SleepCoacher: Combining Computational and Clinician-Generated Sleep Recommendations

An honors thesis submitted for the Bachelor of the Arts Degree

in the Department of Computer Science at Brown University

by

Danaë Metaxa-Kakavouli

May 2015

SLEEPCOACHER:

COMBINING COMPUTATIONAL AND CLINICIAN-GENERATED SLEEP RECOMMENDATIONS

Danaë Metaxa-Kakavouli

Brown University 2015

Widespread use of smartphones to track various important aspects of personal health, including sleep, has the potential to empower individuals to better understand and improve their health independently and easily. Access to raw data and visualizations, however, leaves much to be desired for these users. Simple, personalized recommendations incorporating input of healthcare professionals can improve the efficacy of self-tracking and help guide non-expert self-trackers. We present SleepCoacher, a system integrating personalized data-driven recommendations with feedback from professional sleep clinicians. To evaluate SleepCoacher, we collected one month of smartphone-recorded sleep data from 24 undergraduate students. We collect one month of data for each participant, and in the last two weeks, we provide a personalized recommendations to each user every three days. We use the data collected to measure the effect of recommendations on various aspects of sleep including total hours of sleep, daily self-reported sleep quality, frequency of overnight awakenings, and sleep onset latency. We find that following recommendations significantly improves aspects of sleep hygiene in 94% of users, improved the specific aspects of sleep being targeted by these recommendations in 61% of users. This combination of clinical input and user data is the first of its kind, and in the future can be combined into a fully automated system that provides effective and actionable

individualized recommendations in real time based on observed improvements in other users.

From introducing me to the research process to supporting me in writing my first papers, Jeff Huang and Michael Bernstein have been exceptional mentors; working with them has improved my last last two years of college in immeasurably.

For their unconditional support and love, my infinite gratitude to Melia Snodgrass, Stella Kakavouli, Artemis Metaxa-Kakavouli, and Takis Metaxas.

ACKNOWLEDGEMENTS

This work could not have been completed without the contributions, support, advice, and guidance of Nediyana Daskalova, Alexandra Papoutsaki, Nicole Nugent, Julie Boergers, David Laidlaw, Tom W. Doeppner, and Emma P. Funk. Thanks also to CMU's Toss 'N' Turn team for the use of their data and insights into their analysis, and to the makers of Sleep as Android for the use of their app and permission to modify it for our study.

This Honors Thesis is based on work submitted to UIST 2015 with Nediyana Daskalova and Jeff Huang.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

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]. For more severe sleep disorders, people may be admitted to a sleep clinic where they are assessed by a physician while they sleep using polysomnography and actigraphy sensors. However, these methods are obtrusive, costly, and unnecessary for the vast majority of people who experience transitory sleep difficulties that are not secondary to medical or physiological conditions. As a result, a growing movement in personal informatics has led to people engaging in self-tracking to monitor their own sleep behavior. Research has shown that people are most interested in sleep monitoring technology that is unobtrusive and does not require the purchase of additional devices [14], making the smartphone an ideal form factor for sleep monitoring. Further research has shown that people are receptive to recommendations about behaviors preceding sleep to improve their sleep hygiene [8].

Currently, tens of millions of people have downloaded available sleep monitoring apps, with capabilities to sense noise via the microphone and movement via the accelerometer, to show the user their sleep patterns [2, 3]. Research prototypes have shown that the sensed data from smartphones can train a statistical model to predict sleep duration [13, 21], sleep quality [23], and events occurring during sleep [17]. In order to advance the current state-of-the-art in smartphone personal health monitoring, we go beyond the basic description and visualization of sleep patterns to the provision of tailored behavioral feedback in the form of data-driven and clinician-approved actionable recommendations for improving sleep.

The goal of our work is to measure the effect of actionable, clinician-approved, data-driven recommendations on features of smartphone-sensed movement and noise, as well as self-reported information, which can be used to measure various aspects of sleep quality. Specifically, we consider the impact of these recommendations on self-reported sleep quality, number and frequency of awakenings during sleep, sleep onset latency, and other sleep characteristics. In addition, we investigate the user responses to recommendations including the rate with which participants followed recommendations and their perceptions of recommendations received.

Our main contributions are twofold: (1) our sleep app provides users with *personalized recommendations for improving their sleep*, rather than simply showing them visualizations or data related to their sleep; and (2) this work is the first of its kind to *combine computational insights with feedback from healthcare professionals* to give concrete recommendations for users to improve their sleep.

2.1 Actigraphy and Polysomnography

Polysomnography (PSG) is the traditional method of sleep monitoring used to detect sleep disorders [28]. PSG is an overnight study performed in a hospital or sleep clinic. It can cost patients hundreds to thousands of dollars, and requires them to arrive 2 hours before bedtime for set up, which includes placing electrodes on the scalp, eyelids, and chin, have heart rate monitors attached to the chest, and wear clip on the index finger or ear to track blood oxygen levels [10, 5]. Although this is a noninvasive procedure, it is obtrusive, costly, and cannot be conducted frequently.

Other forms of sleep monitoring use specialized equipment which can improve detection of some sleep events, though these methods, too, are only available to the subset of the population who can afford the extra equipment. Actigraphy involves a user-worn electronic device, and has long been a common method for sleep tracking [19, 27, 26]. These existing medical methods are insufficient; they do not allow users to sleep in a naturalistic setting as is possible with self-tracking devices (see Figure 2.2) or track patterns in sleep over many consecutive nights, which misses valuable insights as sleep habits and general lifestyle change and fluctuate over time.

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. Systems like Lullaby [20] present users with data to help users better

Figure 2.1: Participants in traditional polysomnography studies, like the child pictured above, must wear multiple electrodes and other sensors which can be uncomfortable and time-consuming to set up.

understand their own sleep patterns on a nightly basis using such specialized equipment.

