SleepBandits: Guided Flexible Self-Experiments for Sleep

Nediyana Daskalova
Brown University

Jina Yoon
Brown University

Lisa Wang
Brown University

Cintia Araujo
Brown University

Guillermo Beltran
Brown University

Nicole Nugent
Warren Alpert Medical School

John McGeary
Providence VA Medical Center

Joseph Jay Williams
University of Toronto

Jeff Huang
Brown University

ABSTRACT

Self-experiments allow people to explore what behavioral changes lead to improved health and wellness. However, it is challenging to run such experiments in a scientifically valid way that is also flexible and able to accommodate the realities of daily life. We present a set of design principles for guided self-experiments that aim to lower this barrier to self-experimentation. We demonstrate the value of the principles by implementing them in SleepBandits, an integrated system that includes a smartphone application for sleep experiments. SleepBandits guides users through the steps of a single-case experiment, automatically collecting data from the built-in sensors and user input and calculating and presenting results in real-time. We released SleepBandits to the Google Play Store and people voluntarily downloaded and used it. Based on the data from 365 active users from this in-the-wild study, we discuss opportunities and challenges with the design principles and the SleepBandits system.

Author Keywords

personal informatics; self-experiments; sleep tracking.

CCS Concepts

•**Human-centered computing** → **Empirical studies in HCI;** *Empirical studies in ubiquitous and mobile computing;*

INTRODUCTION

"Scientific wellness" is an approach championed by Leroy Hood that combines behavior coaching with insights from an individual's genetics to create highly personalized treatments for reaching tailored goals [38]. It suggests that insights from personal data can be more effective than generalized advice. For example, the National Sleep Foundation's recommendation [18] to "go to bed early" is well-intended, but may harm "night owls" who have biologically different circadian rhythms.

In this work, we build on the idea of "scientific wellness," but focus solely on behavioral data from users who conduct "self-experiments." These are experiments in which the experimenter is also the only subject and chooses what behavior to change in order to observe its effects. We developed a set of design principles for such guided self-experiments where the goal is to maximize user agency by identifying interventions that work specifically for the self-experimenter. We implemented these principles in the domain of sleep since it is a focus of commonly conducted self-experiments, and allows for objective measures such as time to fall asleep and awakenings per hour [13, 45]. Sleep problems are estimated to impact over 70 million people in the United States alone, resulting in $50 billion of lost productivity annually [33].

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, developed, and deployed a mobile app called SleepBandits to the Google Play Store (the published name on the app store is SleepCoacher). 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 [11, 12] 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 these issues while nudging towards higher scientific validity. Rather than a classical experimental approach or a randomized controlled trial, our method is user-centric, focusing on incorporating the flexibility people need to conduct an experiment in their daily lives.

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-

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: 10.1145/3313831.3376584

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.

RELATED WORK

Existing Systems and Frameworks for Self-Experiments

Self-experimentation has been applied to a variety of fields and even led to the discovery of metabolism [16]. Paco [17] and Galileo [50] 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 [45] and SleepCoacher [14] 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) [45]. 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 [14]. Furthermore, both systems were evaluated with participants recruited through campus mailing lists which do not represent the general population.

TummyTrials [23], another self-experimentation app with a focus on irritable bowel syndrome, applied a framework for self-experimentation in personalized health [24]. 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 they might 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 [23].

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 they want to try, which variable they want to focus on, as well as for how long they want 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.

Comprehensible Results

An important design consideration identified by existing systems is how to display the self-experiment results to users. Previous studies have used difference of means or p-values to generate results [14, 23, 45]; however, the statistics surrounding null-hypothesis testing can be confusing for the lay audience [12, 46]. Probabilities, on the other hand, have been shown to be easier to understand if reported reliably [35]. However, it is important to acknowledge that probabilities still require a level of numeracy that not all potential users possess.

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 [31], identifying effective explanations and feedback messages [51, 52], activities for mental health [36], and interventions for behavior change [26]. While there are many algorithms for solving these problems, Bayesian approaches of multi-armed bandit algorithms like Thompson Sampling [1, 8] may be more easily interpreted by users [52, 46]. 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.

Actigraphy and Non-Clinical Sleep Studies

If a person wishes to improve their sleep, they can start by following general sleep hygiene guidelines. ShutEye [2] is one previous system which uses the person's phone to display these guidelines in an actionable way. For more serious issues, a patient must undergo an overnight polysomnographic study (PSG), in which they sleep with various monitors and electrodes attached to their bodies [4, 48], but this procedure cannot be performed frequently as it is expensive and obtrusive. 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 [22, 37, 44], a technique which infers sleep and wake states based on the person's movement patterns. 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 [32, 25, 19, 14]. While the accelerometer is the best feature to use when predicting sleep duration [9], it may be less accurate when placed further away from the body. To overcome these shortcomings, SleepBandits employs both accelerometer and microphone amplitude data to estimate sleep metrics such as time to fall asleep.

