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

Nediyana Daskalova
Computer Science
Brown University
nediyana@cs.brown.edu

Danaë Metaxa-Kakavouli*
Computer Science
Stanford University
metaxa@stanford.edu

Adrienne Tran*
Computer Science
Brown University
adrienne_tran@brown.edu

Nicole Nugent, Julie Boergers
Brown University and
Bradley Hasbro Research Center
{nnugent,jboergers}@lifespan.org

John McGeary
Providence VA Medical Center
and Brown University
john_mcgeary@brown.edu

Jeff Huang
Computer Science
Brown University
uist@jeffhuang.com

ABSTRACT

We present SleepCoacher, an integrated system implementing a framework for effective self-experiments. SleepCoacher automates the cycle of single-case experiments by collecting raw mobile sensor data and generating personalized, data-driven sleep recommendations based on a collection of template recommendations created with input from clinicians. The system guides users through iterative short experiments to test the effect of recommendations on their sleep. We evaluate SleepCoacher in two studies, measuring the effect of recommendations on the frequency of awakenings, self-reported restfulness, and sleep onset latency, concluding that it is effective: participant sleep improves as adherence with SleepCoacher's recommendations and experiment schedule increases. This approach presents computationally-enhanced interventions leveraging the capacity of a closed feedback loop system, offering a method for scaling guided single-case experiments in real time.

Author Keywords

Personal informatics; self-experiments; sleep recommendations; sleep monitoring; mobile devices

ACM Classification Keywords

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

INTRODUCTION

Over 40 million people in the United States suffer from long-term sleep disorders, and an additional 20 million suffer from occasional sleep problems [1], many of whom could potentially improve their sleep by changing certain behaviors, but do not know how. The most common way to improve sleep is to follow generic sleep hygiene recommendations which, while helpful, neglect individual variation. For example, some people need more sleep than others, some are night owls while others are early birds, and some are more sensitive to noise.

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 [9], making the smartphone an ideal form factor for sleep monitoring. Indeed, widespread use of smartphones to track aspects of personal health, including sleep, are on the rise. Tens of 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 [3, 23]. Users of such apps are receptive to recommendations about behaviors preceding sleep to improve their sleep hygiene [2, 5].

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 behavioral recommendations for improving sleep based on sleep sensing data.

The SleepCoacher system is compatible with sleep sensing apps for mobile devices, including a modified commercial app and one developed by the authors. We evaluate SleepCoacher in a preliminary four-week exploratory study and final six-week study. In both studies, participants placed a smartphone on their bed to collect movement and noise

*These authors contributed equally.

UIST 2016, October 16 - 19, 2016, Tokyo, Japan
Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-4189-9/16/10 ...$15.00 DOI: http://dx.doi.org/10.1145/2984511.2984534

data when sleeping. After analyzing this data for potential interventions, the SleepCoacher system sent each participant a text message encouraging a specific sleep behavior change based on correlations in each user's own sleep data. The SleepCoacher framework is closed-loop; after providing recommendations, the system uses data from subsequent nights of sleep to determine whether a behavior change occurred and yielded improvements in targeted aspects of sleep including frequency of awakenings during sleep, self-reported restfulness rating, and sleep onset latency (time to fall asleep).

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. We find that participant sleep improves as adherence to SleepCoacher's recommendations increases.

RELATED WORK

This paper connects existing clinical practices and computational work in the realm of sleep to two branches of research we currently see as open loops, personal informatics and persuasive technology, automating the single-case experiment process to evaluate the effectiveness of data-driven sleep recommendations.

Actigraphy and Polysomnography

In the domain of sleep improvement, existing professional forms of sleep monitoring use specialized equipment to improve detection of some sleep events, though these methods are costly and require professional oversight.

Polysomnography (PSG) is the traditional method of sleep monitoring used to detect sleep disorders [29]. 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 [7, 32]. Although this is a noninvasive procedure, it is obtrusive, costly, and cannot be conducted frequently.

Actigraphy involves a user-worn electronic device, and has long been a common method for sleep tracking [16, 26, 27]. These existing medical methods are insufficient, however; they are expensive, may not allow users to sleep in a naturalistic setting, and require professional expertise in data analysis and interpretation.

As a result of these shortcomings, trackers such as the FitBit and Jawbone UP use accelerometers as lower quality actigraphy devices to detect movement in the wrist as a proxy for sensing asleep or awake states. They simply track data and maybe compute correlations, but they do not give recommendations or evaluate their effectiveness.

