app.

Pre-set list of experiments. Overall, most participants appreciated the list of suggested experiments. P14,

for example, said *"I really like that it had setup experiments because it gives you a place to start if you're like 'I want to improve my health.' But if you just Google 'improve your health,' it'll tell you 1,200 different ways to do that. And that's not particularly helpful."* While the list was one of their favorite aspects, six interviewees said that this feature could be further improved if it also had the option to create your own experiment once you finish your first app-approved one. Some additional enhancements that participants suggested were the ability to see others' custom experiments and the introduction of more challenging and longer-term experiments.

Structural guidance. Most participants agreed that the app was helpful in running a self-experiment because it guided them through: (1) the choice of the experiment, (2) the process of what to do and when to do it, and (3) the input and analysis of their data to provide *"credible results"* (P10). P9 elaborated that *"I don't fully trust myself to design a rigorous self-experiment."* P12 and P15 expressed how they would not have incorporated aspects of scientific rigor on their own, such as using experience sampling methods for collection of user data. P7 said *"the fact that it was constantly telling you what you had to do, it was like having a friend that you know, motivated you to do something we didn't want to do but think you had to do. It was very helpful. You don't get the effectiveness like that with a human, you know?"* P10 expressed that *"saying 'Yeah, I feel better' is less credible compared to using the app, seeing the actual difference, how it actually improves your mood, is definitely better than doing it on my own."*

Low user burden. Another aspect of Self-E that participants appreciated was the fact that it required a low level of effort. Six participants expressed that while they might be capable of conducting such experiments on without the app, it would be too tedious or challenging. P5, for example, said that having all the data in an app, *"as opposed to remembering myself, lets me do something for a longer period of time, when I wouldn't be able to keep track in my head."* This sentiment was echoed by P14, who reported that Self-E made self-experimentation easier since it *"kind of does everything for you."* In particular, participants found the manual logging convenient since it was pushing a single button. They also liked that the notifications gave them structure and had options for random sampling throughout the day. P14 said that *"I work in retail, especially, one check in a day would probably not have given a good average."* Sixty percent of participants liked how brief the data entry questions were, and thought the scales for reporting the effect variables did not need any modification.

6.5.6 Opportunities for Future Improvement

This study surfaced three main opportunities for improving our approach to practical self-experiments. First, we find that users would like to build up to a given health behavior by taking small steps. Both the instinctive way participants design their own experiments and the fact that six of them decreased the goal amount while setting up the experiment suggest that people would like to start with an incremental improvement.

Participants also expressed a desire for more guidance from the app, in three key areas. First, they kept

bringing up confounding variables that might have affected their experiments, and at the same time did not always comply with the randomized schedules from the app. The practical approach could be improved by explaining why randomization is important and how exactly it helps with the confounding variables. Second, some participants were not even sure how to go about applying the suggested interventions in their lives. For example, it would have been helpful to either have a short tutorial on how to meditate or to point people to a specific app that can help them with meditation. Lastly, participants wanted to be able to also create their own experiment after completing the initial one. Future self-experimentation systems can implement this feature while also guiding the user towards making scientifically sound choices.

The third opportunity for improvement is related to faulty data. Some participants mentioned that their experimental data definitely had outliers, which led to them not trusting the results. A possible solution brought up by them was for the app to let users delete the data points, or even just label them as outliers and account for that in the analysis.

Participants also brought a few usability improvements and potential future features that would have made Self-E more helpful for them. One common piece of feedback was that the effect size phrasing was confusing (e.g. *"0.3 levels higher"*), as it was unclear what *"levels"* referred to. P10 and P12 suggested a bar-graph where the average of each condition is shown and the y-axis ticks are the levels of the given scale. Some participants also would have preferred the scale for the effect variables to allow for more nuance through a continuous or a 10-point scale.

Participants even suggested additional features to enhance their future experiences in self-experimentation, such as the ability to receive reminders each morning about which condition to follow. P2 wanted the ability to read more about the interventions before changing experiments, and P11 wanted to learn more about the science behind self-experimentation in general. P6 wanted the ability to be able to revise the timing of the sampling after her experiment began. P4, P10, and P9 needed a clearer indication of what has been logged so far. P4 wanted a more visual and interactive interface of the app that would motivate her, give her rewards, and push her limits. She said that she was hoping the app would be more like a coach/personal trainer: *"let's make a plan, let's push your boundaries, let's do another mile today."*

One major challenge that we did not foresee was people having trouble with the notifications. P3 and P4 did not know how to pull down the notifications banner on the Android mobile interface to tap the notification and get to the app. This was problematic because the only way to log the effect variable was through tapping that notification. Thus, for the purposes of this study, we released an update of the app that let users enter both the effect and intervention information through the *"Log Behavior"* button on the Home page (Figure 6.2(b)).

Most participants were enthusiastic about the idea of sharing their results with friends and family. Interestingly, P14 and P5 saw self-experiments as a potential *"bonding experience"* to see how a given intervention affected people they know: *"I would compare my result, to see the difference between two people doing the same thing... With my brother, we don't live in the same place"* (P5). P14 wanted to share her results with online communities of people who are interested in improving the same effect variable (pain level). She

said that *"People tell us so many garbage things that don't work. I cannot tell you how many times people told me to do yoga."* Other participants did not want compare or share results because *"everyone has their own rate"* of exercising (P6) and they would not be surprised if other people had different results (P15, P16).

6.6 Discussion

Our study revealed insights about how novices intuitively design self-experiments, and how they interact with an app that aimed to guide them through the steps of a practical self-experiment.

6.6.1 Instinctive vs Scientific Experimental Design

6.6.1.1 Building up to a Goal Amount

Most of our participants brought up the notion of building up to a goal amount of the intervention they were trying to implement. The participants' mental model is in accordance with Fogg's Tiny Habits behavioral model, in which one should start a behavior change with the smallest increment possible [68]. However, existing self-experimentation systems such as TummyTrials [94], QuantifyMe [162], SleepCoacher [51], and even Self-E, are not designed to handle this behavior. While we allowed participants to revise the default goal amount at the beginning of the study, they were not able to change it without restarting the experiment. Therefore, future systems for self-experimentation should take this into consideration.

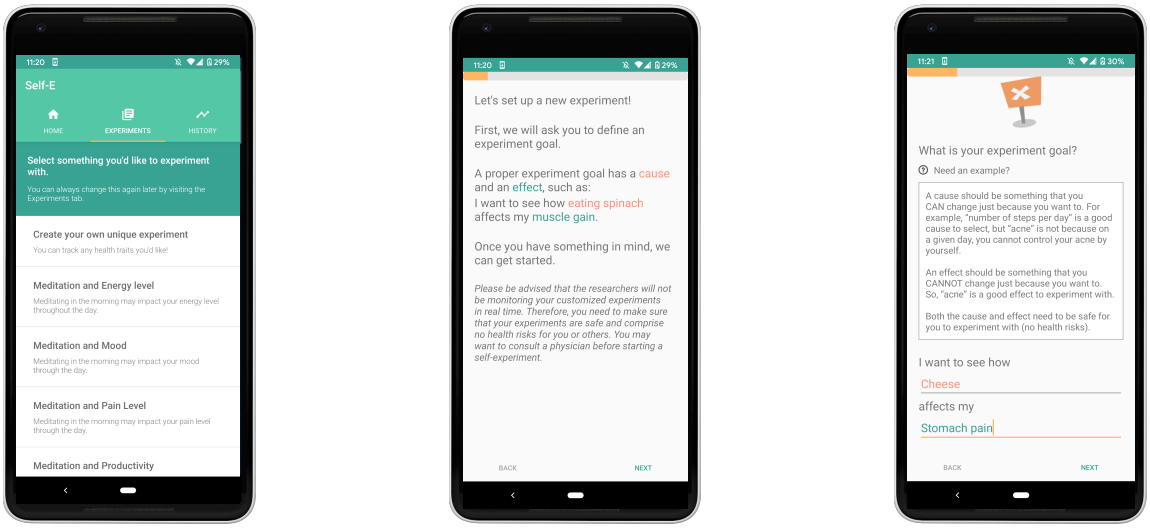
6.6.1.2 Using it as a Starting Point for Behavior Change

Related to building up to a goal amount, participants also used their self-experiments as a way to start implementing a behavior change that they had been thinking about for a while. People's mental models did not match Self-E's attempts to nudge them to implement the intervention on some days, and to avoid it on others. The participants saw the experiment as a catalyst that finally helped them act on the new behavior, and they were reluctant to stop doing it. For them, it was simply easier to continue doing a behavior than have to check and switch between experimental conditions.

Future self-experimentation systems could take this into consideration by giving themselves the role of providing the initial inertia for users to commence a behavior change. Larger block sizes may be a useful parameter for users who wish to balance between the convenience of inertia, and the efficiency of number times the condition is randomized to gain more probabilistic confidence in the causal outcome.

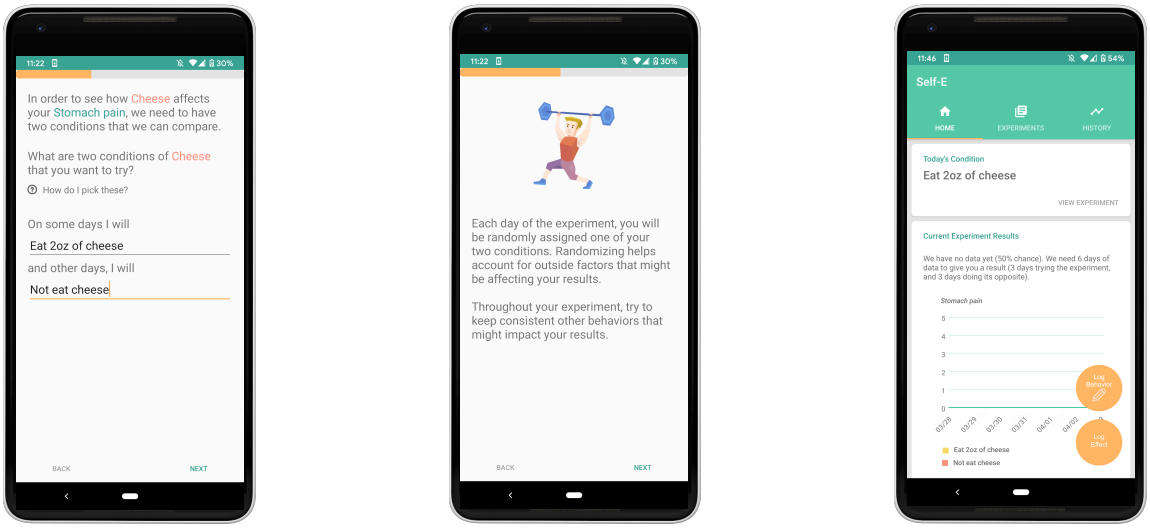
6.6.1.3 Intuitive Assumptions vs App Results

Most participants in our study also had preconceived notions about whether an experiment would be helpful to them before they even started it. Our findings suggest that they thought the intervention would be beneficial,



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

Figure 6.5: Self-E customized experiment flow. (a) The option to create your own customized experiment is listed as one of the options on the Experiments tab. (b) First screen explaining the need for a cause and effect. (c) Users can define their own cause and effect and the guiding text only appears if they tap on the "Need an example?"