2.2 Personal Informatics and Sleep Tracking

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 the smartphone accelerometer, and as Natale et al. discover, smartphones can be just as accurate as medical devices for many sleep metrics, with the exception

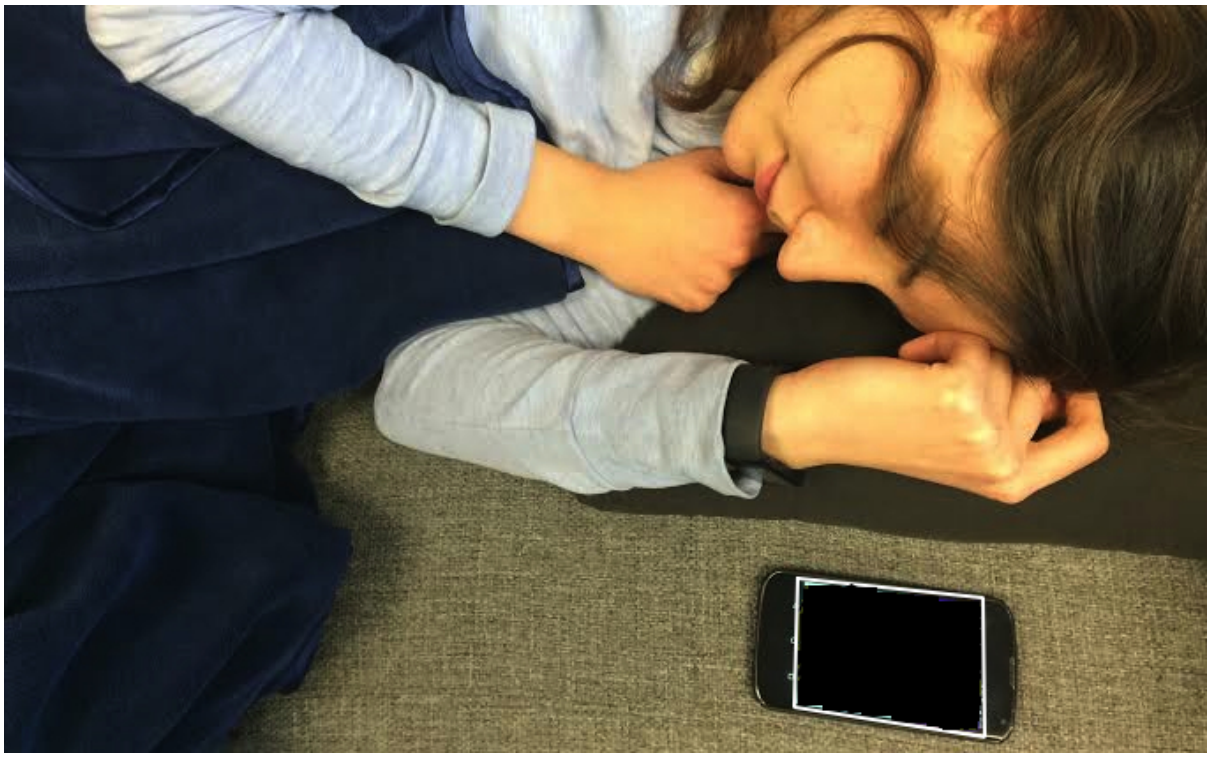

Figure 2.2: Users self-track their sleep by running a sleep app while sleeping, with the phone placed next to them on the bed.

of sleep onset latency [24]. Choe et al. review existing persuasive approaches to improve sleep and discover a lack of prescriptive technology that makes recommendations based on sleep monitoring data [14]. Bauer et al. conduct a study showing that people are indeed interested in recommendations to improve their sleep [8] and offer guidelines based on the time of day to improve sleep hygiene. Sleep hygiene describes behaviors that improve sleep quality or reduce harm to sleep [4].

While users of sleep tracking tools can find value in access to their sleep data, it is a challenge to extract useful sleep events from raw sensor data. Users therefore depend on these 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 [17]. Another app, Toss 'n' Turn, uses smartphone-collected data to train classifiers that detect sleep and predict sleep quality from such sensed data [23]. Other systems use a variety of smartphone sensors to detect the total number of hours slept by a user; Chen et al. find that accelerometer data is the best feature for accurate sleep duration estimation [13].

Health Mashups, by Bentley et al., aims to automatically detect correlations between different forms of personal informatics data and report it to users, though this system is not focused specifically on sleep data [9]. While some users of that system found correlations to be insightful, others found them spurious or obvious. Our work proposes to focus on sleep-related data and take the next step of turning correlations and other markers of sleep events in data into actionable, personalized recommendations.

Our work builds on the aforementioned existing approaches, combining the best of ubiquitous mobile devices' sleep monitoring capabilities with the sleep sensing algorithms developed in prior research, and further working with sleep clinicians to offer actionable recommendations to the users in a scalable way. A study by Lawson et al., supports the claim that a mobile phone application can be designed with the help of clinical sleep researchers [21].

2.3 Sleep and College Students

Though poor sleep is a common complaint among college students, and adversely affects quality of life and academic functioning, few effective interventions for sleep have been developed for college students [18]. Factors contributing to sleep problems among college students include noise and light levels in

communal living environments, substance and caffeine use, widespread use of electronics near bedtime, and delays in the secretion of melatonin during adolescence, which makes it more difficult for individuals in that age range to fall asleep until later in the evening [12, 11]. Lund et al. found that over 60% of students in their study were categorized as poor quality sleepers by the Pittsburgh Sleep Quality Index (PSQI), and over 70% get less than 8 hours of sleep per night [22] (the NIH recommends 7–9 hours of sleep for individuals in this age range [4]). This lack of sleep impairs cognitive development and can lead to a downward spiral of sleep-related issues [15]. Previous sleep research in college students supports the idea that self-monitoring and persuasive technology may be viable strategies to improve sleep behavior in college students [16].

Classifying a user's state as asleep or awake at any given moment based on sensor input is necessary in order to evaluate many aspects sleep quality such as number of overnight awakenings and sleep onset latency. This classification is a challenge faced by both medical technologies and personal informatics approaches. In the building of a system like SleepCoacher, it is necessary to consider existing techniques for sleep/wake state classification and identify a metric that is effective and consistent with best practices in the literature. In this chapter we explore existing medical and personal informatics approaches, and identify a sleep/wake classification algorithm for use by SleepCoacher.

3.1 Current Medical Methods

Due to the involved setup including dozens of electrodes and the capacity to track brain and muscle activity, polysomnography is the most sophisticated and accurate method of detecting sleep/wake states. In particular, the electrodes placed on the scalp are capable of performing electroencephalography—recording electrical activity inside the brain [28].

In contrast, actigraphy relies on accelerometer data and must infer sleep/wake state from that feature. Several viable algorithms have been developed for sleep/wake state detection based on actigraphy data, for example logistic regression classifiers with particular coefficients published in the literature [7]. Another common classifier in actigraphy is the use of a simple threshold with regard to the total amount of motion (integral of the movement curve) or simple level of movement [7].

3.2 Personal Informatics Approaches

Personal informatics methods often draw on methods from the actigraphy literature since both are reliant on accelerometer data, though using personal smartphone devices rather than formal actigraphs necessitates more flexible methods that are robust to the lack of calibration and consistency between different personal devices.