Personal Informatics

Personal informatics is a class of tools that help people collect data for self-monitoring. Early work by Killingsworth and Gilbert, for example, investigated factors involving happiness by developing an iPhone application for people to track their feelings and actions [18]. At a base level, personal informatics tools track data about peoples' lives. Health Mashups has expanded this work by building a tool that detects correlations between different factors in users' lives [6]. While some users found correlations insightful, others found them spurious or obvious. Our work proposes to turn correlations measured on key metrics into actionable, personalized recommendations. Prior work focusing specifically on sleep-related personal informatics, such as Lullaby, has been limited to simply collecting data and displaying it to users so they can look for trends on their own [17]. As with Lullaby, other systems have not developed rigorous methodologies to make recommendations with collected data.

The increased popularity of personal informatics in various aspects of health has led to the use of smartphones in sleep tracking. Sleep can be monitored using a smartphone accelerometer, which can be as accurate as an actigraph accelerometer for many sleep metrics, with the exception of sleep onset latency [22]. Appropriate algorithms, however, are needed to classify sleep and wake states based on actigraphy [28]. While there is currently a lack of prescriptive technology making recommendations based on sleep monitoring data [9], people are interested in recommendations to improve their sleep, such as sleep hygiene guidelines [5]. Handling raw sensor data is challenging for users, who depend on tools to view and interpret data. iSleep is a system that uses smartphone microphone data to detect sleep events overnight [14], while Toss 'n' Turn uses smartphone-collected data to train classifiers that detect sleep and predict sleep quality [21]. Other systems use various smartphone sensors to detect the total number of hours slept by a user; existing literature reports that accelerometer data is the best feature for accurate sleep duration estimation [8].

Our work builds on the aforementioned approaches, combining the mobile devices' sleep monitoring capabilities with sleep sensing techniques from prior research, while working with clinicians to offer actionable and personalized recommendations to users in a scalable way.

Persuasive Technology

Persuasive technology aims to promote changes in users' behaviors or attitudes [13]. Researchers often try to change behavior based on a set of generic guidelines, for example to prompt smoking cessation [12].

One such behavior change system, ShutEye, focuses on displaying sleep hygiene guidelines on a user's mobile phone home screen [5]. 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 [10]. With SleepCoacher we aim to address the lack of personalized tools providing actionable clinician-based feedback on sleep.

Single-Case Experimental Design

Single-case experimental designs allow researchers to evaluate the effectiveness of an intervention on a single participant [15]. Since our recommendations are personalized, each participant in the study is the subject of a single-case design, where the intervention is the action recommended by the SleepCoacher system. Kratochwill et al. outline the standards for single-case intervention research designs to which this research adheres [19].

These standards were compiled by a panel of experts on quantitative methods and single-case design methodology, and suggest that the best design for single-case experiments is an AB phase design, where the A phases correspond to the baseline, and the B phases to intervention periods. The standards suggest a minimum of three attempts to demonstrate the intervention effect, and therefore at least 4 phases ($ABAB$). Each of the phases must have at least 3–5 data points (i.e. 3–5 nights of sleep). The two B phases here are identical as the user follows the same recommendation in both, and we combine the data from them when analyzing study results. We evaluate SleepCoacher following these guidelines. Notably, this individual focus leads us to concentrate our evaluation not on aggregate statistical significance, which is less meaningful for small-scale personalized data collection, but rather to identify whether each of our single-case experiments demonstrated improvement.

Different experimental designs each have unique tradeoffs. The AB design, for example, is susceptible to confounding variables, making conclusions difficult. The ABAB design provides the ideal trade-off between enough days for users to acclimate to recommendations and the least bias. A multiple baseline test would not be appropriate either because it would require 2 of 3 variables (subject, behavior, setting) to remain constant. Keeping the setting fixed would have been nearly impossible for students as their schedules and workload changes weekly, introducing new confounding variables.

Integrated Feedback Loops

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 a integrated closed loop model by tracking the effects of personalized feedback over time.

SLEEPCOACHER

Our integrated system, SleepCoacher, combines automated data collection using smartphones with input from professional clinicians to collect user data and, in return, send daily sleep 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 or not 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 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.