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

Figure 6.6: Self-E customized experiment flow continued. (a) The user has to pick two conditions by filling in the blanks. (b) A screen explains that we will randomize between these conditions every day. (c) On the home screen, the new condition for today is shown at the top.

so when Self-E presented results that contradicted that belief, users found excuses for confounding factors that could have interfered with the data. One possible way to account for this bias while calculating the results is to allow the participants to feed these assumptions as the priors in the Thompson Sampling approach. That way, if participants are very certain something might be helpful for them, the more helpful condition will have a higher prior likelihood than the other condition, which in turn will be reflected in how often they are asked to try each condition. However, we need to be cautious because feeding assumptions as priors might lead to potential biases in the results.

6.6.2 Implications for Self-Experimentation Technology

Our findings demonstrate that participants are aware that their data is sometimes faulty or contains too many outliers, so future self-experimentation systems should be able to handle such cases. For example, users might want to be able to exclude certain data points from analysis. However, this might also lead to biased results, so perhaps systems should let users label extreme data points and then take these data points into account when calculating the result.

While participants thought the original list of experiment was helpful, they also wanted to conduct their own customized ones. Future systems could incorporate this as a feature and increase user agency over time as they complete more self-experiments. However, a completely open self-experimental design might lead to faulty experiments, so such systems should guide users while also providing them flexibility. We have developed a preliminary prototype of that functionality in the Self-E app. To decide on the design, we conducted numerous user experience iterations and consultations with behavior change clinicians. The design of the user flow is portrayed in Figure 6.5 and 6.6.

6.7 Conclusion

This work presents Self-E, an app that guides novice users through the steps of a self-experiment. We conducted a two-week diary study and interviews with 16 participants who used Self-E to conduct general self-experiments in their own lives. We identified several aspects of the system that participants found most helpful, such as the preset list of experiments and the guidance on what to do each day of the experiment. Our qualitative study sought to expand on the field's understanding on how people with little experience in personal analytics perceive self-tracking and self-experimentation with the help of mobile tools. We find that the instinctive way novices conduct self-experiments does not match the implementations in existing systems, so future research can explore ways to help people conduct such experiments in a more intuitive way.

Chapter 7

Investigating the Effectiveness of Cohort-Based Sleep Recommendations

This chapter presents an empirical study investigating the effectiveness of cohort-based recommendations. It is substantially similar to [50], completed as a result of my internship at Microsoft Research, where I was responsible for running and analysing the study, and a majority of the writing.

7.1 Introduction

All the work described in this dissertation so far has been focused solely on each individual's data, using only their records for analysis. However, in this chapter we attempt to leverage the data of groups of similar users in order to identify behaviors that seem to be helpful for a given cohort of users. Overall, existing sleep tracking mobile apps and devices provide mainly summary statistics. Additionally, these systems use only one individual's data, making them prone to over-fitting and omitting valuable context that can be found in other people's sleep patterns.

In this chapter, we build on current sleep tracking methods to leverage data both from each individual and from similar users to provide actionable cohort-based recommendations. Using a collaborative-filtering technique, we identify a cohort for each participant, using a large real-world dataset. We conducted a 4-week study (N = 39), collecting six sleep quality metrics and exit questionnaires from three groups of participants who either received: (1) no recommendation, (2) a general recommendation, or (3) a cohort-based recommendation. We aimed to answer the following research questions: (1) How can we build off of existing collaborative-filtering frameworks to generate cohort-based recommendations for sleep, given a large dataset of other users?, (2) What are the limitations of recommendations based on collaborative-filtering that surface

when people are asked to follow them in their everyday life?, and (3) While grounding our work in existing theory, what design hypotheses can we provide for future systems to overcome these limitations?

The main contribution of this work is the empirical study investigating the effectiveness of cohort-based recommendations by applying collaborative-filtering techniques to the domain of sleep research and leveraging a unique, large-scale real-world dataset. We performed a thematic analysis on the data to learn what makes cohort-based recommendations helpful or not helpful, and we identified design hypotheses for the future cohort-based sleep recommendations. We found that users prefer to be given more control over what their cohorts are based on, and that in order to provide helpful suggestions the recommender system should be able to take into account each user's constraints related to their occupation, schedule, and lifestyle.

7.2 Related Work

7.2.1 Non-Clinical Sleep Studies to Support Healthy Sleep Behaviors

While a smartphone app usually does not require any additional hardware, we chose to use the Microsoft Band for this study as it gave users the opportunity to also track daily exercise. Previous research also points out that it is important to design technology that helps people become more mindful of their sleep, while not imposing impossible sleep goals [36]. To evaluate SleepTight [37], for example, the researchers conducted a four-week study, to determine whether the system with the widgets enabled a higher sleep diary compliance rate than the one without. Our work employed a similar framework by evaluating the effect on sleep of two types of recommendations and a control condition with no recommendations.

7.2.2 Sleep Recommendations

Prior work shows that self-tracking combined with suggestions can improve sleep [48]. However, it is unclear how the few systems that provide recommendations even generate them. Fitbit and Jawbone's insights are limited to general trends and comparisons of the user's sleep to that of others of similar age and gender. However, none of these trackers provide actionable recommendations that have been shown effective for people similar to the user.

Notably, two systems have been developed to provide actionable sleep recommendations beyond simple summary insights. ShutEye [18] focuses on displaying sleep hygiene guidelines on a user's mobile phone home screen. However, these guidelines are based on the general population and might neglect individual differences like those between different age groups [5]. The two systems for recommendation in sleep described in previous chapters, SleepCoacher [51] and SleepBandits [52], also focused only on individual data. Their approach did not use information from other users that might provide helpful insights or motivation for improving someone's sleep.

7.2.3 User-Focused Recommender Systems

To provide cohort-based recommendations, this work employs a recommender system developed to identify groups of similar users. The approach of making predictions or recommendations about a user based on data from other users is known as collaborative filtering [151]. Collaborative filtering is most often based on 'neighborhood models,' in which the unknown ratings from a user are estimated based on ratings from similar users [151]. Neighborhood models are popular because of their simplicity and the intuitive reasoning behind their recommendations [151]. Another benefit, in relation to sleep, is that given the right parameters, these models can be used to generate a recommendation for new users [151], even before any actual sleep tracking data has been recorded. Our cohort-based approach is a type of a neighborhood model, and thus can be further fine-tuned and improved.

Pu and Chen's user-centric framework for evaluating recommender systems emphasizes the need for minimizing the interaction effort for the user while producing useful and trustworthy recommendations in a transparent way [144]. To address this need, some recommender systems focus on getting users' feedback in order to better understand their preferences and provide more accurate recommendations. In particular, a critique-based recommendation system first suggests a few options based on users' current preferences. Then, the user critiques those suggestions, and then the system generates new ones based on the critiques [33]. Such systems help users build a preference profile, but they are task-specific and rely on the content [80]. Alternatively, studies have shown that users are more satisfied when they are given the chance to directly manipulate the attributes as in [80, 23].

In our study, we chose to focus on collaborative filtering techniques to elicit actionable sleep recommendations based on the data from other users. However, as discussed in Section 7.7, future work could explore how critiquing-based systems can provide recommendations that incorporate users' preferences.

7.2.4 Health Recommender Systems

Health recommender systems have been focused on personal health record systems (PHRS), which centralize each person's electronic health data and allow health professionals to access it [174]. PHRS contain too much expert-oriented data, which leads to information overload for the regular user [188]. Thus, health recommender systems have been developed to provide laymen-friendly information to users to better understand their own data [184, 189, 188]. Even outside of PHRS, recommender systems have been used mainly for information filtering [163]. However, such systems have yet to be deployed in the domain of sleep.

7.3 Method

7.3.1 Dataset

Our dataset contained the real-world sleep records collected by the Microsoft Band (MS Band), a wrist-worn fitness tracker, first released in 2014, followed by an updated hardware version in 2015. While there are a total of 40 million sleep records, this study used data gathered between May 2016 and May 2017 (about 1 million users). Overall, MS Band users were from diverse backgrounds, age ranges, gender, occupations, and nationalities. Recent work by Althoff et al. shows that MS Band's sleep data is representative for the general population as the measurements match published sleep estimates [8]. In this study, we used the data only from users who reported being between 18 and 65 years old, were between 50 and 80 inches tall, and weighed under 250 lbs. These criteria are commonly used in sleep and exercise literature [185, 182]. Following similar studies, we also exclude any sleep records with a duration of less than four hours or more than 12 hours [183, 8].

In order to estimate whether the user is asleep or awake, the MS Band uses internally validated proprietary algorithms based on the 3-axis accelerometer, the gyroscope, and the optical heart rate sensor in the Band. Users can manually start and stop sleep tracking by tapping the "Sleep" tile. Otherwise, their sleep is detected automatically. The sleep measurements are:

- Duration: time spent *in bed*
- Sleep Time: amount of time *actually sleeping* (different from time spent in bed)
- Sleep Efficiency: ratio between duration and sleep time
- Time to Fall Asleep
- Number of Wakeups
- Bed and Wake Times
- Amount of Restful/Restless Sleep
- Sleep Recovery Index

The Sleep Recovery Index is calculated via a proprietary algorithm. The value is between 0 and 100, and each quartile is mapped to a score between 0 ("poor") and 3 ("optimal") based on sleep quality.

According to sleep literature, intense exercise for 30 minutes three times a week improves sleep [16]. Exercise data including biking, running, and working out is also collected by the MS Band once users tap on the corresponding activity tile. The Band tracks heart rate using an optical sensor. We compute overall intensity of the activity based on the rate of calories burned (using a proprietary algorithm).

7.3.2 Study Design and Participants

We designed a between-subjects exploratory study to evaluate the effectiveness of cohort-based recommendations and to learn how to make better cohort-based health recommendations in the future. We randomly assigned the participants to one of three conditions. In condition 1, the control "no-recommendations," participants tracked their sleep every night for a month and received no recommendations. In condition 2, participants received a general recommendation halfway through the study. In condition 3, participants also received a recommendation halfway through the study, but it was personalized based on the cohort of similar users. In both conditions with recommendations, participants were sent daily reminders for the last half of the study asking them to follow the same recommendation every day. We describe the algorithm for generating recommendations in Section 7.4.

We recruited participants internally in a large software and services technology company from a pool of beta testers. Due to the remote nature of the study, participants were distributed across the US. Only participants who did not have a self-reported sleep disorder, who could track their sleep for at least 14 nights of the study, who had access to a MS Band (version 1 or 2), and who were not traveling across timezones were allowed to participate in the study. Traveling across timezones affects sleep patterns and causes jetlag, which can interfere with sleep quality and the results of the study [160].

In the initial recruitment email, we asked people to fill out a pre-study questionnaire. which was completed by 77 people. However, five of them were not allowed to participate because they self-identified as having a sleep disorder. We gave out 13 replacement MS bands to the first 13 respondents who said that they either did not have access to one or that theirs was broken. Halfway through the study, only 66 people had functional Bands sending data to the database. Thus, only those people were assigned to one of the three conditions described above. They had various occupations: engineers, program managers, managers, and others such as a sales executive and a business analyst. In a way, this represented a controlled workplace-based cohort, but our investigation focused on behavior- and demographic- based cohorts.

7.3.3 Study Procedure

The pre-study questionnaire had questions from the Epworth Sleepiness Scale (ESS) and the Pittsburgh Sleep Quality Index (PSQI), both of which are commonly used in research studies to evaluate sleep quality [27, 90]. The survey also included questions related to sleep quality and lifestyle.

Participants across all conditions were sent a daily email at the same time asking them to rate their sleep quality between 1 and 5, with 5 being their best sleep ever and 1 being their worst sleep ever. Halfway through the study, the participants in the two conditions with recommendations began to receive recommendations via email. Their daily emails also included a question about whether they followed their recommendation on the previous day. Each week, all participants were also asked through a questionnaire if anything unexpected happened that week that might have affected their sleep, as well as how many times they exercised vigorously

Table 7.1: Questions asked in each of the four types of questionnaires.