The Toss 'N' Turn system from Carnegie Mellon University uses machine learning classifiers to detect sleep/wake states, as well as to predict sleep quality [23]. Depending on the classifier and features used, Toss 'N' Turn finds between 70% and 95% accuracy in detecting sleep/wake state. The supervised learning methods used in that study depend on user-provided labels for the first three or more days in order to achieve these levels of accuracy.

3.3 Classification with Implicit Labels

In developing SleepCoacher, we found it valuable to experiment with different sleep/wake classification strategies based on the relevant actigraphy and personal informatics literature. Mimicking actigraphy strategies and drawing on Toss 'N' Turn's methods, we tested three different classifiers' ability to correctly classify sleep using only acceleration values from Toss 'N' Turn's publicly available data. This style of classifier which relies on only one feature is called a "1R Classifier". Since hand-labels were not available, we tested the efficacy of implicit labels: labeling any time point after the user-designated start of sleep and user-designated end of sleep as "asleep", and other time point as "awake".

We experimented with both generalized classifiers—those trained on the entire body of data—and individual classifiers—those trained only on a specific user's data. The three types of classifiers were a logistic regression (LOG), support vector machines (SVM), and a Naive Bayesian classifier (NB). Training was done with a ten-fold cross-validation.

We found that none of these classifiers reached accuracy near that of the Toss 'N' Turn system, which reported 89% accuracy on a 1R individual-acceleration-trained classifier, likely because of the lack of highly specific, hand-annotated labels. We found that the individually-trained logistic regression classifier reported the highest test accuracy, at 70.5%. Detailed numbers are reported in Table 3.1.

	Generalized Model		**Individual Model**	
Classifier	**Cross Val**	**Test**	**Cross Val**	**Test**
LOG	0.649	0.650	0.770	0.705
SVM	0.619	0.620	0.780	0.682
NB	0.597	0.636	0.643	0.645

Table 3.1: 1R logistic regression (LOG), support vector machines (SVM), and Naive Bayes (NB) classifiers were trained to predict sleep/wake states based on acceleration data across 27 participants (generalized) or within each specific user (individual). Here we report ten-fold cross-validation and testing accuracies.

Examining the parameters of the most effective classifier, the individually-trained logistic regression, it is clear that the classifier is essentially learning a threshold for distinguishing sleep from wake states. This is consistent with the actigraphy literature and with the raw results from Toss 'N' Turn's classifiers. As a result, we used a simple zero-crossing threshold for categorizing sleep/wake states with SleepCoacher. Users were declared awake when more than one movement occurred per two minute period of time, and asleep at the

start of a period of 20 minutes of inactivity. This was determined experimentally to meaningfully characterize sleep/wake states in user data.

4.1 Introducing SleepCoacher

We developed an integrated system, SleepCoacher, that combines automated data collection using Android smartphones with manual input from professional sleep clinicians to collect user data and in return provide personalized suggestions to users for improving sleep. The SleepCoacher process is a cycle consisting of data collection, data transmission to servers we can access, statistical analysis of data, generation of recommendations under clinician review, and delivery of recommendations to the user—all ostensibly before the next night's sleep (Figure 4.1).

4.2 Sensing and Data Processing

To collect data, we worked with developers of a popular Android sleep self-tracking app, Sleep as Android [2], which has over 10 million downloads (of which 1.5 million are active users). We received a modified version of their publicly available app that captures higher-resolution movement data, and we made our own 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.

The app collected bed and wake times, movement data through the accelerometer at 10-second intervals, noise levels from the microphone at approximately 5–10 minute intervals, coarse location coordinates, the user's self-

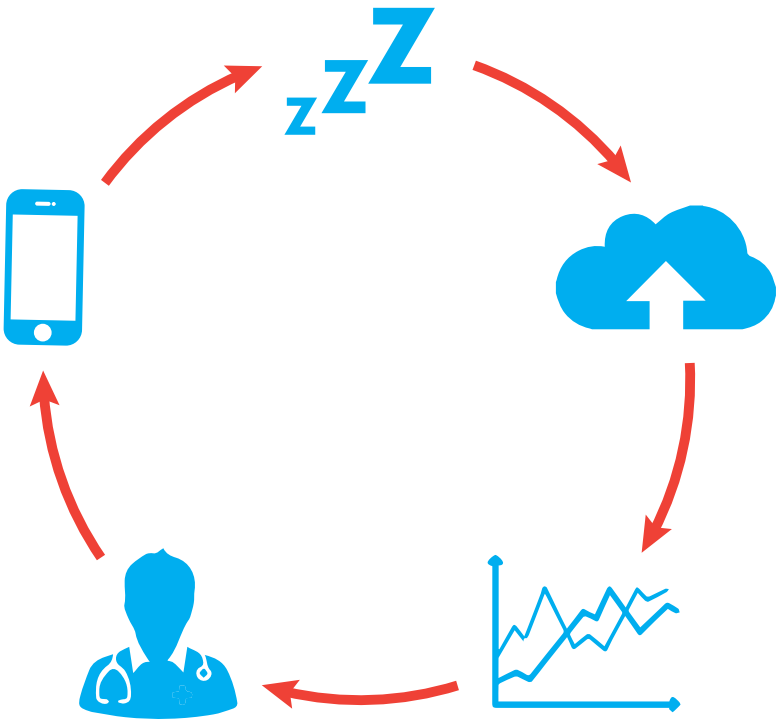

Figure 4.1: We propose a feedback loop: a user's sleep data is uploaded to the cloud and presented to sleep clinicians in the form of charts and correlation tables. Next, clinicians give recommendations, which are sent back to users, who adjust their sleep habits accordingly.

reported sleep rating per night, times of any alarms set, and tags users associated with the night's sleep. From these features, we computed the sleep onset latency (time to fall asleep) using a heuristic common in sleep actigraphy literature [6] and awakenings throughout the night also using a heuristic similar to ones described in sleep actigraphy literature [6, 25]. In essence, users were marked awake if they had more than one movement activity in the past 2 minutes, and they were deemed to have fallen asleep at the beginning of a period with no movement for 20 minutes.

When an application user indicates that they are awake, they stop tracking and enter a rating and a comment. Then, tracking stops and the app immediately uploads the data to servers that we can access, identified by an anonymous identifier. We use a script to download all the sleep data files from all users simultaneously. Then we compute Pearson correlations and statistical tests on the independent and dependent sleep factors (see Table 4.1). Finally, to analyze the data more accurately we analyze the data for specific trends that would suggest a user probably followed the given recommendation.