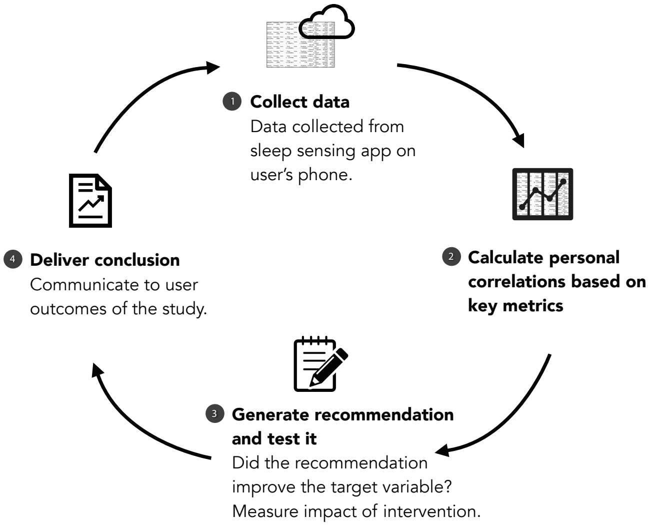
1 Collect data
Data collected from sleep sensing app on user's phone.
2 Calculate personal correlations based on key metrics
3 Generate recommendation and test it
Did the recommendation improve the target variable? Measure impact of intervention.
4 Deliver conclusion
Communicate to user outcomes of the study.

Figure 1. SleepCoacher employs a closed feedback loop: a user's data is uploaded to the cloud and presented to clinicians in the form of charts and correlation tables. Next, clinicians give recommendations, which are sent back to users, who adjust their sleep habits accordingly.

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.

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) [3]. 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 experiment, removing visualizations and extra options that could confuse users or influence their usage of the app and perception of recommendations and changing the frequency with which noise data is collected.

The application collects bed and wake times, accelerometer movement data at 10-second intervals, microphone noise levels at approximately 5–10 minute intervals, the user's self-reported rating of how refreshed they felt upon waking up, times of any alarms set and snoozed, 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 [4, 25]. Our algorithms take raw sensor data as input and record as active any movement with acceleration over $0.98 m/s^2$. Data is labeled "awake" if more than one activity occurred in the previous 2 minutes, and "asleep" at the beginning of a period with no active movement for 20 minutes. Upon waking up, users stop tracking by manually indicating they are awake and are then given the opportunity to enter a rating to how refreshed they felt as well as to add pre-defined or personal tags with the tap of a button. 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. We then compute Pearson correlations to determine which intervention suggestion to send to each user from a collection of recommendations provided 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.

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 actively engaged in research on 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 collected feedback from these clinicians in two different ways. For our Preliminary Study, which served as a pilot for the iterative recommendation process, we generated statistical visualizations for each user's data, adjusting data presentation to mimic actigraph sensor data, in a format intuitive for the clinicians (see Figure 2). These visualizations displayed noise overlaid on movement, vertically aligning multiple nights of a user's sleep to compare variation in a week's nights of sleep. This platform enabled clinicians to provide recommendations by comparing the relationships between our independent variables (including sleep schedule, movement, and noise), and our dependent variables (ratings, awakenings, and sleep onset latency). Though it is not scalable, this process taught us that clinical insights could be pattern-matched into a collection of recommendations.

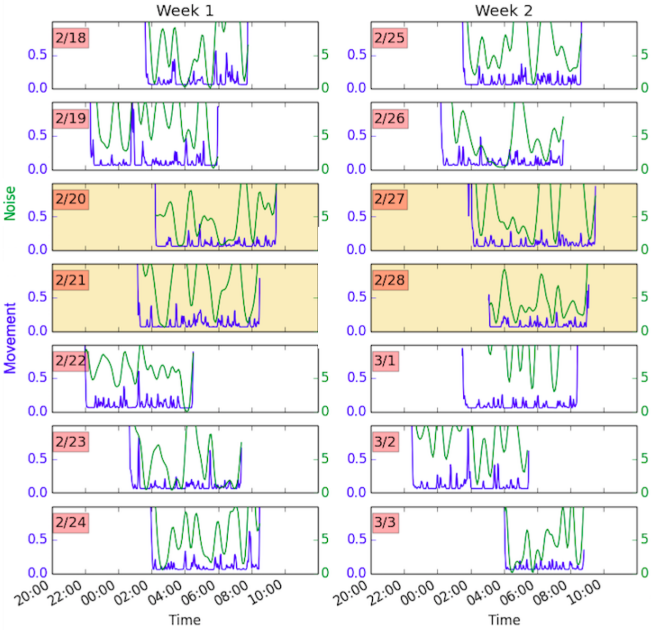

Figure 2. For ease of analysis by clinicians in the Preliminary Study, movement and noise data for each user was combined into a single visualization of all nights. The date of each night is labeled, and weekends are highlighted against a pale yellow background.