Overall, existing consumer apps provide mainly descriptive statistics and do not guide users through self-experiments. The system presented in this paper, SleepBandits, is the first system to implement a Bayesian approach to guided self-experiments.

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 [14, 45, 23, 24], 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 novice self-experimenters 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 [14, 23, 39, 45]:

- **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 [12, 14, 23]). This can be accomplished through providing experiment length suggestions, a short-list 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 [12]). 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 [45] 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 [14, 23]). One way to do that is to present probabilities generated from Bayesian analysis [28, 46], such as Thompson Sampling.

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.

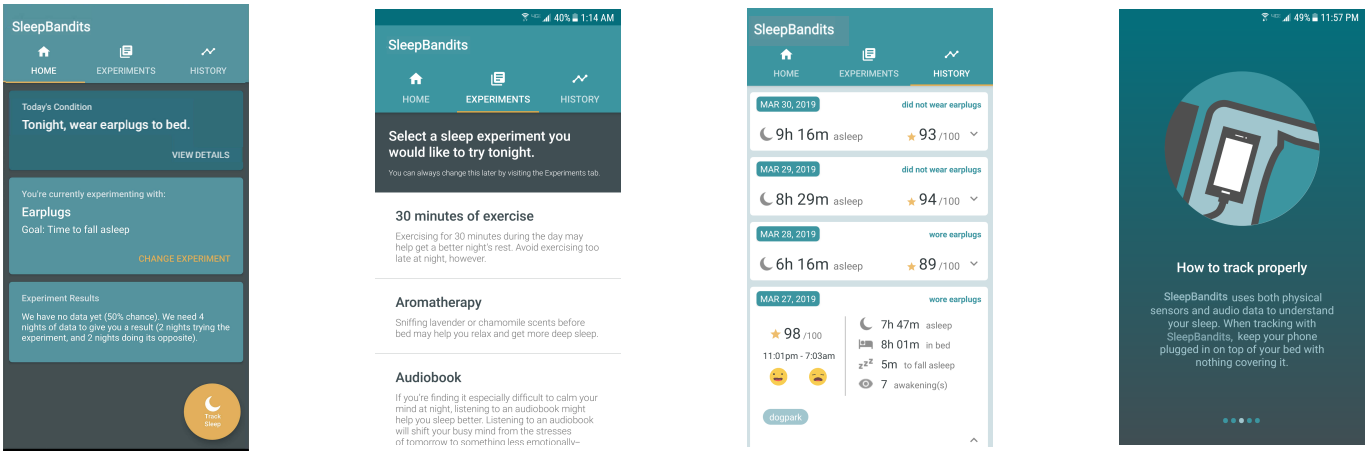
The SleepBandits mobile application was designed to work without any interaction with the researchers and run on various Android OS versions and Android smartphone models. The application collects sleep data by using the device's built-in microphone and accelerometer to track sound amplitude and the user's movements during the night, so the phone must be placed on the bed overnight (Figure 1(d)). 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 1(a)) to begin collecting data. When they wake up in the morning, they tap the "Wake up" button to stop tracking (Figure 2(b)), 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 1(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.

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 1(b). We developed this list by using general sleep hygiene guidelines [18] 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.,



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

Figure 1: 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. (d) Onboarding for new users: explains what to expect from the app and to keep the phone on the bed while sleeping.

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 2(d)). However, once in the app, users can see all 26 in the "Experiments" tab and change to a new one at any time.

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 [44, 5]. 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 [42].