4.3 Design Considerations

We iterated on the design of the visualizations and data that would be produced by SleepCoacher and presented to sleep clinicians. Before we began the study, we had two sessions with the sleep clinicians to look at example visualizations, and discuss modifications. In each session, we presented visualizations of data released by a sleep study performed at Carnegie Mellon University [23] but in the style that we would use for the data generated by SleepCoacher. Our clinicians provided feedback to help us present data in ways that were more intuitive for them—for example displaying noise overlaid on movement, multiple nights for each users on a single page, and descriptive summary statistics for each user. In the second session, we addressed those issues, and presented new visuals and data to the clinicians. The feedback in this round led us to vertically align the nights so that clinicians could compare variation in sleep times, horizontally align all of the same days of the week in the same row, and compute measures for the number of awakenings (they would look particularly for if the user awoke more than three times per night), sleep onset latency, and

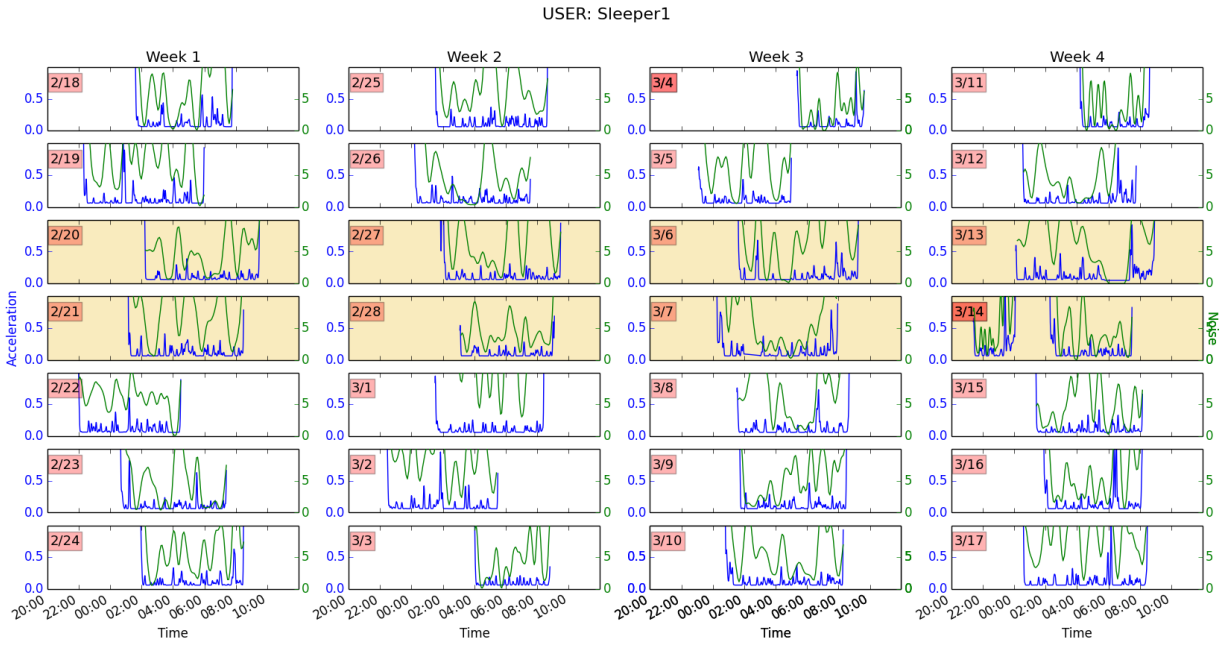

Figure 4.2: For ease of analysis by clinicians, acceleration and noise data for each user was combined into a single visualization encapsulating all nights. The date of each night is labeled, and weekends are highlighted against a pale yellow background.

highlight weekends versus weeknights. In our final version of SleepCoacher, we addressed all those issues to generate intuitive visuals and correlations for the sleep clinicians.

SleepCoacher then presents these correlations, graphs of the individual night data (Figure 4.2), and summary statistics for that user (Table 4.1) on a single page, for interpretation by a clinician or researcher. In further iterations when removing possible confounding factors for user behavior is not critical, these visuals can also be presented to users through the app interface.

Because every person has different responses and outcomes to different sleep recommendations, as well as different natural sleep patterns, this system allows users to receive highly personalized recommendations. One goal of Sleep-

Sleeper1	Noisiness	Rating	Awakenings	Onset Latency
Napped Hours	-0.34	-0.04	-0.28	-0.17
Is Weekend	0.01	0.23	0.06	0.06
Hours Slept	*0.05	-0.11	*0.74	0.4
Noisiness		0.12	*0.43	0.05
Bedtime Consistency	-0.26	-0.17	-0.37	-0.39
#work	0.08	0.04	0.31	0.33
#food	0.3	-0.34	-0.06	0.07

Table 4.1: Correlations between variables were presented to sleep clinicians in the form of a table, with independent variables in the first column and dependent variables in the first row. Significant correlations between independent and dependent variables were denoted with an asterisk.

Coacher is to identify which recommendations work and which do not, both across the entire population of users and specifically with regard to subgroups of users whose sleep patterns are similar to each other. This will allow users to conduct small-scale experiments on their own sleep, adjusting different independent sleep factors and allowing SleepCoacher to learn and adjust its recommendations based on the results, in a rapid feedback cycle.

USER STUDY

After developing SleepCoacher, we conducted a month-long user study to measure the effect of personal recommendations on user sleep and self-tracking behaviors. In the first half of the study we simply collect data from users, and in the second we combine data-driven and clinician-generated sleep recommendations to help users improve sleep.

5.1 Study Procedure

We recruited undergraduate students over age 18 who use an Android smartphone (version 2.2+) as their primary phone. We restricted participants to those without diagnosed sleep problems or other medical barriers that would put them at risk or prevent them from following our recommendations. Our participants, 11 women and 13 men, were all undergraduate students between 18 and 22 years of age.

Undergraduate students 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.

Participants were instructed to use the sleep app nightly for 28 continuous nights, placing the phone flat on their bed near shoulder-level (but not under the pillow), as shown inFigure 2.2. To begin tracking, participants pressed a button to start tracking upon getting into bed and stop it upon waking up. Further, upon waking up participants recorded various qualitative features relevant to their sleep the previous night. Each participant had the option to give their sleep a qualitative sleep rating (1–5 stars), a comment, and a personalized tag

Figure 5.1: Each morning, participants were give the opportunity to rate and comment on the previous night's sleep.

(e.g. "#exercise", "#coffee", "#stress"), as seen in Figure 5.1.

Following the study, participants were given an exit survey asking whether they followed each recommendation, found each recommendation helpful, and had any other comments on the app or recommendations. Participants were also asked about their preferred timing for receiving recommendations, and whether and 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 for their participation; those who did not meet these standards were paid $25. Of our 24 participants, 22 participants recorded their sleep for at least 80% of the duration of the study. Data from the remaining two participants was excluded from analysis.