Questionnaire	Questions
Pre-study questionnaire	Demographics, ESS, PSQI, other qualitative questions
Daily questionnaire	Sleep quality for previous night. If they received a recommendation: did they follow the recommendation (included a reminder of what their recommendation was)
Weekly questionnaire	Did anything out of the ordinary happen, how many times did they exercise this week
Post-study questionnaire	ESS, PSQI, other qualitative questions

for at least 30 minutes. All the questions in the questionnaires are summarized in Table 7.1. At the end of the study, participants filled out a questionnaire similar to the pre-study one, with questions from the ESS and PSQI, and additional ones about the recommendations and their experiences during the study.

7.4 Cohort-Based Recommendations

7.4.1 Finding Users with Similar Profiles

To generate cohort-based recommendations, we first identified a cohort of users similar to each participant. Then, we detected which dependent variable was affecting this cohort's sleep the most.

7.4.1.1 Features for Cohort Selection

We collected the participants' height, weight, and gender. Due to privacy restrictions, we did not collect their age. According to the National Sleep Foundation, body mass index (BMI, calculated based on a person's height and weight) plays a vital role in sleep quality, as obesity can cause undiagnosed sleep-disordered breathing [6]. Previous studies have shown that gender differences also affect sleep [178, 25]. We also asked participants about the average number of days per week they exercised vigorously for 30 minutes or more. Previous research has shown that exercise affects sleep quality [170, 179]. While other external factors such as caffeine consumption can affect sleep, we chose to focus on exercise because it can be tracked with the MS Band. Lastly, we asked participants to rate their overall sleep quality between "very good," "fairly good," "fairly bad," and "very bad." We calculated the average Sleep Recovery Index of each user in the MS Band dataset and mapped the score to a scale of 0 to 3, matching the sleep quality rating from the study participants. Thus, a score of 3 would map to "very good," and 0 would be "very bad." We also calculated the average number of times per week each user exercised vigorously for at least 30 minutes. Thus, the features used to identify a cohort of similar users were: height, weight, gender, number of days per week they exercised vigorously for at least 30 minutes, and their overall sleep quality (as a proxy to the Sleep Recovery Score).

Table 7.2: Template of the recommendation text in each of the four categories.

Category	Recommendation Text
Consistency	People are most affected by how consistently they go to bed and wake up. Doing both within half an hour of a consistent time, including on weekends, could improve sleep. During the last two weeks, you went to bed anywhere between *XB* pm and *YB* am, with *ZB* pm as the most common. You woke up between *XW* am and *YW* am, with *ZW* am as the most common.
Winding down	People are most affected by how much they relax before bed. Setting a comfortable pre-bedtime routine, such as taking a warm bath or shower, or meditating, could improve sleep by making it easier to fall asleep. People typically take *MP* minutes to fall asleep. Typically, it takes you *MY* minutes to fall asleep.
Exercise	People are most affected by how much they exercise. You said you exercised *X* times per week. Exercising for at least 30 minutes three times a week could improve sleep. But vigorous exercise close to bedtime may cause difficulty falling asleep.
Duration	People are affected by how much they sleep at night. They typically slept *HP* hours and *MP* minutes. During the last two weeks, you typically slept for *HY* hours and *MY* minutes.

7.4.1.2 Nearest Neighbor Search

We used the features described above to identify a subset of users who had similar demographic and habit profiles. This was achieved by a nearest neighbor search in k-dimensional space. First, we constructed an anonymous k-d tree from all users using normalized versions of height, weight, gender, frequency of exercise, and average sleep quality. Given the large number of users in our dataset, the height, weight, and BMI of the nearest neighbors that our algorithm returned were almost identical to those of the study participant. Future work can evaluate whether BMI is more effective in identifying appropriate nearest neighbors for specific groups of users, such as those with a high BMI who are particularly susceptible to poor sleep [55].

Next, we used a fast nearest neighbor search library for Approximate Nearest Neighbor Searching (ANN) [129], implemented in an R package called FNN [21], to identify users with similar profiles in the k-d tree. We experimented with a wide range of numbers (2 to 50) of nearest neighbors. We created a simple interface that took as inputs the 5 variables we used for identifying the cohorts, and checked how long it took for our system to return the set of nearest neighbors for each number between 2 and 50. While the MS Band dataset included millions of sleep records, we wanted to use a small enough number that could be attained easily in future studies. In case our cohort-based framework was successful, we wanted to make sure that it was reasonable enough to be used in real-life scenarios where users would perhaps input their information in an online system and get a recommendation in real time. We found that $N = 5$ is optimal for real time scenarios (based on performance and quality considerations) and $N = 30$ for offline recommendations in this study. We picked 30 because that is a common sample size used in user studies, and generally leads to the possibility of a more robust statistical analysis for $N \geq 30$.

7.4.1.3 Selecting the Recommendation

Once we had identified each participant's cohort of nearest neighbors, we determined which recommendation would be most appropriate. To do so, we calculated the participant's median value for each dependent variable, and looked for its quartile rank among the values of the nearest neighbors (a low rank meant that this participant was doing worse than most of his/her neighbors). The dependent variables were (1) time to fall asleep, (2) number of wakeups per hour, (3) Sleep Recovery Index, and (4) sleep time.

Next, we selected the variable with the lowest rank to be the target of the recommendation. For example, if a participant takes 34 minutes to fall asleep, but 34 minutes is worse than 90% of the neighbors, compared to the other three variables which are comparatively worse than 20% of their neighbors, then their recommendation would be geared towards changing something that affects the time to fall asleep.

Once the dependent variable was selected, we identified which of the following was affecting it the most: sleep duration (the time spent in bed, but not necessarily asleep), consistency, exercise frequency, or heart rate. There was a recommendation for each one of these options, as shown in Table 7.2. The user received the recommendation for the option that had the highest effect towards improving the target variable. Continuing the example from above, if a participant is doing the poorly on the time to fall asleep metric compared to their cohort, we identified what factor is affecting the time to fall asleep of the cohort users the most. If it turned out to be their bedtime consistency, for example, we would send the participant the recommendation from the 'consistency' category.

In order to avoid confounding the effects of multiple recommendations, we only recommended one behavior change to each participant. Halfway through the study, the participants received the first email with their recommendation. Then, every day for the remaining two weeks, they received a reminder of the recommendation. Previous studies show that such periodic prompts increase adherence [71]. Similar to previous studies [51, 79, 35], the intervention consisted of a single behavior change, as it takes time for the change to have an effect on sleep and to avoid confounding effect of multiple interventions. While the focus of our study was not on behavior change, we wanted to design a study that maximized the benefit to the users in case the recommendation was helpful and they followed its advice.

7.4.2 Text of the Recommendations

Previous work has shown that people prefer recommendations containing more details about their own sleep [51]. In this study, we selected four main recommendation categories based on the MS Band data. Table 7.2 shows the text of each category (both recommendation conditions used the same template). 'Winding down' refers to doing something to relax before bed to let your body recognize that it is time to slow down.

Table 7.3: The average number of nights tracked per condition, and the average percentage difference in sleep time before and after the recommendation period per condition. The sleep time of the cohort-based recommendations condition increased the most.

Condition	Before Rec.	After Rec.	Percentage Difference in Sleep Time
No recommendations	11	11	0.6%
General	13	6	0.2%
Cohort-based	12	8	4.2%

7.5 Quantitative Summary

Sixty-six participants were initially assigned across the three conditions. During the course of the study, 13 people traveled across timezones and another 13 had less than 14 data points (this includes participants who had to stop tracking because their Bands stopped working and there were no more replacements available). Thus, only 40 people completed the study and met our pre-defined inclusion criteria. We excluded one of them from analysis because he did not track his sleep properly so he was missing some metrics.

Thus, we had 39 participants (8 female) for data analysis; 15 in "no recommendations," 11 in "general recommendations," and 13 in "cohort-based recommendations." Given our small sample size and the dropout rate caused by reasons such as travel and malfunctioning MS bands, we focus our analysis mainly on the qualitative feedback from participants, and limit the details of the quantitative analysis.

We calculated summary statistics for the three conditions in six sleep metrics: (1) sleep time, (2) time to fall asleep, (3) number of awakenings per hour, (4) subjective sleep quality, (5) ESS score, and (6) PSQI score. In the general condition, an approximately even number of each recommendation was initially assigned to participants at random. All the recommendations were derived from general sleep hygiene guidelines, so they were meant to be helpful for the general public. We sent the participants in the general condition a random one to evaluate whether the targeted cohort-based recommendation was more helpful than a general one. In the cohort-based recommendations condition, participants received the recommendation that was thought to be most helpful specifically for them, so the distribution of recommendation categories was not even like in the general condition. However, the final number of participants in the two recommendation conditions was 24. Thirteen out of 22 participants completed the study in the cohort-based condition, while 11 out of 21 completed it in the general condition. Given the small sample and exploratory nature of this study, we focus on the differences between the cohort-based and the general condition, rather than breaking down the analysis per recommendation category.

7.5.1 Microsoft Band Data

Table 7.3 shows the number of nights tracked per condition. For the conditions with recommendations, we employed a similar analysis procedure to SleepCoacher [51]: only the days that the participants said they

followed the recommendation were used for analysis, which explains the drop in the number of data points. It is worthwhile to note that participants may have had troublesome days, making them unable to follow the recommendation, and that stress would have affected their sleep. However, since our goal was to evaluate the effectiveness of the cohort-based recommendations in comparison to that of the general ones, we are keeping the analysis as consistent as possible, and excluding the days when they did not follow the recommendation in both conditions.

We used a t-test to determine the significance of the differences between the conditions. We had one hypotheses for each of the four dependent variables: that each of the four dependent variables would be significantly improved for the cohort-based recommendations condition compared to the those of the other two conditions. However, while all four variables for the cohort-based condition improved the most, the differences were not significant when we applied the Bonferonni correction to the 4 hypotheses.

Figure 7.1 shows the average sleep time for each condition before and after the recommendation. The condition without any recommendations has the shortest sleep time, whereas the cohort-based recommendations one has the longest sleep time. Furthermore, the highest percentage of people who increased their sleep time is in the cohort-based recommendations condition (63%), along with the highest average percentage of improvement (4.2%). In comparison, only 55% of people in the general condition and 53% of people in the no recommendations condition increased their sleep time.

7.5.2 PSQI and ESS Sleep Measures

We collected the PSQI and ESS scores of participants both before and after the study. Initially, the scores between the three conditions were not significantly different. A higher score on the PSQI (out of 21 points) and ESS (out of 24 points) scales is interpreted as worse sleep, and thus a decrease reflects improvement when looking at the change in these metrics. The PSQI scores of the no recommendations condition increased by 2.25 points on average, compared to 1 point for general condition, and only 0.42 points for the personalized condition. We summarize the changes of PSQI scores in Figure 7.2. However, both questionnaires are based on self-reported sleep quality factors such as duration and latency. Previous studies have indicated that people report higher awareness of their habits after they start self-tracking [51, 118, 37]. Therefore, these results might indicate that participants were more conscious of their sleep patterns when filling out the end-of-study questionnaire, causing their worsened sleep scores.