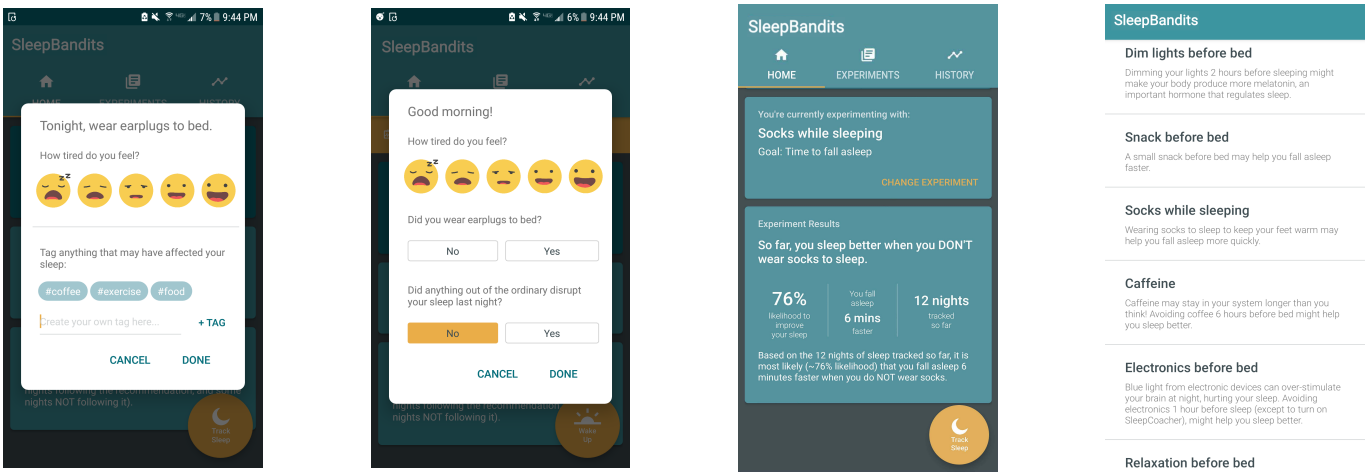
To determine how long a user takes to fall asleep, SleepBandits employs a heuristic from the sleep literature that was previously used in SleepCoacher [14]. The limitation of this heuristic is that it uses a static threshold to determine whether someone is awake or asleep. If a user places their phone 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.

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 1(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 2(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 2(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 2(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



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

Figure 2: (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. (d) The user initially has six curated experiments to choose from.

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 [20], 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.

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 2(c). This design was 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. 2(c)). 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 [12, 47]. 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.

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, users had to comply with at least 5 nights of each condition. We selected these group lengths based on the standards for single-case intervention research design by Kratochwill et al. [27] 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 [30]. 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.

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 [29]. 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 [10, 12], 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% [34], the one-month retention rate is 43% [30], and 21% of users only use an app once [30]. 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 [6] 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 [41] or Citizen Science, but rather than using a survey or online questionnaire, we offer benefits from using a sleep-tracking application. We accept the challenge by Bernstein et al. [3], "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."

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 sleep is worse than the threshold). Figure 3 shows an example with success and failures compared to the threshold.

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., prob=0.7 for audiobook, prob=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). In

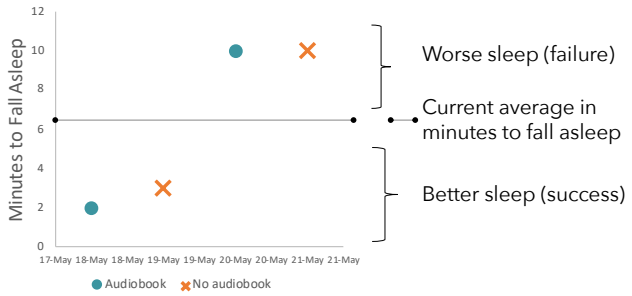


Figure 3: Example where the Thompson Sampling result indicates a success or failure based on whether the time to fall asleep is above or below the average threshold so far.

this case, 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 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 [43].

$$x_t \leftarrow argmax_{x \in \chi} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x] \quad (1)$$

As shown in Figure 4, 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).

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."

Experiment Adherence and Duration

In a rigorous scientific setting and in previous studies [45, 14], 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 2(b), each day, participants were asked whether they followed the app's suggestions. However, as seen in previous research [14, 23, 45], users experience unexpected life events that prevent them from adhering to the correct experimental condition for that night. 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 [14], so this trades off some control in the randomized controlled trial configuration to accommodate natural user behavior.

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.

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 [12]. Unlike participants in the TummyTrials study [23], 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. 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."

Adherence to Instructions: Balancing Scientific Rigor and Everyday Life

Previous studies show that people do not intuitively randomize their conditions [12], even though it helps decrease the effect of confounding variables [21]. 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 such studies ranged from 22.5% in QuantifyMe [45], to 53% in SleepCoacher [14], and 95% in TummyTrials [23]. In comparison, the median adherence reported in randomized controlled

Figure 4: 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%.