In the Preliminary study, clinicians provided recommendations for each user on a per correlation basis. 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 in order of quality, and each rank was integrated into the system as a weight on the likelihood of a user receiving the corresponding recommendation.

Collection of Recommendation Templates

In the second study, we focused on three dependent variables that were measurable: 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 dimensions of lifestyle: environment (specifically factors affecting sleep such as light and noise); physical state (including health, 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. Recommendations had to be measurable (easy to observe and tag), easy for users to comply, and empirically supported by prior sleep research. These criteria are ordered from most to least important. 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.

Ranking and Recommendation Selection Algorithm

Once we had an extensive list of recommendation templates, sleep clinicians individually ranked the generated recommendations by importance. Recommendations were only included in the final collection when all three clinicians were able to agree on ranking. We created a collection of templates for the 114 possible correlations of independent and dependent variables; based on the aforementioned procedure, each possible combination was mapped to at least one potential recommendation.

On the day before a recommendation was due to be delivered, SleepCoacher calculated the 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. If more than one recommendation was possible, the algorithm factored in the weight of each one. This allowed us to assign suggestions algorithmically by weighting relevant recommendations according to clinician ranking. Using this system, the top recommendation for each independent-dependent variable combination had a 75% chance of being selected.

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, number of alarm rings). Optimal hours of sleep for each individual were determined by taking the population of data points where the restfulness rating was 4/5 or more and determining the average hours slept at that high rating.

Correlations in the Experimental Setup

SleepCoacher aims to provide personalized recommendations, since every person has different responses to given sleep recommendations, as well as different natural sleep patterns. The methodology we are presenting allows users to conduct small-scale experiments on their own sleep, adjusting various independent sleep factors and allowing SleepCoacher to learn and improve its recommendations based on the results, in a rapid feedback cycle.

To analyze participant data, each independent variable of sleep behavior and dependent variable representing a sleep outcome are correlated. SleepCoacher computes Pearson correlations and performs statistical tests on these sleep factors (see Figure 6 later for examples). We considered an approach that leveraged Bayesian Statistics and Support Vector Machines to better tune our system, but found that they were unnecessarily complex for self-experiments and did not provide appropriate information about relationship variables. While correlations are simple, they are a powerful measure of the relationship between the independent and dependent variables.

USER STUDIES

We performed two studies: the Preliminary Study (an exploratory study of 28 continuous nights), and a longer 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.

For both studies, we recruited undergraduate students over the age of 18 who use Android smartphones (version 2.2+) as their primary mobile device. We restricted participants to those without medical barriers that would put them at risk or diagnosed sleep problems that might prevent them from participating in our interventions. The two sets of study participants were disjoint.

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 any other comments about the experience. Participants were also asked whether and (if so) how they felt participating had affected their sleep habits. Participants who tracked their sleep for at least 80% of the duration of the study and completed the exit survey were paid $50; those who did not meet these standards were paid $25.

Preliminary Study

The participants, 11 women and 13 men, were all undergraduate students between 18 and 22 years of age. Of our 24 participants, 22 recorded their sleep for at least 80% of the duration of the study, and the remaining two were excluded from data analysis.

After about 20 days of simply tracking their sleep to establish baselines, participants also started receiving recommendations based on their individual data. Each participant received a total of three recommendations—one every three days until the end of the study. We chose this interval to account for the time it takes for behavior change to be reflected in a user's quality of sleep.

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 shows the study setup based on the single-case design (SCD) standards format of the *ABAB* phase design, 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 previous 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 round the study 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.

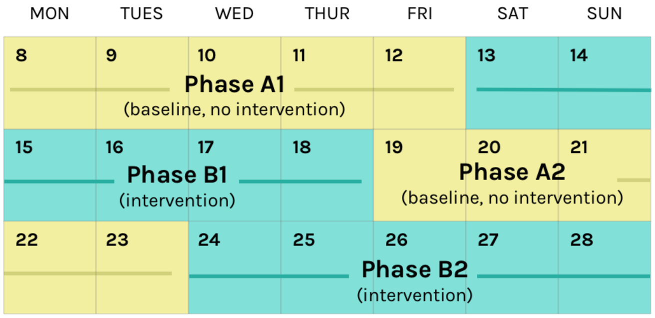

Figure 3. In the *ABAB* phase design of our Final Study, *A* phases (yellow) were non-intervention days, and *B* phases (blue) were intervention days.

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.

Recommendations and Daily Feedback

In both studies, participants were asked to track their sleep every night study 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 1 includes two of those options.

