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 admitted to a sleep clinic for observation by a physician using costly and obtrusive sensor technology such as polysomnography and actigraphy. In contrast, prior research has shown that people are most interested in sleep monitoring technology that is unobtrusive and does not require the purchase of additional devices [9], making the smartphone an ideal form factor for sleep monitoring. Indeed, widespread use of smartphones to track important aspects of personal health, including sleep, are on the rise. Tens of millions of people have downloaded available 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, however, current app-based solutions lack many of the features which make clinical methods successful, including personalized analysis and professional guidance. Our system, SleepCoacher, addresses this deficit by implementing a self-experimentation framework based on clinician-generated sleep improvement recommendations. SleepCoacher goes beyond the basic 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 smartphone 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 data when sleeping. After analyzing this data for potential interventions, the SleepCoacher system sent each participant a text-based 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: (1) we propose 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) we present SleepCoacher, a system implementing this framework for the purpose of improving sleep.

We evaluate these contributions and 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, 27, 26]. 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, some fitness trackers such as the FitBit and Jawbone UP use accelerometers as lower quality actigraphy devices to detect movement in the user’s wrist as a proxy for sensing whether the user is asleep or awake. 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 information 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 on this work by building a tool that detects correlations between different factors in users’ lives [6]. While some users found correlations to be 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, and smartphones can be as accurate as an actigraph accelerometer for many sleep metrics, with the exception of sleep onset latency [22]. Appropriate algorithms, however, need to be used 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 guidelines for improving sleep hygiene [5]. Handling raw sensor data is challenging for most users, who depend on tools to help them view and interpret data. iSleep is an example of a system that uses smartphone microphone data to detect sleep events during the night [14]. Toss ’n’ Turn uses smartphone-collected data to train classifiers that detect sleep and predict sleep quality from such sensed data [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 best of ubiquitous mobile devices’ sleep monitoring capabilities with the sleep sensing algorithms developed in prior research, and further 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 they 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). 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, and rather seek to identify whether each of our single-case experiments showed improvement.

**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 an 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.

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.

**Sensing and Data Processing**

SleepCoacher’s underlying framework can be applied to sleep improvement on top of any app which already collects motion and noise data. For this study, we worked with developers of a popular 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, coarse location coordinates, 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 [Anonymized] Hospital and a psychiatry and sleep researcher from [Anonymized] Hospital 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 smaller 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.

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 first generated over 100 recommendations based on all possible correlations between independent and dependent sleep variables. The clinicians then 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 after displaying a relevant correlation in behavior.

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.
Sleep clinicians then individually ranked the generated recommendations by importance. Recommendations were only included 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. Recommendations were then distributed to users with frequency relative to the rank of the recommendation. 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.

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 5). 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.

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.

USER STUDIES
We performed two studies: the Preliminary Study (an exploratory study of 28 continuous nights’ duration), 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 the user followed the recommendation, found it helpful, or had any other comments about the experience. Participants were also asked whether and (if so) how they felt participating had improved their sleep habits. Participants who tracked their sleep for at least 80% of the duration of the study and also 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 to send them at 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 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 also tracked participants for a final day in order to round the study to a full three 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.

Each participant received one new recommendation every 21 days, receiving two unique recommendations in total throughout the 6-week study duration.

**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.

**FINDINGS**

**Most Common Recommendations**

The most common recommendation we sent in the Preliminary Study connected noisiness (the independent variable) to awakenings (the dependent variable). Twenty of the twenty-two total participants received it at some point 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 v rating, and noisiness v sleep onset latency. As the very different rates of these frequent recommendations reflect, SleepCoacher sent users a greater diversity of suggestions in the Final Study. Another recommendation in the Preliminary Study was weekend sleep and rating, and it prompted users to change their week sleep to be more like their weekend sleep. However, compliance rate for it 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 or not they wanted to follow the recommendations. When surveyed, users reported following 32 of 66 recommendation cases. There were some recommendations, such as wearing earplugs, for which 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 the user 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
Figure 5. 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 for the user on the left, whereas the one on the right has a strong negative correlation for the same variables.

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
When deciding which key factors influence sleep and other aspects of life quality, research has shown that individuals show great variation [6]. Figure 6 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 5, 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 more refreshed 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.

**Participant Perspectives**

We conducted participant exit surveys following each study. Overall, users felt their sleep habits were positively influenced by their use of 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 are increasingly able to recognize as personalized. The first recommendation was given based on 5 or 6 nights of sleep, whereas the second was given based on all the data we had at that point (about 26).

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 these cases. Conversely, in the other 14 recommendations, participants felt like there was no improvement in their sleep, but the data pointed to an improvement in 8 of these cases. This is a common finding with psychophysiological 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 major 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 explain justify each good 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,
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 small experiments and their outcomes we can give better recommendations to similar users in the future through a rapid feedback cycle. The best results in sleep improvement 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.

We also note that undergraduates are particularly susceptible to inconsistencies in their sleep habits, as much of their schedule is determined by the academic calendar, and that participants in this study were not necessarily intrinsically motivated to improve their sleep. The question of encouraging an intrinsically motivated user to engage in sustained behavior change is out of the scope of this work, as we instead focus on direct and immediate feedback on areas for potential improvement.

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.

Future Work
Our Final Study addressed many of the challenges uncovered during the Preliminary Study, but also pointed to new ones. There are many more ways to personalize recommendations, for instance varying the frequency and time with which recommendations are sent, which we hope will be addressed in future work. Furthermore, based on the most recent participant feedback, systems like SleepCoacher need to be more flexible and give more autonomy to users by letting them choose interventions. Additionally, this work suggests the potential for matching users with others with similar sleep patterns in order to provide novel recommendations proven to work on similar sleepers. Finally, systems like SleepCoacher will need to help motivate users to maintain behavior change. Lastly, a future methodology improvement suggestion is to focus on effect size and evaluate its usefulness as a signifier of causality.

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.

ACKNOWLEDGMENTS
Anonymized for review.
REFERENCES