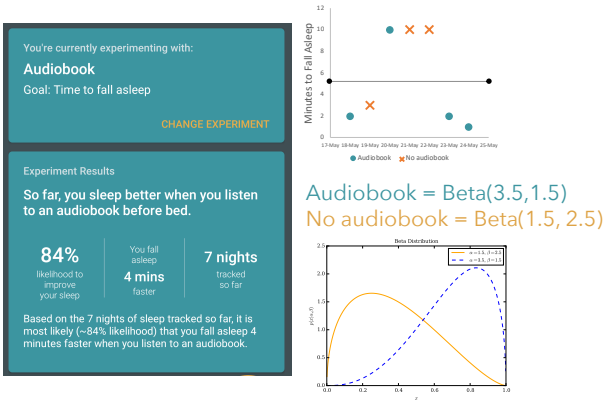


Figure 5: After a few more nights of data, the beta distributions changed shapes for both conditions, and the likelihood audiobooks are helpful is now 84%.

trials was 88.4% (range: 48%–100%) [53]. SleepBandits employed an "as-treated analysis," meaning that 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 [20]. 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.

Effect of Minimum Experiment Length on Completion

In rigorous in-lab studies with systems like SleepCoacher and QuantifyMe [14, 45], which employed AB phase designs, participants were asked to conduct a single self-experiment over the course of weeks (16 days in [45] and 21 days in [14]), before they saw a result. TummyTrials [23] 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 6, 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. For the purposes of this paper, 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 ($p < 0.05$).

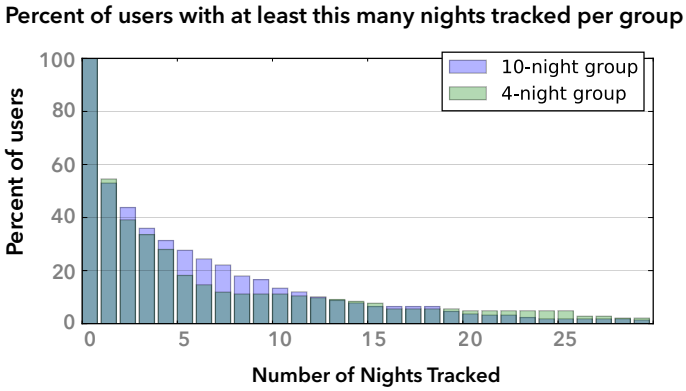


Figure 6: The overall number of nights tracked followed a similar trend in both groups of users. Around the fifth night, however, there is a dip in the percentage of users from the 4-night group that tracked their sleep (which is when those users saw a result in the app).

Reasons for Users Ending the Experiment

Once users reached the required minimum number of days in each condition, they were shown a result (Figure 2(c)), representing the likelihood that the intervention was helping them.

At that point, they were 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.

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.

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 [40].

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.

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.

DISCUSSION

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 results of the experiment 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. SleepBandits calculated the results for users in the shorter group after just 2 days per condition, a duration that might be too short for statistical rigor. With the language and framing of the result 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 keep take this implication into account, as we need to further explore the role of technologies like SleepBandits.

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 6, 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 [15, 7]. 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.

Nudging Users Towards Most Helpful Recommendations

As noted before, participants were more likely to choose one of the first recommendations, 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 [49] which states that the healthiest choices should be those that are 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. As we saw in our results, if users manage to improve their target variable in their first experiment, they are more inclined to conduct another one. Additionally, future systems can even identify cohorts of similar users and recommend experiments that other people comparable to the given user found helpful.

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.

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 [14, 45, 27]. In this work we found that users prefer to conduct shorter self-experiments, but two days per condition might be too short, so we will increase the minimum number of days in the system to 3 per condition (6 nights), as recommended by the standards in [27]. 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.

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.

ACKNOWLEDGEMENTS

This research is funded in part by the Brown University Seed Award, National Science Foundation IIS-1656763, Brown University Data Science Institute, and the Office of Naval Research (#N00014-18-1-2755).

REFERENCES

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

[2] J.S. Bauer, S. Consolvo, B. Greenstein, J. Schooler, E. Wu, N.F. Watson, and J. Kientz. 2012. ShutEye: Encouraging Awareness of Healthy Sleep Recommendations with a Mobile, Peripheral Display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. 1401–1410. DOI:http://dx.doi.org/10.1145/2207676.2208600

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

[4] A. Blaivas. 2014. Polysomnography. (2014). Retrieved August 25, 2018 from http://www.nlm.nih.gov/medlineplus/ency/article/003932.htm.

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

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

[7] Marta E Cecchinato, John Rooksby, Alexis Hiniker, Sean Munson, Kai Lukoff, Luigina Ciolfi, Anja Thieme, and Daniel Harrison. 2019. Designing for Digital Wellbeing: A Research & Practice Agenda. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, W17.

[8] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*. 2249–2257.

[9] Z. Chen, M. Lin, F. Chen, N.D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A.T. Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*. IEEE, 145–152.

[10] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. 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*. ACM, 173–182.