Example Final Study Scenario

In phase A1, a participant tracks her sleep and adds comments and ratings. On the fifth day, SleepCoacher computes correlations and finds the highest one of 0.7 between bedtime and onset latency. The system looks through the recommendation templates for those 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."

FINDINGS

Most Common Recommendations

The most common recommendation in the Preliminary Study was the independent-dependent variable pair of noisiness to awakenings: 21 of the 22 participants received it during the study. The second most common was hours slept and rating (received by 11/22). In the Final Study, the two most common recommendations, both sent to only 5/17 people, were: hours slept → rating, and noisiness → sleep onset latency. As the different frequencies of these recommendations reflect, SleepCoacher sent a greater diversity of suggestions in the Final Study. Another recommendation in the Preliminary Study was weekend sleep and rating, prompting users to change their week sleep to be more like their weekend sleep. However,

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 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.

its compliance rate was too low and did not lead any actionable change, so we did not use it in the Final Study.

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. However, in the Preliminary Study we simply trusted participants when they said they followed or not followed a given recommendation.

In 16 of the 32 cases where recommendations were followed in the Preliminary Study, we saw improvement over the course of three nights of sleep in the key metrics targeted by the recommendation. This improvement, however, 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, as seen in Table 1. Since each of the 17 participants received two recommendations, we had 34 cases to observe the effect of a recommendation 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 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 sleep variables were improved in all 7 of the cases when adherence was higher than 80%.

Compliance Rate and Reasons for Non-Adherence

In both studies, users used the app on average for 94–95% of the nights (0.5–0.6 SD). Similarly, users rated their sleep an average of 85-88% of the nights. A slightly higher percentage of Final Study users used the app for more than 95% of the nights although the study was 2 weeks longer—11 of 17 users in the Final compared to 13 of 22 in the Preliminary study.

In the Preliminary Study's exist survey, many participants confessed to not following the recommendations in the first study (this is expected, since participation was not mandatory). In the Final Study's exist 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 account for that possibility, potentially by suggesting recommendations that do not necessarily concern exact and drastic changes, but rather start with incremental improvements.

Individual Differences in Correlations

Research has shown that individuals show great variation in which key factors influence sleep and other aspects of life quality [6]. Figure 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 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 the 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 in a future system, before accumulating sufficient personal data for a user, the 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.

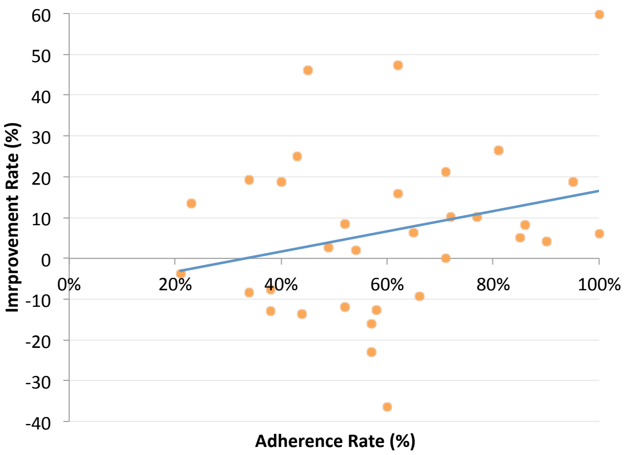

Figure 4. The more a participant adhered to the suggested experimental outline in the Final Study, the more their target sleep variable improved. All participants with adherence rate higher than 80% improved their sleep.

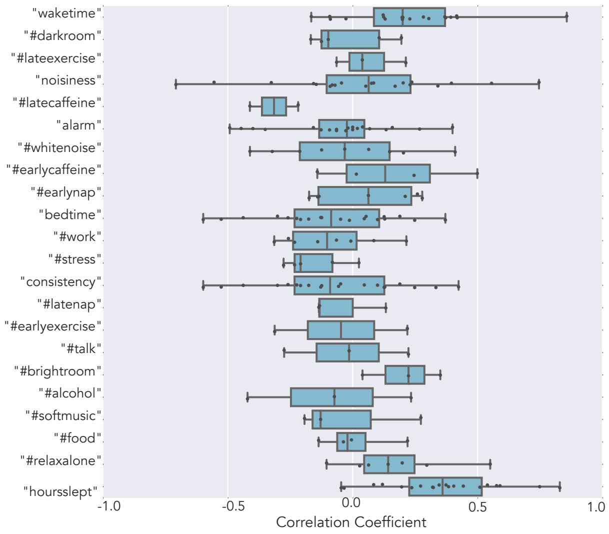

Figure 5. Aggregate rating correlations across all participants show large individual variation for some variables, but not for others. Every dot is a user in our study. Each bar represents the lower bound, first quartile, second quartile (median), third quartile, and upper bound, respectively. The variables with "#" are either pre-defined or personal tags.