7.6 Qualitative Findings

In addition to the data collected by the MS Band and the ESS & PSQI questionnaires, participants provided written responses in the post-study questionnaire. We adopted an inductive approach to analyze the responses by performing a thematic analysis [24] on the data. Three researchers independently coded the written

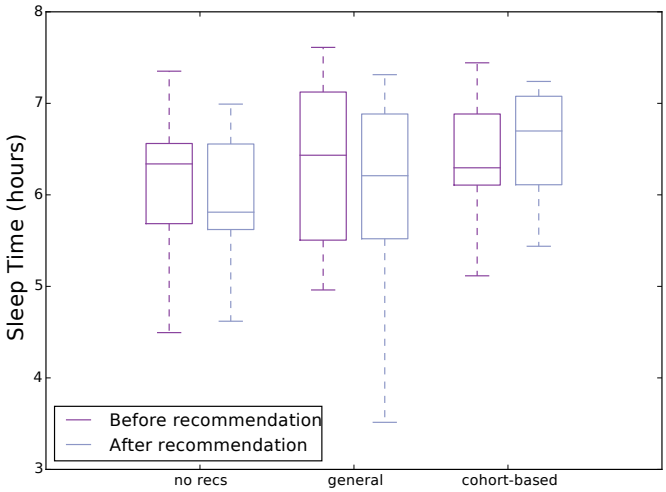


Figure 7.1: Sleep time amounts per condition before and after the recommendation. While there was high variance, the three groups were not significantly different before the recommendation. However, after the recommendation, cohort-based recommendations resulted in longer sleep times.

responses. Then, the coding schemes were discussed until a consensus was reached on the overall themes and sub-themes (Figure 7.3). The major themes were focused on whether participants learned something from being a part of the study, whether they thought the recommendations were helpful, and their main reasons for following or not following the recommendations. The participants who adhered to the recommendations the most were the ones that increased their sleep time the most.

While the quantitative data analysis focused only on the participants who followed the recommendations for at least one day, the qualitative analysis also explored the answers of those who did not follow the recommendations at all to gain a deeper understanding of why participants chose to follow or not follow the suggestions. Due to the qualitative nature of the data, we did not seek measurable differences between the groups, so here we describe findings from both conditions – the most helpful aspects of the recommendations as well as the aspects that need further improvement. We include some nuances of the participants' experiences to highlight the complexity of adhering to recommendations. We further elaborate on these findings in the context of cohort-based recommendations in Section 7.7.

7.6.1 Helpful Aspects of the Recommendations

Similar to previous studies such as ShutEye [18] and SleepTight [37], we found that sleep tracking systems (1) serve as reminders for well-known healthy sleep behaviors and increase people's awareness about their current sleep habits, and (2) help users identify patterns in the effects of various behaviors on their sleep. However, we also identified a type of self-reflection insight due to cohort formation: participants in our study were comparing themselves to the 'people' mentioned in the text of the recommendations.

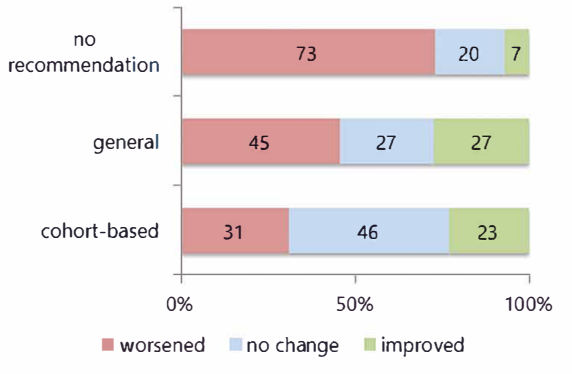


Figure 7.2: The percentage of participants whose PSQI score changed in each direction per condition. The PSQI scores of the no-recommendations condition worsened the most.

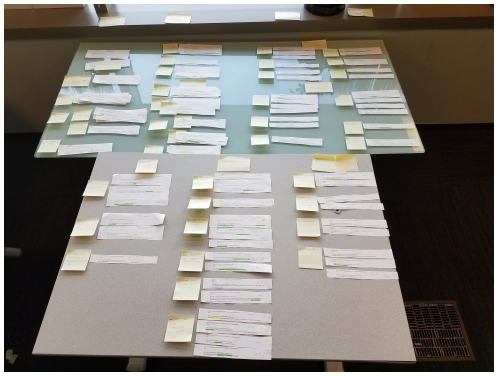


Figure 7.3: Three of the authors performed a thematic analysis on the final survey data, which resulted in a few major themes discussed in the Qualitative Findings section.

7.6.1.1 Increasing Consciousness about Current Sleep Habits

The most common reason for following the recommendation, reported by five participants, was that it was already something that they were trying to do but the study increased their attention and adherence to it. Thus, it was *"a good reminder of what to do even though I have heard the recommendation before,"* as one participant put it.

Five other participants' reason for following the recommendations was that sleep was important to them and they wanted to try the suggested behavior change to see if it would *"make a difference."* However, receiving the extra nudge from the study was what triggered the increase in *"attention and focus,"* as P9 pointed out. Overall, participants found the recommendation helpful because it made them more conscious about their current sleep habits in general, which is a common finding to previous sleep studies such as ShutEye [18] and SleepCoacher [51]. P29, who received a cohort-based recommendation, said that *"I consciously stepped away from screens before bedtime and started doing meditation to relax."* P7, who received a cohort-based recommendation for consistency, pointed out that it made him *"more conscious about getting sleep—trying to adjust my bedtime."*

7.6.1.2 Emphasizing Impact of Various Factors on Sleep

On the one hand, two participants distinctly noticed that they slept better when they were following the recommendations. P11, for example, said that *"I found that if I did go to bed as suggested, I felt more rested on the next day and getting up was easier."* On the other hand, some also inferred causal effects about interferences to sleep, such as P15, who received a general recommendation about consistency, actually learned that *"temperature fluctuation definitely affected my sleep."* Another participant P9, whose recommendation

was to wind down before bed, noticed that *"meditating makes a huge difference, but little else seems to."* Overall, five participants specifically pointed out that their sleep habits were "wrong" and should do something to change that, which matches the "trend" reflection type from Choe et al. [38]. P21, for example, found that *"I lack a lot of sleep, and need to to go bed earlier."* Thus, participants noticed patterns in how the recommendations affected their sleep.

7.6.1.3 Using Social Comparison as Behavior Change Motivation

The recommendations stated statistics about other people, which led some participants to compare themselves to them. Thus, some participants found out that they were not doing so well compared to others. P33 had gained Choe et al.'s "comparison" type insight [38] that *"I am getting less sleep than the body of users I am being compared against."* We consider implications of this finding in Section 7.7.

7.6.2 Lessons Learned About the Shortcomings of the Recommendations

In this section, we identify aspects of the recommendations that detracted participants from following them every day. We address possible suggestions on how to mediate those issues in the context of cohort-based recommendations in the Discussion section.

7.6.2.1 Prior Commitments Made It Difficult to Fit the Recommendation in Daily Schedule

The most common reason for not following the suggestion, reported by 12 of the participants, was because it was too difficult to adjust their schedules. Nine of them followed it only three times or less per week. They had social and work commitments that prevented them from going to bed early enough to get more sleep or from waking up later. Unfortunately, according to sleep literature [5], a consistent bedtime and waketime schedule will have the best effects only when followed every day, including on weekends.

An interesting insight came from P7, who was recommended to keep a consistent wake time. He reported that he needs to *"get up by a certain time in order to take the children to school,"* and that his children's schedule *"is not flexible, so suggesting I sleep later was not helpful."* This adds nuance to his earlier comment that the recommendation was helpful because it made him adjust his bedtime. We explore implications in the Discussion section, but it is important to point out that a good sleep recommender should be able to take such constraints into account before giving recommendations.

Even without having to adjust their bed and wake schedule, some participants pointed out that it was difficult to fit the recommended action into their routine. P1 was recommended to exercise 3 times a week, but only did so once a week because *"that is as much exercise as I am able to put on my schedule."*

7.6.2.2 The Perceived Effect of the Recommendation Did Not Match Required Effort

Three participants pointed out that they did not follow the recommendation because the burden of adhering to it outweighed its benefits. P4 said that *"it's hard to adjust the schedule – the impact is moderate compared to other factors."* P12 further reported that it was *"not very useful, it was a difference of 10 minutes in my sleep."* Three participants, including P13, pointed out that the recommendation was "not specific enough," so it did not entice them to adhere to it.

7.6.2.3 The Recommendation Did Not Seem Trustworthy nor Encouraging

Three participants expressed mistrust in the recommendation, caused by one of the following: (1) they doubted it was based on the correct metric, (2) the phrasing was unconvincing, or (3) it did not match their preconceptions about what good sleep, in particular, about amount of sleep. P18 thought that the time to fall asleep was not measured properly. P14, who received a cohort-based recommendation for sleep duration, actually pointed out that *"it didn't seem like a real person looked at that statement."* P10 received a similar recommendation, but was disappointed because *"there wasn't any encouragement like FitBit to actually try to change my sleep."* P27, on the other hand, did not trust the recommendation because the suggested hours of sleep seemed inaccurate. She strongly believed that *"people need 7.5 hours of sleep,"* so it did not seem logical that the recommendation was for a different amount.

7.6.2.4 The Recommendation Was Not Novel or Was Not Related to What They Wanted to Improve

In contrast to the participants from Section 6.1.1. who were inspired to follow the recommendation specifically because they had seen it before and were thus enticed to finally try it, other participants expressed their disappointment that they did not learn anything new from the recommendation, and that is why they did not adhere to it. Two participants from the recommendations conditions reported that they had been tracking their sleep for a while previously, so they were already aware of what affects their sleep. P6, for example, said *"I've been tracking my sleep for the last few years so not much was new here."*

Participants reported insights about observations that confirmed previous knowledge about themselves. P3 was recommended to keep a consistent sleep schedule to which he replied that *"I guess I already knew it, but it was nice to see the data,"* whereas another participant said that *"I confirmed I don't sleep enough."* This insight adds nuance to the reasons why people are not improving their sleep: even when participants acknowledge the helpful things they can and should be doing to have proper sleep hygiene, they are still not necessarily following them.

Alternatively, two participants did not find the recommendation helpful because they were hoping to improve a different aspect of their sleep that they were having trouble with. P20, for example, who was recommended to wind down before bed, said that *"I have trouble staying asleep and getting quality sleep,*

and I have no trouble falling asleep. The recommendation was too generic and doesn't actually apply to my sleeping habits."

7.6.2.5 The Recommendation Did Not Lead to Immediate Improvement

Finally, two participants pointed out that they did not think the recommendation was helpful because they felt better on the days that they did not follow it. Specifically, both of those participants were referring to the fact that they would rather wake up naturally than use an alarm clock for a preset wake up time. Furthermore, another participant pointed out that *"the exercise does affect sleep but other things could affect it way more."* This quote also represents similar feedback from three other participants who did not think that what the recommendation was suggesting was really the cause of their sleep issues.

7.7 Discussion

Our 4-week study leverages a rich real-world dataset and builds on existing collaborative-filtering frameworks to provide sleep recommendations based on a cohort of similar users. As such, our work identifies limitations of these techniques that surface when participants are asked to adhere to the recommendations in real life. In this section, we discuss some design hypotheses that might overcome these limitations, as well as their implications for future studies that evaluate how cohort-based recommendations in the health space could be improved. Based on the complexities we identified, we also discuss the need for incorporating various personal constraints and for phrasing the recommendations in an appropriate way. Our work can be used as a basic framework for future work, which could also search for more effective phrasings of the recommendations.