5.2 Sleep Clinician Review

In combination with the existing sleep app, we worked with two sleep clinicians from [Omitted for Anonymity] Hospital to analyze each users' sleep data and generate individual-specific as well as more general recommendations for improving sleep. Over a series of meetings we developed intuitive visualizations of each user's sleep patterns as well as color-coded tables highlighting significant correlations between dependent and independent sleep attributes in the data. Our visuals (as shown in Table 4.1 and Figure 4.2) were organized to mimic the visual structure of professional polysomnography and actigraphy data which clinicians are used to studying (an example of a traditional polysomnography chart is shown in Figure 5.2).

After these preliminary meetings, halfway through the study we presented the clinicians with figures and charts for each user, as well as possible recommendations for improving sleep derived from the data. The clinicians screened these recommendations to ensure they were safe, relevant, and effectively worded, and suggested additional recommendations based on each user's charts (see Figure 4.2). This process resulted in a total of 27 unique recommendations, 19 (70%) of which were generated by the sleep clinicians after viewing user data, and 8 (30%) of which were clinician-approved.

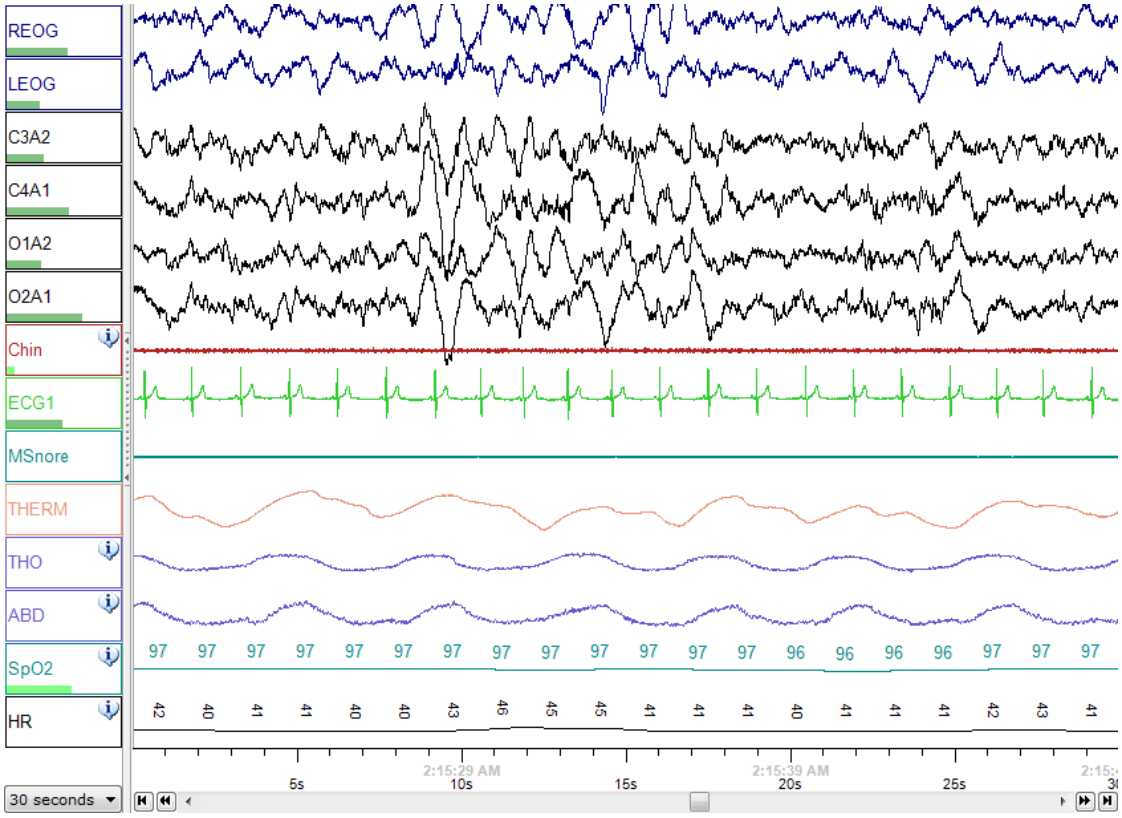

Figure 5.2: Sleep clinicians have experience interpreting traditional polysomnography charts, such as the one pictured above.

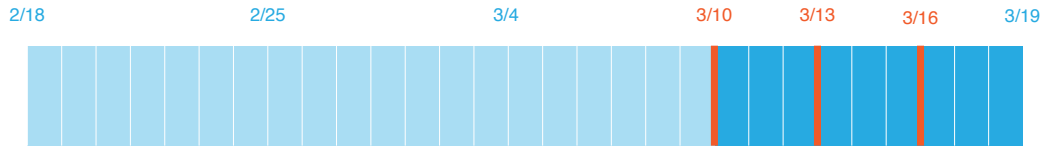

Figure 5.3: Data was collected for one month (2/18–3/18); recommendations were given to participants once every 3 days for the last 10 days of the study.

After eliciting these recommendations, our users received a total of three personalized recommendations, roughly one per three days, for the remaining ten days of the study, as shown in Figure 5.3. Recommendations were always distributed to participants at 10pm by a text message send to their mobile phones.

RESULTS

6.1 Aggregate Results

In the latter part of the study, we sent each user three recommendations, one per three days, for a total of 72 recommendations. When surveyed, users reported following 39 of them (54%). By examining the data, we can confirm that with certainty that 46% of these (18 recommendations) were followed. In 61% of the users that followed recommendations we saw an improvement over the course of three sleeps in the specific dependent outcomes targeted with the preceding recommendation. More generally, 94% of the users that followed the recommendations saw some kind of improvement in their sleep outcomes.

Before analyzing data to compare those who followed recommendations with those who did not, we filter out those users who reported following recommendations but who we can confirm did not change their behavior. For half of all followed recommendations, for example when a user was asked to change a self-reported sleep attribute (e.g. "#stress", "#deeptalks") or when behavior changes would not appear in the data (e.g. wearing earplugs), it was not possible to ascertain whether or not users had made changes, but this was an effective strategy for the other half of followed recommendations (e.g. having a more consistent bedtime, sleeping longer).

We separated the recommendations in two groups: those sent to only one user ("individual"), and those sent to more than one user ("general"). Overall, we sent 13 individual recommendations, 4 of which we could confirm were followed from the data (31%), and 59 general ones, 14 of which we could confirm

were followed (31%).