Participant Perspectives

We conducted exit surveys following each study. Overall, users felt their sleep habits were positively influenced by SleepCoacher. Even users who reported making no effort to follow recommendations noted that they were more aware of their sleep habits and the influence their daily activities on sleep, which is consistent with previous research on self-monitoring and suggestions [10].

In the Final Study's survey, participants were also asked how personalized they thought each of the recommendations they received was on a scale from 1 as the least personalized to 5 as the most. The average score was 2.94 (SD 0.9) for the first round of recommendations, and 3.76 (SD 0.66) for the second round. This further strengthens the intuition that as we collect more data for users, they receive recommendations that they are increasingly able to recognize as personalized. The first recommendation was based on 5 or 6 nights of sleep, whereas the second was based on all the data until then (about 26 nights).

In the Preliminary Study, we also asked participants whether they thought each recommendation improved their sleep. In 20 of the 34 recommendations, participants felt an improvement when following the recommendation. However, the data showed an improvement in only 11 of

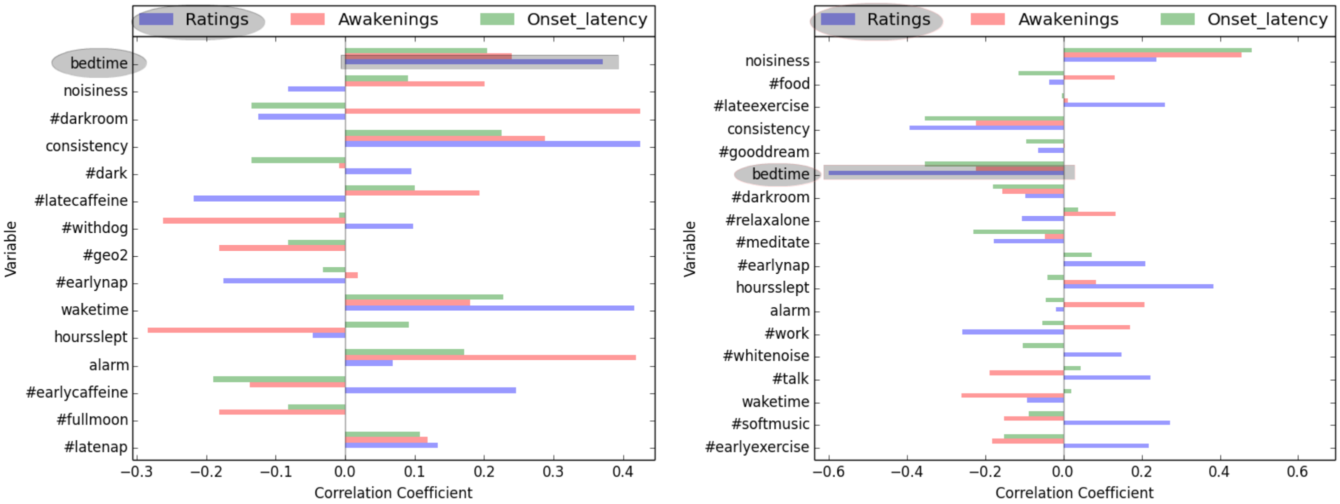

Figure 6. There is large individual variation across correlations between independent and dependent variables. Here, the sleeper on the left has a strong positive correlation between bedtime and rating, whereas the one on the right has a strong negative correlation for the same variables.

these cases. Conversely, in the other 14 cases, participants felt like there was no improvement in their sleep, but the data pointed to an improvement in 8 of them. This is a common finding with psycho-physiological research [24, 30]. Participants may not subjectively realize that they are better or worse on some index. Inability to recognize a subjective difference, in other words, does not necessarily reflect objective improvements.

Areas for Improvement

Feedback from participants 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, 7/22 participants asked for more specific and personalized recommendations, and an additional 2 said they would prefer to receive recommendations more frequently. The lack of concrete metrics drawn from participants' 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 specific and personalized by adding some actual statistics for the user's sleep as can be seen in Table 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 1. In the Final Study, 4/17 participants said the daily feedback would be better if they included more than just 1 statistic, but rather combined all the information we had for the previous night. In the future we are planning to experimentally determine an optimal level of information that remains informative, yet not overwhelming. On the other hand, there was one participant who said he would rather receive a weekly summary rather than daily feedback texts. This suggests the need for personalization of the frequency of sleep feedback.