7.7.1 Selecting a Cohort of Similar Users

Collaborative filtering recommender systems use a variety of algorithms to recommend items to a user that other similar users have liked [150]. As in our approach, the similar users are clustered together based on relevant attributes. Usually that means that for each user, a set of nearest neighbors is found with whose past ratings there is the strongest correlation [13]. However, while those recommender systems are based on what other users have liked, our method is based on what other users with better sleep quality have done. We chose to base cohorts on the simple model of demographic information (BMI and gender), self-perceived sleep quality, and self-reported exercise level. While our approach used only a nearest neighbors classifier, previous studies have shown that clustering can be added as a pre-processing step to increase efficiency [192].

Below, we discuss four more complex ways for generating the cohorts based on our findings: (1) ask users for their constraints and base the cohort on users with similar constraints, (2) ask users about what activities they engage in or are interested in trying, (3) let the users select what they want their cohorts to be based on, and (4) provide more details about who the other people in their cohort are.

As stated in Section 7.6.2.1, twelve of the twenty-four participants who received any kind of recommendation brought up the issue that the suggested behavior change was not attainable due to constraints in their schedule. Nevertheless, a part of them attempted to incorporate it in their lives as much as possible. The participants who followed their recommendations the most were the ones that improved their sleep the most. However, for the majority of participants, the suggestion was not actionable enough, so they did not follow it or benefit from it. Therefore, the main deterrent to improving sleep was the practicality of the recommendations. Thus, in this section, we focus on ways we can improve the framework to generate actionable and impactful recommendations, inspired by social and behavioral theory.

To provide more actionable recommendations, a future system can ask the user for what limitations already exist in their daily routine and build a cohort based on people with similar restrictions. Furthermore, the system could also ask users for what activities (e.g., exercise and relaxation) they engage in to build on existing habits, rather than attempt to introduce a completely new one. That way, the cohorts would be based more realistically on people whose daily lives include similar opportunities for improvement and challenges. Finally, the system can also ask the user what they want to improve to make sure it finds a cohort that worked towards a similar goal, addressing participant feedback from Section 7.6.2.4.

A further implication, brought up by two participants, is to let users select what they want their cohorts to be based on: while some might prefer to focus just on their gender and age group, others might pick people with similar occupations, schedules, fitness activities, dependents, or lifestyles. This might increase their trust in the recommendation, and make sure that their cohorts are people they consider similar to themselves.

Additionally, while the "people like you" in recommender systems like Netflix are usually hidden [194], we found that participants in our study generally wanted to know more about who those people are. Therefore, perhaps in the health domain the framework for cohort-based recommendations needs to provide details about how the cohort of nearest neighbors is similar to the given user. It is important to be able to balance a sense of cohesion in a cohort while preserving individual privacy.

Selecting the right cohort that the user identifies with is critical for inspiring behavior change. According to social cognitive theory [14], self-efficacy is the belief in one's ability to perform certain behaviors in a given context, and that they will have an effect on one's life. We can use this notion to show users that people with similar constraints and schedules are still changing their behavior and improving their sleep. This could potentially motivate the user to also implement the changes and, in turn, increase their self-efficacy.

7.7.2 Phrasing of the Cohort-Based Recommendations

Since the only recommendations in sleep research studies so far have been either based just on the individual user [51] or on the general sleep hygiene guidelines, there was no clear direction on how to phrase the cohort-based recommendations. We chose to frame the recommendations as general advice, followed by specific averages for the cohort and the individual user. While the focus in this study was not on how to best

phrase the recommendations, participants provided us with valuable feedback on how to improve them in the future. Below, we focus on six main improvements based on our findings: (1) increase trustworthiness of the recommendation, (2) include information about how helpful and worthy this recommendation is, incorporating statistics from people in the cohort who have attempted it previously, (3) start with small actionable steps, (4) explore less known sleep recommendations to introduce novelty, (5) phrase them differently for collectivist vs individualist societies, and (6) be clear about how the cohort is similar to user.

The trustworthiness of the recommendation was brought up in a few ways from different participants in Section 7.6.2.3. One important tension that we identified was between general sleep hygiene guidelines and personalize recommendations. For example, P27 strongly believed in the general advice that "people need 7.5 hours of sleep," so when she received her specific cohort-based recommendations for a different amount, she deemed it untrustworthy. Thus, the discrepancy between the well-known tips and the new recommendations caused skepticism. One way to mitigate that would be to include the source of the suggestion, as discussed in ShutEye [18]. Similarly, cohort-based recommendations might benefit from including short snippets with facts from sleep literature, presented in layman's terms to educate the participant.

Furthermore, recommendations might be more trustworthy if it was obvious that they helped similar users. As illustrated in Section 7.6.2.2, participants have a preconception of what is considered an impactful and worthy outcome for which to change their behavior. Five participants reported that they ignored the recommendations because they considered them unworthy of the effort required for the behavior change. Further research is needed to identify where that boundary lies. The trustworthiness of the recommendations can be increased by being more transparent about how each metric is calculated. This is also related to participants' desire for more specific details. However, this brings up another interesting tension: what is the right balance between including all the details that we have to make the recommendation believable and at the same time making sure the user is not being overloaded with information.

The third improvement we suggest is grounded in behavior change literature: the phrasing of the intervention affects the way the patient understands it [159]. The Nudge Theory, for example, emphasizes that the intervention must be simple and easy to follow, such as presenting fruit at an eye level and fried chips on a top shelf [175]. Similarly, the recommendations in the domain of health have to be as clear as possible. According to Fogg's Tiny Habits, in addition to being clear, they also need to start with the smallest actionable step possible [68]. Exercise seems to be the most difficult recommendation to follow, as it involves introducing a new habit in the cases when participants do not usually workout. The participants in our study that improved their sleep the most were the ones that followed their recommendation most often. This leads to an interesting question of whether following the recommendation on just a few nights leads to enough of an effect and if the frequency can potentially be built up from there. If it is not enough, then the question is whether the behavior change is worth pursuing at all.

Section 7.6.1.1 described that participants liked receiving recommendations because they served as a reminder of a health behavior they already knew they should be engaging in. It is not surprising that

participants were already aware of these suggestions since we specifically chose recommendations that were part of the general sleep hygiene guidelines. However, not all general guidelines will work for everyone, so another improvement in could be to explore whether less well-known guidelines based on the specific behaviors of the cohort are appealing enough to be followed.

An interesting study by Cialdini et al. assessed the impact of two social influence principles: 1) commitment/consistency and 2) social proof on participants' decisions. They found Americans and Poles are impacted differently by the two principles, with Americans being more impacted by the commitment/consistency principle [41]. This discrepancy implies that behavior change recommendations might need to be phrased differently according to the user's background. Specifically, cohort-based recommendations that rely on parallels between the individual and other similar users might be most effective in collectivist societies.

Lastly, based on participant feedback, the recommendations would also be more dependable if they knew how the other people in their cohort were similar to them (as described in Section 7.7.1).

7.7.3 Social Comparison and Interaction

According to the theory of social comparison, people compare themselves to others with similar opinions or abilities [65]. In the context of our study, this emphasizes the need of a participant to know who the other people in their cohort are in order to know how closely related they are to them. Some participants in Section 7.6.1.3 specifically mentioned that they compared themselves to the cohort. Past applications such as Shakra [10] and Houston [43] helped users compare their fitness data to that of their friends, but the cohorts in our study were strictly strangers. Future work could explore whether cohorts based on people we know give a basis for better sleep recommendations.

Previous research has shown that sharing fitness information with friends was helpful [10], that online social networks may be effective in behavior change interventions [124], and that human interaction was successful in promoting increased physical activity among middle-aged and elderly people [105]. The benefit of cohort-based recommendations specifically are that participants can be following the same recommendations. Additionally, users might benefit from being able to interact with their cohort: both for giving reminders to each other and for inspiration for behavior change.

7.7.4 Improving the Recommendation Generation

In this study, we used an algorithm to generate recommendations based on nearest neighbors. One suggested improvement to our framework is to use only recently generated data from current active users. In the case that such information is not available, we could also use a similar time frame and location from previous years. This could ensure similar weather trends to give more accurate recommendations as seasonal variations can have a great impact on people's sleep. Another possible improvement is to adjust the number of nearest neighbors: future studies can explore the effect of a cohort size.

Finally, since two participants in Section 7.6.2.4 received a recommendation for a variable different from what they were hoping for, future systems could ask users what they would like to focus on and provide intervention suggestions specifically for that variable. This design hypothesis is in agreement with goal-setting theory [122], according improvement is highest when the user sets the goals. Thus, the system could identify the people in this user's cohort that are doing better on the target variable, and suggest a recommendation based on what they are doing differently from the user.

7.7.5 Limitations

The results of this work are limited by the population demographics. All participants were employees of one technology company, and while they were from diverse occupations, most were software engineers, and the gender distribution was not balanced. Furthermore, the month-long study was conducted during the spring, which is generally considered a transition period when children are on break or people travel for vacation. However, the goal of the study was to evaluate the effect of the cohort-based recommendations and to explore people's reactions to them, so we still gained valuable insights. Further work can focus on applying these methods to a broader population. Another limitation is that the sample is already somewhat biased, as they all owned an MS Band previously and are thus already conscious or interested in tracking their health. Even with this limitation of a specific population from the technology sector, however, the study was designed to evaluate the effectiveness of cohort-based recommendations compared to general ones.

7.8 Conclusion

We presented the findings of a four-week study that explored the effectiveness of cohort-based sleep recommendations. To evaluate the effects of these recommendations, we compared the sleep quality of participants in three conditions who either received: (1) no recommendation, (2) a general recommendation, or (3) a cohort-based recommendation. We learned that participants' sleep time increased by 16 minutes (4.2%) on average when they received cohort-based recommendations, whereas it increased by less than one minute (0.18%) on average for participants who received general ones. Based on the participant feedback, we identified and discussed design hypotheses that can be tested in future cohort-based sleep recommender systems. We found that users preferred to be given more control over the selection of their cohorts, and wished that the recommender system considered their constraints related to their occupation, schedule, and lifestyle. Our work adds to the growing body of knowledge on how to make the recommendations more trustworthy, and how to incorporate social comparison to make them more engaging. This study opens a new direction of investigation of what happens when sleepers are put into cohorts, to try and sleep better together.

Chapter 8

Conclusions & Future Directions

This dissertation presents novel systems that aim to make self-experimentation accessible to novices. First, in Chapter 3, we presented SleepCoacher, an automated system that lets users track their sleep and then provides personalized data-driven recommendations in the form of an experiment. In Chapter 4, we presented findings about how people conduct self-experiments for the first time on their own, as well as the guidelines we developed to help them navigate this process.

In Chapter 5, we presented a set of design principles for building systems that guide users through the steps of a flexible self-experiment. We implemented those principles in SleepBandits, a self-contained system that incorporates a robust sleep tracking app and a back-end server that uses Thompson Sampling to analyze the data from the experiment. Chapter 6 builds on the findings from the study of SleepBandits, and implements the guided approach to flexible self-experiments in domains beyond sleep, in the form of the Self-E system. Based on the qualitative study we conducted, we learn valuable lessons about how to evolve such systems in order to match the mental models of users. Lastly, Chapter 7 moves away from analyzing only one user's data to leveraging the trends found across cohorts of similar users in order to provide behavior change recommendations based on collaborative filtering.