Across all participants, the average rating was 3.5, average awakenings was 1.6/hour, average onset latency was 24.9 minutes, and average hours of sleep was 7.4. Overall, users who followed two or more recommendations rated their sleep slightly higher on average (3.8), but otherwise showed minimal difference.

6.2 Recommendation-Specific Findings

The first round of recommendations, sent roughly halfway through the study, addressed a general problem across all 24 participants. Since awakenings are always accompanied by noise, which may either be a cause or result of the awakening, users were instructed, "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, etc.)". Five users followed this recommendation, and we saw the expected decrease in awakenings in two of these users and other sleep improvements such as increase in hours slept and lower latency among all five users. Users who followed this recommendation were significantly more likely to see an increase in average hours slept when comparing the three days following the recommendation to that of all previous nights ($p < 0.05$).

The second recommendations were more personalized, and included suggestions to maintain a more consistent bedtime, sleep longer, and stabilize weekend and weekday sleep trends. 7 of 24 users followed this suggestion, and in 5 of these 7 cases we saw the predicted change in dependent sleep outcomes (which varied according to the recommendation). In the other two cases, we saw improvements different from the ones predicted. Users who followed

Round	Example	Followed	Improved	Benefited
1	"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, etc.)"	5	40%	100%
2	"When your bedtime is variable, you have more trouble falling asleep. Try to go to bed around the same time every night."	7	71.4%	100%
3	"When you reported #alcohol, you were less satisfied with your sleep. Try to avoid #alcohol within 2–3 hours of bedtime."	6	66.7%	83.3%

Table 6.1: The 3 rounds of recommendations sent to participants during the study. "Followed" represents the number of participants who followed the recommendation, "Improved" represents the percentage of participants with improvements in the target variable, and "Benefited" represents the percentage of participants with improvements in any area.

this recommendation were significantly more likely to see a decrease in sleep onset latency ($p < 0.01$).

The third recommendations were the most personalized, and focused on user-generated tags, when available. These recommendations included suggestions to avoid "#alcohol", increase "#talk", and decrease "#deeptalks" to increase various dependent outcomes. We saw improvements in the targeted outcomes for 2/3 of users, and other improvements in one user. Users who followed the third recommendation were significantly more likely to see a decrease in the number of awakenings per hour during sleep ($p < 0.05$). Across

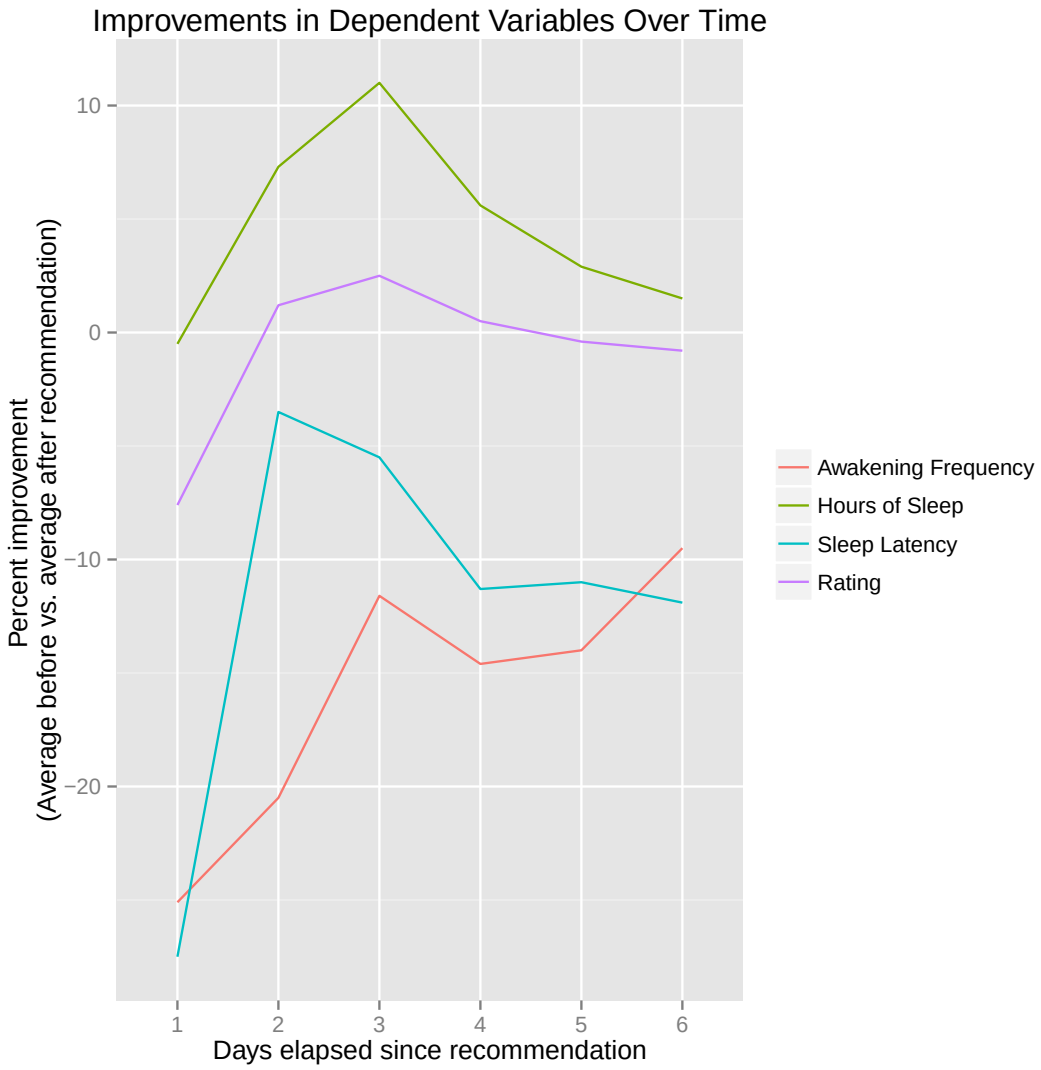

Figure 6.1: Comparing the average value of our independent variables before the recommendation with the average value in the subsequent x days, we see average improvements peaking at around 3 days.

all recommendations, only one user who followed the recommendation did not see any improvement over the subsequent 3 days.

Figure 6.1 compares the average values of four dependent outcomes before the first recommendation to the average values x days following that recom-

mendation ($1 \leq x \leq 6$). Across all four outcomes (awakening frequency, hours of sleep, sleep onset latency, and rating), the difference between the difference in averages before and after the recommendation shows steady improvement until the second or third day when a new recommendation was sent, after which improvement drops off but remains positive compared to pre-recommendation values.