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

[12] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons Learned from Two Cohorts of Personal Informatics Self-Experiments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 46.

[13] Nediyana Daskalova, Nathalie Ford, Ann Hu, Kyle Moorehead, Benjamin Wagnon, and Janet Davis. 2014. Informing Design of Suggestion and Self-Monitoring Tools through Participatory Experience Prototypes. In *International Conference on Persuasive Technology*. Springer, 68–79.

[14] Nediyana Daskalova, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. 2016. SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 347–358.

[15] Gavin Doherty, David Coyle, and John Sharry. 2012. 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*. ACM, 1421–1430.

[16] Garabed Eknoyan. 1999. Santorio Sanctorius (1561–1636)–founding father of metabolic balance studies. *American journal of nephrology* 19, 2 (1999), 226–233.

[17] Bob Evans. 2019. PACO: The Personal Analytics Companion. (2019). https://pacoapp.com/

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

[19] T. Hao, G. Xing, and G. Zhou. 2013. iSleep: Unobtrusive Sleep Quality Monitoring Using Smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys '13)*. Article 4. DOI: http://dx.doi.org/10.1145/2517351.2517359

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

[21] Mieke Heyvaert and Patrick Onghena. 2014. 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 (2014), 51–64.

[22] G. Jean-Louis, D.F. Kripke, R.J. Cole, J.D. Assmus, and R.D. Langer. 2001. Sleep detection with an accelerometer actigraph: comparisons with polysomnography. *Physiology & Behavior* 72, 1 (2001), 21–28.

[23] 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. 2017. 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*. ACM, 6850–6863.

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

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

[26] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. 2018. Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 95.

[27] T. R. Kratochwill, J. H. Hitchcock, R. H. Horner, J. R. Levin, S. L. Odom, D. M. Rindskopf, and W. R. Shadish. 2012. Single-case intervention research design standards. *Remedial and Special Education* (2012), 0741932512452794.

[28] Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman, and Eric B Hekler. 2017. 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*. ACM, 6837–6849.

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

[30] Localytics. 2018. Mobile Apps: What's A Good Retention Rate? (Mar 2018). `http://info.localytics.com/blog/mobile-apps-whats-a-good-retention-rate`

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

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

[33] Lung National Heart, Blood Institute, National Institutes of Health, and others. 2005. Your Guide to Healthy Sleep (NIH Publication No. 06-5271). (2005).

[34] Business of Apps. 2018. Mobile app uninstall rate after 30 days. (May 2018). Retrieved August 25, 2018.

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

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

[37] C. Pollak, W. Tryon, H. Nagaraja, and R. Dzwonczyk. 2001. How Accurately Does Wrist Actigraphy Identify the States of Sleep and Wakefulness? 24, 8 (2001), 957–965.

[38] Nathan D Price, Andrew T Magis, John C Earls, Gustavo Glusman, Roie Levy, Christopher Lausted, Daniel T McDonald, Ulrike Kusebauch, Christopher L Moss, Yong Zhou, and others. 2017. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature biotechnology* 35, 8 (2017), 747.

[39] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. 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*. ACM, 707–718.

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

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

[42] Consumer Reports. 2016. Why Americans Can't Sleep. (Jan 2016). `https://www.consumerreports.org/sleep/why-americans-cant-sleep/`

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

[44] A. Sadeh, P.J. Hauri, D.F. Kripke, and P. Lavie. 1995. The role of actigraphy in the evaluation of sleep disorders. *Sleep* 18, 4 (1995), 288–302.

[45] Akane Sano, Sara Taylor, Craig Ferguson, Akshay Mohan, and Rosalind W Picard. 2017. QuantifyMe: An Automated Single-Case Experimental Design Platform. In *International Conference on Wireless Mobile Communication and Healthcare*. Springer, 199–206.

[46] Jessica Schroeder, Ravi Karkar, James Fogarty, Julie A Kientz, Sean A Munson, and Matthew Kay. 2019. A Patient-Centered Proposal for Bayesian Analysis of Self-Experiments for Health. *Journal of healthcare informatics research* 3, 1 (2019), 124–155.

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

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

[49] Richard H Thaler and Cass R Sunstein. 1975. Nudge: Improving Decisions About Health, Wealth, and Happiness. (1975).

[50] San Diego University of California. 2019. Galileo: Design and Run Experiments with people from around the world. (2019). https://galileo-ucsd.org/galileo/home

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

[52] Joseph Jay Williams, Anna N Rafferty, Dustin Tingley, Andrew Ang, Walter S Lasecki, and Juho Kim. 2018. Enhancing Online Problems Through Instructor-Centered Tools for Randomized Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 207.

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