Overall, the systems described in this dissertation aim to lower the barrier to self-experimentation for a wider audience of users by prioritizing agency and flexibility while maintaining scientific rigor. Most users liked that our systems were optimized for the simplest experiments possible, starting with a single intervention and target variable at a time. However, as they learn more about how to conduct basic self-experiments, some users become interested in more sophisticated ones and look for features that SleepBandits and Self-E were not designed to include. Thus, our systems might need to focus on those aspects in order to promote a wider adoption of self-experimentation. Future research in the field can expand on the systems we have developed by encouraging users to track for longer and providing them with even more actionable information that leads to interesting and useful personalized insights.

Going forward, we can apply the lessons we learned from this work to domains beyond health, and help people optimize other aspects of their lives. Thanks to the ubiquity of data collection, we have large amounts of data available for each person from various sources. Thus, we could, for example, use email patterns or social media data to determine when users are most productive or most focused. By leveraging the abundant amount of data about themselves that people have access to, we can explore how to conduct non-traditional self-experiments that lead to meaningful and actionable insights.

While SleepBandits and Self-E are currently only based on each individual's data, our findings suggest that self-experimentation can be enhanced through incorporating a social aspect in the process. Going further, we can leverage theory from social psychology to encourage users to continue with their experiments through social interaction and collaboration. One way could be to have users who are conducting the same experiment be able to reach out to each other and provide guidance and support to others. Another way could be to connect users who already have a social bond in real life and allow them to conduct their experiments in parallel and compare results. Thus, we can investigate new human-centered ways to conduct social self-experiments.

Furthermore, research suggests that techniques and models from artificial intelligence have the potential to improve health and wellness tools [165]. SleepBandits and Self-E are a first step in that direction with their implementation of the Thompson Sampling algorithm. However, further work is needed to explore how we can visualize or phrase the self-experiment and Thompson Sampling data in a way that clearly conveys its meaning. We can delve deeper into this topic and bridge findings from biostatistics, artificial intelligence, and psychology in order to determine the best way to analyze and present data from self-experiments.

Bibliography

[1] Fitbit. http://www.fitbit.com/.

[2] Healthy sleep tips, 2014. Retrieved November 23, 2014 from http://sleepfoundation.org/sleep-tools-tips/healthy-sleep-tips.

[3] Tips for better sleep. https://www.cdc.gov/sleep/about_sleep/sleep_hygiene.html, Jul 2016.

[4] Sleep as android, 2017. Retrieved September 6, 2018 from http://sleep.urbandroid.org/.

[5] Sleep hygiene tips - research & treatments - american sleep assoc. https://www.sleepassociation.org/patients-general-public/insomnia/sleep-hygiene-tips/, 2017.

[6] Obesity and sleep. https://sleepfoundation.org/sleep-topics/obesity-and-sleep, 2018.

[7] Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of COLT*, volume 23, pages 39.1–39.26, 2012.

[8] Tim Althoff, Eric Horvitz, Ryen W White, and Jamie Zeitzer. Harnessing the web for population-scale physiological sensing: A case study of sleep and performance. In *Proceedings of the 26th International Conference on World Wide Web*, pages 113–122. International World Wide Web Conferences Steering Committee, 2017.

[9] Sonia Ancoli-Israel, Roger Cole, Cathy Alessi, Mark Chambers, William Moorcroft, and Charles P Pollak. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 26(3):342–392, 2003.

[10] Ian Anderson, Julie Maitland, Scott Sherwood, Louise Barkhuus, Matthew Chalmers, Malcolm Hall, Barry Brown, and Henk Muller. Shakra: Tracking and sharing daily activity levels with unaugmented mobile phones. *Mobile networks and applications*, 12(2-3):185–199, 2007.

[11] Amid Ayobi, Paul Marshall, Anna L Cox, and Yunan Chen. Quantifying the body and caring for the mind: self-tracking in multiple sclerosis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6889–6901. ACM, 2017.

[12] Amid Ayobi, Tobias Sonne, Paul Marshall, and Anna L Cox. Flexible and mindful self-tracking: Design implications from paper bullet journals. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 28. ACM, 2018.

[13] Marko Balabanović and Yoav Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[14] Albert Bandura. *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ, US: Prentice-Hall, Inc, 1986.

[15] David H Barlow, Nock K Matthew, and Michel Hersen. *Single case experimental designs: Strategies for studying behavior for change*. Pearson, 2008.

[16] Kelly Glazer Baron, Kathryn J Reid, and Phyllis C Zee. Exercise to improve sleep in insomnia: Exploration of the bidirectional effects. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, 9(8):819, 2013.

[17] Colin Barr, Maria Marois, Ida Sim, Christopher H. Schmid, Barth Wilsey, Deborah Ward, Naihua Duan, Ron D. Hays, Joshua Selsky, Joseph Servadio, Marc Schwartz, Clyde Dsouza, Navjot Dhammi, Zachary Holt, Victor Baquero, Scott MacDonald, Anthony Jerant, Ron Sprinkle, and Richard L Kravitz. The preempt study-evaluating smartphone-assisted n-of-1 trials in patients with chronic pain: Study protocol for a randomized controlled trial. *Trials*, 16(1):67, 2015.

[18] Jared S Bauer, Sunny Consolvo, Benjamin Greenstein, Jonathan Schooler, Eric Wu, Nathaniel F Watson, and Julie Kientz. ShutEye: Encouraging awareness of healthy sleep recommendations with a mobile, peripheral display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1401–1410. ACM, 2012.

[19] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(5):30, 2013.

[20] Michael S Bernstein, Mark S Ackerman, Ed H Chi, and Robert C Miller. The trouble with social computing systems research. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 389–398. ACM, 2011.

[21] Alina Beygelzimer, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li. Fnn. `https://cran.r-project.org/web/packages/FNN/index.html`, 2013.

[22] Allen J. Blaivas. Polysomnography, 2014. Retrieved August 25, 2018 from `http://www.nlm.nih.gov/medlineplus/ency/article/003932.htm`.

[23] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. TasteWeights: A visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 35–42. ACM, 2012.

[24] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[25] Sarah A Burgard and Jennifer A Ailshire. Gender and time for sleep among us adults. *American sociological review*, 78(1):51–69, 2013.

[26] Daniel J Buysse. Sleep health: Can we define it? Does it matter? *Sleep*, 37(1):9–17, 2014.

[27] Daniel J Buysse, Charles F Reynolds, Timothy H Monk, Susan R Berman, and David J Kupfer. The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry res*, 28(2):193–213, 1989.

[28] Kelly Caine. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 981–992. ACM, 2016.

[29] Colleen E Carney, Daniel J Buysse, Sonia Ancoli-Israel, Jack D Edinger, Andrew D Krystal, Kenneth L Lichstein, and Charles M Morin. The consensus sleep diary: Standardizing prospective sleep self-monitoring. *Sleep*, 35(2):287–302, 2012.

[30] Marta E Cecchinato, John Rooksby, Alexis Hiniker, Sean Munson, Kai Lukoff, Luigina Ciolfi, Anja Thieme, and Daniel Harrison. Designing for digital wellbeing: A research & practice agenda. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page W17. ACM, 2019.

[31] Ting-Ray Chang, Eija Kaasinen, and Kirsikka Kaipainen. What influences users' decisions to take apps into use?: A framework for evaluating persuasive and engaging design in mobile apps for well-being. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, page 2. ACM, 2012.

[32] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson Sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[33] Li Chen and Pearl Pu. Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1-2):125–150, 2012.

[34] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T Campbell. Unobtrusive sleep monitoring using smartphones. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 145–152. IEEE, 2013.

[35] Mi-Yeon Cho, Eun Sil Min, Myung-Haeng Hur, and Myeong Soo Lee. Effects of aromatherapy on the anxiety, vital signs, and sleep quality of percutaneous coronary intervention patients in intensive care units. *Evidence-Based Complementary and Alternative Medicine*, 2013, 2013.

[36] Eun Kyoung Choe, Sunny Consolvo, Nathaniel F Watson, and Julie A Kientz. Opportunities for computing technologies to support healthy sleep behaviors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3053–3062. ACM, 2011.

[37] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A Kientz. SleepTight: Low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 121–132. ACM, 2015.

[38] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. Understanding self-reflection: How people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 173–182. ACM, 2017.

[39] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1143–1152. ACM, 2014.

[40] Chia-Fang Chung, Elena Agapie, Jessica Schroeder, Sonali Mishra, James Fogarty, and Sean A Munson. When personal tracking becomes social: Examining the use of instagram for healthy eating. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1674–1687. ACM, 2017.

[41] Robert B Cialdini, Wilhelmina Wosinska, Daniel W Barrett, Jonathan Butner, and Malgorzata Gornik-Durose. Compliance with a request in two cultures: The differential influence of social proof and commitment/consistency on collectivists and individualists. *Personality and Social Psychology Bulletin*, 25(10):1242–1253, 1999.

[42] John Concato, Nirav Shah, and Ralph I Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342(25):1887–1892, 2000.

[43] Sunny Consolvo, Katherine Everitt, Ian Smith, and James A Landay. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 457–466. ACM, 2006.

[44] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3207–3216. ACM, 2015.

[45] Felicia Cordeiro, Daniel A Epstein, Edison Thomaz, Elizabeth Bales, Arvind K Jagannathan, Gregory D Abowd, and James Fogarty. Barriers and negative nudges: Exploring challenges in food journaling. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1159–1162. ACM, 2015.

[46] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 175–182. ACM, 2004.

[47] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. Lessons learned from two cohorts of personal informatics self-experiments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):46, 2017.

[48] Nediyana Daskalova, Nathalie Ford, Ann Hu, Kyle Moorehead, Benjamin Wagnon, and Janet Davis. Informing design of suggestion and self-monitoring tools through participatory experience prototypes. In *International Conference on Persuasive Technology*, pages 68–79. Springer, 2014.

[49] Nediyana Daskalova, Eindra Kyi, Kevin Ouyang, Andrew Park, Nicole Nugent, and Jeff Huang. Self-E: Guided self-experimentation beyond sleep. In *[submission]*.

[50] Nediyana Daskalova, Bongshin Lee, Jeff Huang, Chester Ni, and Jessica Lundin. Investigating the effectiveness of cohort-based sleep recommendations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–19, 2018.

[51] Nediyana Daskalova, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. SleepCoacher: A personalized automated self-experimentation system for sleep recommendations. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 347–358. ACM, 2016.

[52] Nediyana Daskalova, Jina Yoon, Yibing Wang, Cintia Araujo, Guillermo Beltran, Nicole Nugent, John McGeary, Joseph Jay Williams, and Jeff Huang. Sleepbandits: Guided flexible self-experiments for sleep. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20. ACM, 2020.

[53] Orianna DeMasi, Sidney Feygin, Aluma Dembo, Adrian Aguilera, and Benjamin Recht. Well-being tracking via smartphone-measured activity and sleep: Cohort study. *JMIR mHealth and uHealth*, 5(10):e137, 2017.

[54] Joan Didion. On keeping a notebook, 1968. Retrieved September 24, 2015 from `https://penusa.org/sites/default/files/didion.pdf`.

[55] John B Dixon, Linda M Schachter, and Paul E O'brien. Sleep disturbance and obesity: Changes following surgically induced weight loss. *Archives of internal medicine*, 161(1):102–106, 2001.

[56] Gavin Doherty, David Coyle, and John Sharry. Engagement with online mental health interventions: An exploratory clinical study of a treatment for depression. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1421–1430. ACM, 2012.

[57] Markéta Dolejšová and Denisa Kera. Soylent diet self-experimentation: Design challenges in extreme citizen science projects. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2112–2123. ACM, 2017.

[58] Naihua Duan, Ian Eslick, Nicole Gabler, Heather Kaplan, Richard Kravitz, Eric Larson, Wilson Pace, Christopher Schmid, Ida Sim, and Sunita Vohra. *Design and Implementation of N-of-1 Trials: A User's Guide*. Agency for Healthcare Research and Quality, 2014.