In the exit survey users indicated what time of the day they would prefer to receive the recommendations. 8/24 said they would like to receive them when they wake up, 6/10 before they go to bed, and the majority, 10/24, preferred getting them throughout the day.

6.3 Participant Feedback

Following the study, we conducted an exit survey to understand which recommendations users followed, why they did or did not follow them, and how they felt about the overall experience. A key insight from these surveys was the range of reasons participants gave for neglecting to follow recommendations. Their reasons fell into two main groups: participants were often not intrinsically motivated, and/or found it difficult to follow concrete suggestions given the college lifestyle. When users found the effort- or time-cost of following a recommendation to be low, many were happy to follow recommendations; one user reported, "I followed it because it wasn't that hard to do; I just downloaded a white noise app and tried it out." In other cases, however, users were deterred by the overhead of having to adjust to a new sleep behavior, or found the cost too high: "I tried to use the white noise generator built into the app but found it

distracted me from getting to sleep. I have never been a fan of earplugs because they are uncomfortable." Many users reported following recommendations "as much as possible", but being unable to do so fully because of constraints imposed by the college lifestyle: "I just think [keeping a consistent bedtime] isn't a very feasible thing for a college student."

The most common suggestion for improvement we received was the desire for more personalized and frequent feedback. The lack of concrete metrics drawn from participants' data made some participants less convinced that recommendations were indeed based in person-specific patterns. Despite this, personalization was effective, particularly when users' own added tags were used in data-driven recommendations. These allowed users to focus on factors which they felt were important and relevant in their own lives. For example, when given a recommendation to decrease "#stress", a participant noted, "I definitely felt the affects of stress in areas of my life other than sleep, so the tip did not come as a surprise. The tip encouraged me to act on that stress by trying to manage it better. The most helpful result probably was connecting socially which helped alleviate some of the loneliness I felt during this period." This kind of introspection made many participants feel more motivated to use the app and satisfied at the outcomes.

Overall, users felt their sleep habits had been positively influenced by their use of the app. 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 have on sleep. Users noted insights including, "It has made me aware of how much time I spend working on my laptop before going to bed," "[The app has] definitely made me more conscious of the times/length

of my sleep," and "It's made me a little more reflective about my sleep habits, because while they seem relatively healthy, maybe I could actually be doing a lot better."

DISCUSSION

7.1 Turning Correlation into Causation

Throughout our study, we saw correlations between sleep attributes, for example between onset latency and bedtime consistency. We based our recommendations on those correlations with the assumption that there might also exist a causal relationship between such attributes. In 61% of users who followed recommendations targeted at improving specific dependent outcomes such as sleep onset latency, rating, and frequency of awakenings we found that following the recommendation indeed led to improvements in the predicted dependent outcome. We are limited by our relatively small number of users and short time-frame, given that many of these improvements emerge gradually and with consistent effort, but this result supports the hypothesis that there exist causal relationships between many of the independent and dependent attributes studied, and that improving sleep hygiene can cause better sleep outcomes.

7.2 Self-Experiments

In order to further strengthen the evidence for the causal relationships we hypothesize, further development and deployment of the app will allow users to conduct small-scale personal experiments, altering an attribute related to their sleep and tracking the results of that change over time. This provides an even deeper level of personalization. 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 long term through a rapid feedback cycle.

7.3 Consistency and Improvement

The improvements in the dependent outcomes over time shown in Figure 6.1 provide us with important feedback—users who follow the recommendations see the most improvement on days 2 and 3. It is important to keep in mind that users received the next recommendation on day 3, so we might have some confounding effects; they might have thought that they no longer need to follow the previous recommendation, for example.

Notably, the difference in averages between days prior to the recommendation and days after the recommendation is sometimes negative, suggesting that participants were experiencing poorer sleep as the study continued; this is consistent with the timing of the study, which ended just before spring break, after midterms. However, participants who followed recommendations saw such negative effects become less severe over time.

Participants' difficulty with maintaining consistent changes over time also speaks to a strength of self-tracking relative to existing medical methods. Sleep habits and aspects of lifestyle impacting sleep necessarily change and fluctuate with time. Undergraduates, as we saw in our study, are particularly susceptible to inconsistencies in their sleep habits, as much of their schedule is externally determined by the academic calendar.

7.4 Risks of Automation

As in any automated system, attempts to influence user behavior algorithmically must be done with care. A system that attempts to force changes in user behavior may quickly be perceived as annoying and as a result fall into disuse. Instead of such 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.

Even with the noble goal of empowering users through information, providing feedback to anonymous users through an automated system can lead to unintended consequences. In our first recommendation, we suggested that users buy earplugs or use a white noise machine to decrease wakenings during the night. One user gently informed us, "I am hearing impaired and take out my hearing aids when I sleep," so this recommendation was inappropriate. Further, 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 suggestions that are ineffective or simply do not apply to particular users. In order to better support a wide range of people, these systems must be developed conscientiously, with the flexibility to accommodate such differences.

7.5 Limitations

Improvements in sleep can take time to come into effect, and require consistency—something a population of undergraduates without intrinsic motivation to use a sleep app might struggle with. Similarly, the fluctuations of student life due to the academic calendar impacted our participants' ability to make conscious improvements to their sleep. We heard as much from participants, who noted, "As a college student, I think it's particularly difficult to follow a recommendation such as [sleeping consistently between weekdays and weekends] because we have a set class schedule, set outside activities, etc.", and, "It's hard to get more sleep. :(".

This study, while anticipating causal relationships between different aspects of sleep, was not conducted with randomized experiments, which would be needed to prove causal relationships between the independent and dependent aspects of sleep. Doing so, however, would require a much larger population size, a worthy goal for a different study. It is also intrinsic to the process of allowing users to determine independently whether or not to follow recommendations. From the perspective of our users, the approximation of causation derived by predicting what sleep outcomes will be impacted by a recommendation and then confirming that impact is sufficiently motivational when recommendations are personalized and specific.

7.6 Future Work

Our results suggest the promise of combining expert insights—medical professionals, in the case of health-related tracking—with data-driven recommendations to build more interactive, powerful personal informatics systems.

The widespread prevalence of sleep problems as well as increasing interest in self-tracking with smartphones suggests that a natural and promising step forward for SleepCoacher is growth to massive scale. In order to meet the needs of millions of users, SleepCoacher will need to be adjusted in several ways. Primarily, clinician recommendations will need to be retrieved automatically. This will require a curated database of recommendations, each marked with the data-related trend that led to its generation by sleep clinicians. Upon registering with SleepCoacher and recording their sleep for a short calibration period, features extracted from user data would match users' particular sleep patterns with relevant recommendations and send those recommendations automatically.