Further suggestions for improving the daily feedback were related to phrasing, adding more diverse feedback (we had only four possible feedback texts, one of which was randomly selected each day), and adding immediate feedback on whether the person followed the recommendation last night or not. One participant suggested, "Change 'make sure to follow your recommendation tonight' to 'continue to follow your recommendation tonight' if the person followed it the previous night. It made me worry that I'd forgotten to tag my sleep with #earplugs."

DISCUSSION

Helping Users Help Themselves

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 Preliminary Study, users received recommendations every third day, and so effects may have compounded across recommendations. Recommendations were also sent later in the evening, not giving time for advance planning. Thus, in order to meet the single-case design standards and address these challenges, we designed the Final Study around 5-day phases.

In the Final Study, participants conducted small-scale personal experiments, altering an attribute related to their sleep 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. The best improvements in sleep quality are observed when the participant adheres to the suggested study design at least 90% of the time.

Significance in Personalized Micro-Experiments

To evaluate SleepCoacher in our Final Study we observed whether each participant's target variable improved after following the recommendation in contrast to when the user did not. The summary text sent to that user was based solely on this result, regardless of statistical significance, as statistical significance is a less relevant metric with a relatively few (21) data points. According to single-case design literature, a better measurement of the effect of the intervention is a calculation of the effect size [31, 15]. Hedge's g is a standardized-mean differences approach used to compute effect size for single-case designs [31]. Data from such studies is autocorrelated, but according to Manolov and Solanas, this kind of effect size calculation is least affected by autocorrelation [20]. A future system could calculate the Hedge's g effect size and 95% confidence interval, which shows that a result will be in the interval with probability of 0.95 for repeated experiments. If the effect size is larger than 0.5, it is considered "medium" [11], and combined with confidence bounds within the range for a dependent variable's improvement, this strongly suggests causality [33].

Using Hedge's g effect size in the Final Study, 2 of the 7 cases with adherence higher than 80% show a causal relationship and none of the 9 cases with adherence between 60% and 80% show causality, further supporting the claim that participants following study design are more likely to see effects. However, it also shows correlations are not always causal and that experimentation is necessary for determining causality.

Empowering Users through Computation

As with any automated system, attempts to force changes in user behavior may quickly be perceived as annoying and as a result fall into disuse. Instead of using a prescriptive model of feedback, recommendation systems should aim to empower users by helping them become as informed as possible about their own behaviors and the anticipated 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.

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 able to analyze data for 81% of the nights, which contributed to the relatively low adherence rate.

Imperfect tailoring of recommendations occasionally had unintended consequences. One recommendation suggested that users wear earplugs or use a white noise machine to decrease awakenings. One user gently informed us, "I am hearing impaired and take out my hearing aids when I sleep," so this recommendation was inappropriate. The user explained: "when I wake up it is from vibrations in my house from the room below or above me." Anonymous and automated recommendation systems may, with inadequate knowledge about users, provide ineffective or inappropriate suggestions. In order to better support a diversity of users, these systems must be developed conscientiously, with the flexibility to accommodate such differences.

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 sleep 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 benefit, specifically seeing improvements of sleep hygiene including 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.

Clinicians seek to tailor general health guidelines to their individual patients, but are limited by reliance on the individual'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 and an empathetic professional clinician's holistic understanding of human needs. Furthermore, the self-experimentation system we develop has the potential to impact other domains, from nutrition to education.

REFERENCES

1. 2014. Brain Basics: Understanding Sleep. (2014). Retrieved September 24, 2015 from http://ninds.nih.gov/disorders/brain_basics/understanding_sleep.htm.

2. 2014. Healthy Sleep Tips. (2014). Retrieved November 23, 2014 from http://sleepfoundation.org/sleep-tools-tips/healthy-sleep-tips.

3. 2015. Sleep as Android. (2015). Retrieved September 24, 2015 from https://sites.google.com/site/sleepasandroid/.

4. S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. Pollak. 2003. The role of actigraphy in the study of sleep and circadian rhythms. American Academy of Sleep Medicine Review Paper. *Sleep* 26, 3 (2003), 342–392.

5. 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

6. F. Bentley, K. Tollmar, P. Stephenson, L. Levy, B. Jones, S. Robertson, E. Price, R. Catrambone, and J. Wilson. 2013. Health Mashups: Presenting Statistical Patterns Between Wellbeing Data and Context in Natural Language to Promote Behavior Change. *ACM Trans. Comput.-Hum. Interact.* 20, 5 (2013), 30:1–30:27. DOI: http://dx.doi.org/10.1145/2503823

7. A. Blaivas. 2014. Polysomnography. (2014). Retrieved April 13, 2015 from http://www.nlm.nih.gov/medlineplus/ency/article/003932.htm.