[59] Eugene S. Edgington. Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology*, 34(4):437–442, 1987.

[60] Daniel A Epstein, Nicole B Lee, Jennifer H Kang, Elena Agapie, Jessica Schroeder, Laura R Pina, James Fogarty, Julie A Kientz, and Sean Munson. Examining menstrual tracking to inform the design of personal informatics tools. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6876–6888. ACM, 2017.

[61] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 731–742. ACM, 2015.

[62] Deborah Estrin and Ida Sim. Open mHealth architecture: An engine for health care innovation. *Science*, 330(6005):759–760, 2010.

[63] Bob Evans. Paco: The personal analytics companion. https://pacoapp.com/, 2019.

[64] Miguel Farias and Catherine Wikholm. Has the science of mindfulness lost its mind? *BJPsych bulletin*, 40(6):329–332, 2016.

[65] Leon Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.

[66] Brianna S Fjeldsoe, Alison L Marshall, and Yvette D Miller. Behavior change interventions delivered by mobile telephone short-message service. *American journal of preventive medicine*, 36(2):165–173, 2009.

[67] Brian J Fogg. Persuasive technology: Using computers to change what we think and do. *Ubiquity*, 2002(December):5, 2002.

[68] Brian J. Fogg. Tiny habits, 2015. Retrieved September 21, 2018 from http://tinyhabits.com/.

[69] National Sleep Foundation. Healthy sleep tips, 2014. Retrieved November 23, 2018 from http://sleepfoundation.org/sleep-tools-tips/healthy-sleep-tips.

[70] Susannah Fox and Maeve Duggan. Tracking for health. https://www.pewinternet.org/2013/01/28/tracking-for-health/, Jan 2013.

[71] Jillian P Fry and Roni A Neff. Periodic prompts and reminders in health promotion and health behavior interventions: Systematic review. *Journal of medical Internet research*, 11(2), 2009.

[72] Andrew G Miner, Theresa M Glomb, and Charles Hulin. Experience sampling mood and its correlates at work. *Journal of Occupational and Organizational Psychology*, 78(2):171–193, 2005.

[73] Martin J. Gardner and Douglas G. Altman. Confidence intervals rather than p values: Estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, 292(6522):746–750, 1986.

[74] Anne Germain, Robin Richardson, Douglas E Moul, Oommen Mammen, Gretchen Haas, Steven D Forman, Noelle Rode, Amy Begley, and Eric A Nofzinger. Placebo-controlled comparison of prazosin and cognitive-behavioral treatments for sleep disturbances in us military veterans. *Journal of psychosomatic research*, 72(2):89–96, 2012.

[75] Cynthia Graber and Nicola Twilley. Diet for one? Scientists stalk the dream of personalized nutrition. *The New York Times*, Jun 2019.

[76] Mathias Haefeli and Achim Elfering. Pain assessment. *European Spine Journal*, 15(1):S17–S24, 2006.

[77] Andria Hanbury, Katherine Farley, Carl Thompson, Paul M. Wilson, Duncan Chambers, and Heather Holmes. Immediate versus sustained effects: Interrupted time series analysis of a tailored intervention. *Implementation Science*, 8(1):130–147, 2013.

[78] Tian Hao, Guoliang Xing, and Gang Zhou. iSleep: Unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, pages 1–14, 2013.

[79] VR Hariprasad, PT Sivakumar, V Koparde, S Varambally, J Thirthalli, M Varghese, IV Basavaraddi, and BN Gangadhar. Effects of yoga intervention on sleep and quality-of-life in elderly: A randomized controlled trial. *Indian journal of psychiatry*, 55(Suppl 3):S364, 2013.

[80] F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 3–10. ACM, 2015.

[81] Daniel Harrison, Paul Marshall, Nadia Bianchi-Berthouze, and Jon Bird. Activity tracking: Barriers, workarounds and customisation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 617–621. ACM, 2015.

[82] Eiji Hayashi and Jason Hong. A diary study of password usage in daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2627–2630. ACM, 2011.

[83] Steven C Hayes. Single case experimental design and empirical clinical practice. *Journal of consulting and clinical psychology*, 49(2):193, 1981.

[84] Eric B Hekler, Winslow Burleson, and Jisoo Lee. A DIY self-experimentation toolkit for behavior change. In *Personal Informatics in the Wild: Hacking Habits for Health & Happiness at the ACM-CHI Conference*. ACM, 2013.

[85] Miguel A Hernán and Sonia Hernández-Díaz. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials*, 9(1):48–55, 2012.

[86] Mieke Heyvaert and Patrick Onghena. Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*, 3(1):51–64, 2014.

[87] Alexis Hiniker, Sungsoo Ray Hong, Tadayoshi Kohno, and Julie A Kientz. Mytime: Designing and evaluating an intervention for smartphone non-use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4746–4757. ACM, 2016.

[88] Bradley E. Huitema, Ron Van Houten, and Hana Manal. Time-series intervention analysis of pedestrian countdown timer effects. *Accident Analysis & Prevention*, 72:23–31, 2014.

[89] Girardin Jean-Louis, Daniel F Kripke, Roger J Cole, Joseph D Assmus, and Robert D Langer. Sleep detection with an accelerometer actigraph: Comparisons with polysomnography. *Physiology & behavior*, 72(1-2):21–28, 2001.

[90] Murray W Johns. A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. *Sleep*, 14(6):540–545, 1991.

[91] Paula Johnson, Therese Fitzgerald, Alina Salganicoff, Susan F Wood, and Jill M Goldstein. Sex-specific medical research: Why women's health can't wait. *A report of the Mary Horrigan Connors Center for Women's Health & Gender Biology at Brigham and Women's Hospital. Brigham and Women's Hospital*, 2014.

[92] Tero Jokela, Jarno Ojala, and Thomas Olsson. A diary study on combining multiple information devices in everyday activities and tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3903–3912. ACM, 2015.

[93] W Paul Jones. Single-case time series with Bayesian analysis: A practitioner's guide (methods, plainly speaking). *Measurement and evaluation in counseling and development*, 36(1):28–40, 2003.

[94] Ravi Karkar, Jessica Schroeder, Daniel A Epstein, Laura R Pina, Jeffrey Scofield, James Fogarty, Julie A Kientz, Sean A Munson, Roger Vilardaga, and Jasmine Zia. TummyTrials: A feasibility study of using self-experimentation to detect individualized food triggers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6850–6863. ACM, 2017.

[95] Ravi Karkar, Jasmine Zia, Roger Vilardaga, Sonali R Mishra, James Fogarty, Sean A Munson, and Julie A Kientz. A framework for self-experimentation in personalized health. *Journal of the American Medical Informatics Association*, 23(3):440–448, 2015.

[96] Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel Watson, Sunny Consolvo, and Julie A Kientz. Lullaby: A capture & access system for understanding the sleep environment. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 226–234. ACM, 2012.

[97] Matthew Kay, Gregory L Nelson, and Eric B Hekler. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4521–4532. ACM, 2016.

[98] Joseph Jofish Kaye, Mary McCuistion, Rebecca Gulotta, and David A Shamma. Money talks: Tracking personal finances. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 521–530. ACM, 2014.

[99] Christina Kelley, Bongshin Lee, and Lauren Wilcox. Self-tracking for mental wellness: Understanding expert perspectives and student experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 629–641. ACM, 2017.

[100] John M Kelley and Ted J Kaptchuk. Group analysis versus individual response: The inferential limits of randomized controlled trials. *Contemporary clinical trials*, 31(5):423–428, 2010.

[101] Ian Kerridge. Altruism or reckless curiosity? A brief history of self experimentation in medicine. *Internal medicine journal*, 33(4):203–207, 2003.

[102] Young-Ho Kim, Eun Kyoung Choe, Bongshin Lee, and Jinwook Seo. Understanding personal productivity: How knowledge workers define, evaluate, and reflect on their productivity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 615. ACM, 2019.

[103] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. TimeAware: Leveraging framing effects to enhance personal productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 272–283. ACM, 2016.

[104] Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. OmniTrack: A flexible self-tracking approach leveraging semi-automated tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):67, 2017.

[105] Abby C King, Robert Friedman, Bess Marcus, Cynthia Castro, LeighAnn Forsyth, Melissa Napolitano, and Bernardine Pinto. Harnessing motivational forces in the promotion of physical activity: the community health advice by telephone (chat) project. *Health education research*, 17(5):627–636, 2002.

[106] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3063–3072, 2011.

[107] Predrag Klasnja, Beverly L Harrison, Louis LeGrand, Anthony LaMarca, Jon Froehlich, and Scott E Hudson. Using wearable sensors and real time inference to understand human recall of routine activities. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 154–163. ACM, 2008.

[108] Ping-Ru T Ko, Julie A Kientz, Eun Kyoung Choe, Matthew Kay, Carol A Landis, and Nathaniel F Watson. Consumer sleep technologies: A review of the landscape. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, 11(12):1455, 2015.

[109] Judy Kopp. Self-monitoring: A literature review of research and practice. In *Social Work Research and Abstracts*, volume 24, pages 8–20. Oxford University Press, 1988.

[110] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. Rotating online behavior change interventions increases effectiveness but also increases attrition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):95, 2018.

[111] Thomas R Kratochwill, John H Hitchcock, Robert H Horner, Joel R Levin, Samuel L Odom, David M Rindskopf, and William R Shadish. Single-case intervention research design standards. *Remedial and Special Education*, 34(1):26–38, 2013.

[112] Quantified Self Labs. About the quantified self, 2012. Retrieved September 08, 2015 from `http://quantifiedself.com/about/`.

[113] Gary P Latham. Goal setting: A five-step approach to behavior change. *Organizational Dynamics*, 32(3):309–318, 2003.

[114] Dan Ledger and Daniel McCaffrey. Inside wearables: How the science of human behavior change offers the secret to long-term engagement. *Endeavour Partners*, 200(93):1, 2014.

[115] Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman, and Eric B Hekler. Self-experimentation for behavior change: Design and formative evaluation of two approaches. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6837–6849. ACM, 2017.

[116] Ian Li, Anind Dey, and Jodi Forlizzi. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 557–566, New York, NY, USA, 2010. ACM.

[117] Ian Li, Anind Dey, and Jodi Forlizzi. Understanding my data, myself: Supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 405–414. ACM, 2011.

[118] Zilu Liang and Bernd Ploderer. Sleep tracking in the real world: A qualitative study into barriers for improving sleep. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, pages 537–541. ACM, 2016.

[119] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized medicine*, 8(2):161–173, 2011.

[120] John A List and Steven D Levitt. What do laboratory experiments tell us about the real world. *NBER working paper*, pages 14–20, 2005.

[121] Localytics. Mobile apps: What's a good retention rate? `http://info.localytics.com/blog/mobile-apps-whats-a-good-retention-rate`, Mar 2018.

[122] Edwin A Locke and Gary P Latham. *A theory of goal setting & task performance.* Prentice-Hall, Inc, 1990.

[123] J Derek Lomas, Jodi Forlizzi, Nikhil Poonwala, Nirmal Patel, Sharan Shodhan, Kishan Patel, Ken Koedinger, and Emma Brunskill. Interface design optimization as a multi-armed bandit problem. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4142–4153. ACM, 2016.

[124] Carol A Maher, Lucy K Lewis, Katia Ferrar, Simon Marshall, Ilse De Bourdeaudhuij, and Corneel Vandelanotte. Are health behavior change interventions that use online social networks effective? A systematic review. *Journal of medical Internet research*, 16(2), 2014.