Additionally, once users are confirmed to have followed a recommendation and the effects of their behavior changes are confirmed, our system can provide even better recommendations to new users: recommendations which have been shown to improve sleep in other users with similar patterns. These suggestions will be more likely to have positive impacts, and also give the user more information about the specific effects of following a recommendation and the likelihood it will lead to those improvements.

CONCLUSION

Existing medical methods for studying sleep—while valuable for those with diagnosed, severe sleep troubles—are insufficient for the majority of people who have mild to moderate problems. For these individuals, whose sleep patterns and lifestyles we saw change and fluctuate over time, personal sleep trackers offer the opportunity to improve sleep hygiene by identifying identifying personal areas for improvement through regular, low-cost recording of their sleep in a naturalistic setting.

This work sets out to show the value and viability of a mixed computational and clinical strategy of providing personalized, actionable recommendations to self-trackers. By conducting a one-month user study with 24 participants, we have shown that users who follow recommendations are significantly more likely to see improvements in aspects of sleep hygiene including hours of sleep, sleep onset latency, and frequency of awakenings during sleep. Further, we see improvements in the specific sleep outcomes our recommendations were were targeting in 61% of users who followed recommendations, and saw more general improvements in 94% of those users. SleepCoacher is the first step towards a personalized sleep coach for every user, with the capabilities of an automated data-driven learning algorithm and a professional sleep clinician with empathy and holistic understanding of human needs.

BIBLIOGRAPHY

[1] Brain basics: Understanding sleep. NIH Publication No.06-3440-c.

[2] Sleep as android. https://sites.google.com/site/sleepasandroid/.

[3] Sleepcycle. http://www.sleepcycle.com/.

[4] Healthy sleep tips, 2014. http://sleepfoundation.org/sleep-tools-tips/healthy-sleep-tips.

[5] Polysomnography (sleep study), 2014. Retrieved: 2015-04-13.

[6] S Ancoli-Israel, R Cole, C Alessi, M Chambers, W Moorcroft, and C Pollak. The role of actigraphy in the study of sleep and circadian rhythms. american academy of sleep medicine review paper. *Sleep*, 26(3):342–392, 2003.

[7] S Ancoli-Israel, R Cole, C Alessi, M Chambers, Moorcroft W, and Pollak C. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 26(3):342392, 2003.

[8] J.S. Bauer, S. Consolvo, B. Greenstein, J. Schooler, E. Wu, N.F. Watson, and J. Kientz. Shuteye: Encouraging awareness of healthy sleep recommendations with a mobile, peripheral display. In *Proc SIGCHI*, CHI '12, pages 1401–1410, New York, NY, USA, 2012. ACM. http://doi.acm.org/10.1145/2207676.2208600.

[9] F. Bentley, K. Tollmar, P. Stephenson, L. Levy, B. Jones, S. Robertson, E. Price, R. Catrambone, and J. Wilson. Health mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *TOCHI '13'*, 20(5):30, 2013.

[10] A. Blaivas. Polysomnography, 2014. http://www.nlm.nih.gov/medlineplus.

[11] M. Carskadon, C. Acebo, G. Richardson, B. Tate, and R. Seifer. An approach to studying circadian rhythms of adolescent humans. *J Biol Rhythms*, 12(3):278–89, 1997.

[12] M. Carskadon, S. Labyak, C. Acebo, and R. Seifer. Intrinsic circadian period of adolescent humans measured in conditions of forced desynchrony. *Neurosci Lett.*, 260(3):129–32, 1999.

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

[14] E.K. Choe, E. Consolvo, N.F. Watson, and J.A. Kientz. Opportunities for computing technologies to support healthy sleep behaviors. In *Proc. CHI '11*, pages 3053–3062. ACM, 2011.

[15] M. Curcio G., Ferrara and L. De Gennaro. Sleep loss, learning capacity and academic performance. *Sleep Med Rev*, 10(5):323–37, 2006. http://www.ncbi.nlm.nih.gov/pubmed/16564189.

[16] N. Daskalova, N. Ford, A. Hu, Moorehead K., B. Wagnon, and J. Davis. Informing design of suggestion and self-monitoring tools through participatory experience prototypes. In *Proc. Persuasive Tech.*, 2014.

[17] T. Hao, G. Xing, and G. Zhou. isleep: unobtrusive sleep quality monitoring using smartphones. In *Proc SenSys '13'*, pages 4:1–4:14, 2013.

[18] Shelley D Hershner and Ronald D Chervin. Causes and consequences of sleepiness among college students. *Nature and science of sleep*, 6:73, 2014.

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

[20] M. Kay, E.K. Choe, J. Shepherd, B. Greenstein, N. Watson, S. Consolvo, and J.A. Kientz. Lullaby: a capture & access system for understanding the sleep environment. In *Proc UbiComp*, 2012.

[21] S. Lawson, S. Jamison-Powell, A. Garbett, C. Linehan, E. Kucharczyk, S. Verbaan, D.A. Rowland, , and K. Morgan. Validating a mobile phone application for the everyday, unobtrusive, objective measurement of sleep. In *Proc. SIGCHI*, CHI '13, pages 2497–2506, New York, NY, USA, 2013. ACM. http://doi.acm.org/10.1145/2470654.2481345.

[22] H. Lund, B. Reider, A. Whiting, and R. Prichard. Sleep patterns and predictors of disturbed sleep in a large population of college students. *Adolescent Health*, 46(2):124–132, 2010. http://www.jahonline.org/article/S1054-139X(09)00238-9/abstract.

[23] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong. Toss 'n' turn: Smartphone as sleep and sleep quality detector. In *Proc. SIGCHI*, CHI '14, pages 477–486, New York, NY, USA, 2014. ACM. http://doi.acm.org/10.1145/2556288.2557220.

[24] Vincenzo Natale, Maciek Drejak, Alex Erbacci, Lorenzo Tonetti, Marco Fabbri, and Monica Martoni. Monitoring sleep with a smartphone accelerometer. *Sleep and Biological Rhythms*, 10(4):287–292, 2012. http://dx.doi.org/10.1111/j.1479-8425.2012.00575.x.

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

[26] C. Pollak, W. Tryon, H. Nagaraja, and R. Dzwonczyk. How accurately does wrist actigraphy identify the states of sleep and wakefulness? In *Sleep*, volume 24, pages 957–965, 2001.

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

[28] H. Shin, B. Choi, D. Kim, and J. Cho. Robust sleep quality quantification method for a personal handheld device. *Telemedicine and e-Health*, 20(6):522–30, 2014.