8. 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.

9. E. K. Choe, S. Consolvo, N. F. Watson, and J. A. Kientz. 2011. Opportunities for Computing Technologies to Support Healthy Sleep Behaviors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. 3053–3062. DOI: http://dx.doi.org/10.1145/1978942.1979395

10. N. Daskalova, N. Ford, A. Hu, Moorehead K., B. Wagnon, and J. Davis. 2014. Informing Design of Suggestion and Self-Monitoring Tools through Participatory Experience Prototypes. In *Proc. Persuasive Tech.* DOI:http://dx.doi.org/10.1007/978-3-319-07127-5_7

11. J. A. Durlak. 2009. How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology* (2009), jsp004.

12. B. S. Fjeldsoe, A. L. Marshall, and Y. D. Miller. 2009. Behavior change interventions delivered by mobile telephone short-message service. *American journal of preventive medicine* 36, 2 (2009), 165–173.

13. B. J. Fogg. 2002. Persuasive Technology: Using Computers to Change What We Think and Do. *Ubiquity* 2002, December, Article 5 (2002). DOI: http://dx.doi.org/10.1145/764008.763957

14. 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, 4:1–4:14 pages. DOI: http://dx.doi.org/10.1145/2517351.2517359

15. M. Heyvaert and P. 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.

16. 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.

17. M. Kay, E. K. Choe, J. Shepherd, B. Greenstein, N. Watson, S. Consolvo, and J. A. Kientz. 2012. Lullaby: A Capture & Access System for Understanding the Sleep Environment. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. 226–234. DOI: http://dx.doi.org/10.1145/2370216.2370253

18. M. A. Killingsworth and D. T. Gilbert. 2010. A wandering mind is an unhappy mind. *Science* 330, 6006 (2010), 932–932.

19. 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.

20. R. Manolov and A. Solanas. 2008. Comparing N= 1 effect size indices in presence of autocorrelation. *Behavior Modification* 32, 6 (2008), 860–875.

21. J. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. 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 (CHI '14)*. 477–486. DOI: http://dx.doi.org/10.1145/2556288.2557220

22. V. Natale, M. Drejak, A. Erbacci, L. Tonetti, M. Fabbri, and M. Martoni. 2012. Monitoring sleep with a smartphone accelerometer. *Sleep and Biological Rhythms* 10, 4 (2012), 287–292. DOI:http://dx.doi.org/10.1111/j.1479-8425.2012.00575.x

23. Northcube. 2015. SleepCycle. (2015). Retrieved September 24, 2015 from http://www.sleepcycle.com/.

24. W. Ooteman, M. Koeter, R. Vserheul, G. M. Schippers, and W. Brink. 2006. Measuring craving: an attempt to connect subjective craving with cue reactivity. *Alcoholism: Clinical and Experimental Research* 30, 1 (2006), 57–69.

25. J. Paquet, A. Kawinska, and J. Carrier. 2007. Wake detection capacity of actigraphy during sleep. *Sleep* 30, 10 (2007), 1362.

26. 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.

27. 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.

28. A. Sadeh, K. M. Sharkey, and M. A. Carskadon. 1994. Activity-Based SleepWake Identification: An Empirical Test of Methodological Issues. *Sleep* 17, 3 (1994), 201–207.

29. H. Shin, B. Choi, D. Kim, and J. Cho. 2014. Robust Sleep Quality Quantification Method for a Personal Handheld Device. *Telemedicine and e-Health* 20, 6 (2014), 522–30.

30. L. L. Sievert, K. Begum, R. Sharmeen, O. Chowdhury, S. Muttukrishna, and G. Bentley. 2008. Patterns of occurrence and concordance between subjective and objective hot flashes among Muslim and Hindu women in Sylhet, Bangladesh. *American Journal of Human Biology* 20, 5 (2008), 598–604.

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

32. Mayo Clinical Staff. 2014. Polysomnography (sleep study). (2014). Retrieved April 13, 2015 from http://www.mayoclinic.org/tests-procedures/polysomnography/basics/definition/prc-20013229.

33. G. M. Sullivan and R. Feinn. 2012. Using effect size-or why the P value is not enough. *Journal of graduate medical education* 4, 3 (2012), 279–282.