[125] David Mant. Can randomised trials inform clinical decisions about individual patients? *The Lancet*, 353(9154):743–746, 1999.

[126] Ellyn E Matthews, J Todd Arnedt, Michaela S McCarthy, Leisha J Cuddihy, and Mark S Aloia. Adherence to cognitive behavioral therapy for insomnia: A systematic review. *Sleep medicine reviews*, 17(6):453–464, 2013.

[127] John D Mayer and Yvonne N Gaschke. The brief mood introspection scale (BMIS). 1988.

[128] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I Hong. Toss'n'Turn: Smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 477–486. ACM, 2014.

[129] David M. Mount and Sunil Arya. Ann. `http://www.cs.umd.edu/~mount/ANN/`, 2010.

[130] National Heart, Lung, and Blood Institute and National Institutes of Health and U.S. Department of Health and Human Services. Your guide to healthy sleep (NIH Publication No. 06-5271), 2005.

[131] Gina Neff and Dawn Nafus. *Self-tracking*. MIT Press, 2016.

[132] Rosemery O Nelson and Steven C Hayes. Theoretical explanations for reactivity in self-monitoring. *Behavior Modification*, 5(1):3–14, 1981.

[133] Allen Neuringer. Self-experimentation: A call for change. *Behaviorism*, 9(1):79–94, 1981.

[134] Northcube. Sleepcycle, 2015. Retrieved September 21, 2018 from http://www.sleepcycle.com/.

[135] Business of Apps. Mobile app uninstall rate after 30 days. http://www.businessofapps.com/mobile-app-uninstall-rate-after-30-days-is-28-according-to-appsflyer/, May 2018. Retrieved August 25, 2018.

[136] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

[137] Leysia Palen and Marilyn Salzman. Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 87–95. ACM, 2002.

[138] Jean Paquet, Anna Kawinska, and Julie Carrier. Wake detection capacity of actigraphy during sleep. *Sleep*, 30(10):1362–1369, 2007.

[139] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. PopTherapy: Coping with stress through pop-culture. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, pages 109–117. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.

[140] Sun Young Park and Yunan Chen. Individual and social recognition: Challenges and opportunities in migraine management. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1540–1551. ACM, 2015.

[141] Gemma Phillips, Lambert Felix, Leandro Galli, Vikram Patel, and Philip Edwards. The effectiveness of M-health technologies for improving health and health services: A systematic review protocol. *BMC Research Notes*, 3(1):250, 2010.

[142] Charles P Pollak, Warren W Tryon, Haikady Nagaraja, and Roger Dzwonczyk. How accurately does wrist actigraphy identify the states of sleep and wakefulness? *Sleep*, 24(8):957–965, 2001.

[143] James O Prochaska and Wayne F Velicer. The transtheoretical model of health behavior change. *American journal of health promotion*, 12(1):38–48, 1997.

[144] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164. ACM, 2011.

[145] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. MyBehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings*

of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 707–718. ACM, 2015.

[146] Roni Caryn Rabin. The drug-dose gender gap. *The New York Times*, Jan 2013.

[147] David K Randall. Insomnia: Relax... and stop worrying about lack of sleep. `https://www.theguardian.com/lifeandstyle/2012/sep/22/dreamland-insomnia-sleep-cbt-drugs`, Sep 2012.

[148] Katharina Reinecke and Krzysztof Z Gajos. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1364–1378. ACM, 2015.

[149] Consumer Reports. Why americans can't sleep. `https://www.consumerreports.org/sleep/why-americans-cant-sleep/`, Jan 2016.

[150] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[151] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2011.

[152] Seth Roberts. Self-experimentation as a source of new ideas: Ten examples about sleep, mood, health, and weight. *Behavioral and Brain Sciences*, 27(2):227 – 288, 2004.

[153] Seth Roberts. The unreasonable effectiveness of my self-experimentation. *Medical hypotheses*, 75(6):482–489, 2010.

[154] Seth Roberts. The reception of my self-experimentation. *Journal of Business Research*, 65(7):1060–1066, 2012.

[155] Seth Roberts and Allen Neuringer. Self-experimentation. In *Handbook of research methods in human operant behavior*, pages 619–655. Springer, 1998.

[156] Till Roenneberg, Anna Wirz-Justice, and Martha Merrow. Life between clocks: Daily temporal patterns of human chronotypes. *Journal of biological rhythms*, 18(1):80–90, 2003.

[157] John Rooksby, Parvin Asadzadeh, Mattias Rost, Alistair Morrison, and Matthew Chalmers. Personal tracking of screen time on digital devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 284–296. ACM, 2016.

[158] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[159] Rainer Sachse. The effects of intervention phrasing on therapist-client communication. *Psychotherapy research*, 3(4):260–277, 1993.

[160] Robert L Sack, Dennis Auckley, R Robert Auger, Mary A Carskadon, Kenneth P Wright Jr, Michael V Vitiello, and Irina V Zhdanova. Circadian rhythm sleep disorders: Part I, basic principles, shift work and jet lag disorders. *Sleep*, 30(11):1460–1483, 2007.

[161] Avi Sadeh, Peter J Hauri, Daniel F Kripke, and Peretz Lavie. The role of actigraphy in the evaluation of sleep disorders. *Sleep*, 18(4):288–302, 1995.

[162] Akane Sano, Sara Taylor, Craig Ferguson, Akshay Mohan, and Rosalind W Picard. QuantifyMe: An automated single-case experimental design platform. In *International Conference on Wireless Mobile Communication and Healthcare*, pages 199–206. Springer, 2017.

[163] Hanna Schäfer, Santiago Hors-Fraile, Raghav Pavan Karumur, André Calero Valdez, Alan Said, Helma Torkamaan, Tom Ulmer, and Christoph Trattner. Towards health (aware) recommender systems. In *Proceedings of the 2017 international conference on digital health*, pages 157–161. ACM, 2017.

[164] Jessica Schroeder, Chia-Fang Chung, Daniel A Epstein, Ravi Karkar, Adele Parsons, Natalia Murinova, James Fogarty, and Sean A Munson. Examining self-tracking by people with migraine: Goals, needs, and opportunities in a chronic health condition. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*, pages 135–148. ACM, 2018.

[165] Jessica Schroeder, Ravi Karkar, James Fogarty, Julie A Kientz, Sean A Munson, and Matthew Kay. A patient-centered proposal for Bayesian analysis of self-experiments for health. *Journal of healthcare informatics research*, 3(1):124–155, 2019.

[166] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*, pages 157–180. Springer, 2009.

[167] Steven L Scott. A modern Bayesian look at the Multi-armed Bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

[168] William R. Shadish, David M. Rindskopf, and Larry V. Hedges. The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 2(3):188–196, 2008.

[169] Hangsik Shin, Byunghun Choi, Doyoon Kim, and Jaegeol Cho. Robust sleep quality quantification method for a personal handheld device. *Telemedicine and e-Health*, 20(6):522–530, 2014.

[170] Nalin A Singh, Karen M Clements, and Maria A Fiatarone. A randomized controlled trial of the effect of exercise on sleep. *Sleep*, 20(2):95–101, 1997.

[171] Justin D Smith. Single-case experimental designs: A systematic review of published research and current standards. *Psychological methods*, 17(4):510, 2012.

[172] Timothy Sohn, Kevin A Li, William G Griswold, and James D Hollan. A diary study of mobile information needs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 433–442. ACM, 2008.

[173] Mayo Clinical Staff. Polysomnography (sleep study), 2014. Retrieved August 25, 2018 from `http://www.mayoclinic.org/tests-procedures/polysomnography/basics/definition/prc-20013229`.

[174] Paul C Tang, Joan S Ash, David W Bates, J Marc Overhage, and Daniel Z Sands. Personal health records: Definitions, benefits, and strategies for overcoming barriers to adoption. *Journal of the American Medical Informatics Association*, 13(2):121–126, 2006.

[175] Richard H Thaler and Cass R Sunstein. Nudge: Improving decisions about health, wealth, and happiness, 1975.

[176] John B. Todman and Pat Dugard. *Single-case and small-n experimental designs: A practical guide to randomization tests*. Psychology Press, 2001.

[177] Eric Topol. The A.I. Diet. *The New York Times*, Mar 2019.

[178] Ling-Ling Tsai and Sheng-Ping Li. Sleep patterns in college students: Gender and grade differences. *Journal of psychosomatic research*, 56(2):231–237, 2004.

[179] Sunao Uchida, Kohei Shioda, Yuko Morita, Chie Kubota, Masashi Ganeko, and Noriko Takeda. Exercise effects on sleep physiology. *Frontiers in neurology*, 3:48, 2012.

[180] University of California, San Diego. Galileo: Design and run experiments with people from around the world. `https://galileo-ucsd.org/galileo/home`, 2019.

[181] Niels van Berkel, Jorge Goncalves, Peter Koval, Simo Hosio, Tilman Dingler, Denzil Ferreira, and Vassilis Kostakos. Context-informed scheduling and analysis: Improving accuracy of mobile self-reports. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[182] Perla A Vargas, Melissa Flores, and Elias Robles. Sleep quality and body mass index in college students: The role of sleep disturbances. *Journal of American College Health*, 62(8):534–541, 2014.

[183] Olivia J Walch, Amy Cochran, and Daniel B Forger. A global quantification of "normal" sleep schedules using smartphone data. *Science advances*, 2(5):e1501705, 2016.

[184] Shu-Lin Wang, Young Long Chen, Alex Mu-Hsing Kuo, Hung-Ming Chen, and Yi Shiang Shiu. Design and evaluation of a cloud-based mobile health information recommendation system on wireless sensor networks. *Computers & Electrical Engineering*, 49:221–235, 2016.

[185] Douglas L Weeks, Stacie Borrousch, Andrew Bowen, Lisa Hepler, Mary Osterfoss, Andrea Sandau, and Frank Slevin. The influence of age and gender of an exercise model on self-efficacy and quality of therapeutic exercise performance in the elderly. *Physiotherapy theory and practice*, 21(3), 2005.

[186] Allen B. Weisse. Self-experimentation and its role in medical research. *Texas Heart Institute Journal*, 39(1):51–54, 2012.

[187] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. 'Don't Waste My Time': Use of time information improves focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1729–1738. ACM, 2016.

[188] Martin Wiesner and Daniel Pfeifer. Health recommender systems: Concepts, requirements, technical basics and challenges. *International journal of environmental research and public health*, 11(3), 2014.

[189] Martin Wiesner, Daniel Pfeifer, and Arzu Yilmaz. Satisfying health information needs: A german health exhibition example. In *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on*, pages 1–4. IEEE, 2012.

[190] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 379–388. ACM, 2016.

[191] Joseph Jay Williams, Anna N Rafferty, Dustin Tingley, Andrew Ang, Walter S Lasecki, and Juho Kim. Enhancing online problems through instructor-centered tools for randomized experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018.

[192] Gui-Rong Xue, Chenxi Lin, Qiang Yang, WenSi Xi, Hua-Jun Zeng, Yong Yu, and Zheng Chen. Scalable collaborative filtering using cluster-based smoothing. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 114–121. ACM, 2005.

[193] Ze Zhang, Michael J Peluso, Cary P Gross, Catherine M Viscoli, and Walter N Kernan. Adherence reporting in randomized controlled trials. *Clinical trials*, 11(2):195–204, 2014.

[194] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the Netflix prize. *Lecture Notes in Computer Science*, 5034:337–348, 2008.