Investigating the Effectiveness of Cohort-Based Sleep Recommendations

NEDIYANA DASKALOVA, Brown University, USA
BONGSHIN LEE, Microsoft, USA
JEFF HUANG, Brown University, USA
CHESTER NI, Microsoft, USA
JESSICA LUNDIN, Microsoft, USA

Existing sleep-tracking apps and devices provide simple descriptive statistics or generic recommendations for everyone. In this work, we aim to leverage cohort-based sleep data to provide recommendations to improve an individual's sleep. We report a 4-week study (N = 39) conducted to compare three alternatives: 1) no recommendation, 2) general recommendation, and 3) cohort-based recommendation, using six sleep quality metrics. For the cohort-based recommendation, recommendations were generated based on "similar users" using about 40 million sleep events from Microsoft Band users. Our results indicate that cohort-based systems for health recommendations can prompt a desire for behavior change inspired by social comparison and increased awareness about sleep habits. However, in order to be effective, such systems need to establish their credibility and to be able to generate cohorts based on features that are important to users. Finally, we provide further suggestions and design implications for future cohort-based recommendation systems for healthy sleep.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; *Empirical studies in HCI*; Ubiquitous and mobile computing systems and tools;

Additional Key Words and Phrases: Sleep, self-tracking, behavior change, sleep recommendations, sleep hygiene

ACM Reference Format:
Nediyana Daskalova, Bongshin Lee, Jeff Huang, Chester Ni, and Jessica Lundin. 2018. Investigating the Effectiveness of Cohort-Based Sleep Recommendations. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 101 (September 2018), 19 pages. https://doi.org/10.1145/3264911

1 INTRODUCTION

According to a study by the National Sleep Foundation, 45% of Americans say that poor or insufficient sleep affected their daily activities at least once in the past seven days [1]. However, people might not know exactly how to improve their sleep habits, and prior studies have shown that they are receptive to sleep-related suggestions [27]. The most common suggestion for improving sleep is to follow the generic sleep hygiene guidelines, such as "sleep 7–9 hours" or "avoid caffeine close to bedtime" [3, 4]. While these guidelines may be helpful for the overall population, they fail to acknowledge individual differences and thus might be inappropriate or even detrimental to an individual's sleep. For example, chronotypes [51], the characterization of a person's tendency to wake up early

Authors' addresses: Nediyana Daskalova, Brown University, 115 Waterman Street, Providence, RI, 02912, USA, nediyana@cs.brown.edu; Bongshin Lee, Microsoft, Redmond, USA, bongshin@microsoft.com; Jeff Huang, Brown University, 115 Waterman Street, Providence, RI, 02912, USA, me@jeffhuang.com; Chester Ni, Microsoft, Redmond, USA, chesni@microsoft.com; Jessica Lundin, Microsoft, Redmond, USA, jelundin@microsoft.com.

2474-9567/2018/9-ART101 $15.00
https://doi.org/10.1145/3264911

or feel refreshed when they go to sleep late and wake up late, are not typically incorporated in recommendations. Furthermore, even when people are aware of these recommendations, they often do not adhere to them [44].

There are two major approaches for improving sleep: (1) an obtrusive and expensive overnight study, used for diagnosing major sleep disorders, and (2) a broad selection of sleep tracking mobile apps and devices that provide mainly summary statistics. While the former provides detailed personalized recommendations for improvement, the latter, albeit being more affordable and widely accessible, pay little attention to individual sleep patterns. On the other hand, systems that use only one individual's data are prone to over-fitting and omit valuable context that can be found in other people's sleep patterns.

Bridging these approaches, we build on current sleep-tracking methods to leverage data both from each individual and from similar users to provide actionable cohort-based recommendations. Using a collaborative-filtering technique, we identify a cohort for each participant, using a large real-world dataset. We conducted a 4-week study (N = 39), collecting six sleep quality metrics and exit questionnaires from three groups of participants who either received: (1) no recommendation, (2) a general recommendation, or (3) a cohort-based recommendation. We aimed to answer the following research questions: (1) How can we build off of existing collaborative-filtering frameworks to generate cohort-based recommendations for sleep, given a large dataset of other users?, (2) What are the limitations of recommendations based on collaborative-filtering that surface when people are asked to follow them in their everyday life?, and (3) While grounding our work in existing theory, what design hypotheses can we provide for future systems to overcome these limitations?

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

2 RELATED WORK

2.1 Automatic Sleep Monitoring

To diagnose severe sleep disorders and receive detailed suggestions for improvement, people need to undergo an overnight polysomnography study in a sleep clinic [2]. However, these studies cannot detect occasional sleep problems, and sleep in clinics may not be representative of *in situ*. The alternative sleep-tracking apps and devices, such as Microsoft Band and FitBit, employ actigraphy techniques [34, 47, 54]: they infer sleep and wake states based on accelerometer-detected movement. While these trackers have shortcomings such as limited battery life, non-standardized accuracy, and discomfort [39], research shows that one in 10 Americans owns one [40]. Our study participants used one of these devices, the Microsoft Band 2, to track their sleep over a 4-week period.

Previous sleep studies using smartphones and other devices outside of the regular polysomnography labs have focused mainly on developing systems for detecting sleep and predicting sleep quality. For example, Toss 'n' Turn uses smartphone-collected data to train classifiers [45], and iSleep uses smartphone microphone data to detect sleep events [31]. 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 to estimate sleep duration accurately [19]. While a smartphone app usually does not require any additional hardware, we chose to use the Microsoft Band for this study as it gave users the opportunity to also track daily exercise.

2.2 Non-Clinical Sleep Studies to Support Healthy Sleep Behaviors

Some non-clinical sleep studies focus on evaluating their systems based on the behavior change of users, but Klasnja et al. [38] argue a better focus for early stage HCI technologies would be on the users' experiences with the system. Thus, in our work, we focused on the analysis of the end-of-study qualitative questionnaires that participants completed. This method is in line with previous studies, which use qualitative feedback from participants to evaluate their systems. Lullaby [36], for example, uses temperature, light, audio and motion sensors to collect data and lets users look through graphs for trends in their sleep quality. Somnometer [56] even allows users to share sleep information with others. Previous research also points out that it is important to design technology that helps people become more mindful of their sleep, while not imposing impossible sleep goals [21]. To evaluate SleepTight [22], for example, the researchers conducted a four-week study, to determine whether the system with the widgets enabled a higher sleep diary compliance rate than the one without. Our work employed a similar framework by evaluating the effect on sleep of two types of recommendations and a control condition with no recommendations.

2.3 Sleep Recommendations

Prior work shows that self-tracking and suggestions combined can improve sleep [26]. However, it is unclear how the few systems that provide recommendations even generate them. Fitbit and Jawbone's insights are limited to general trends and comparisons of the user's sleep to that of others of similar age and gender. However, none of these trackers provide actionable recommendations that have been shown effective for people similar to the user.

Notably, two systems have been developed to provide actionable sleep recommendations beyond simple summary insights. ShutEye [12] focuses on displaying sleep hygiene guidelines on a user's mobile phone home screen. However, these guidelines are based on the general population and might neglect individual differences like those between different age groups [4]. SleepCoacher, on the other hand, provides personalized recommendations based on the framework of self-experiments [27]. SleepCoacher users tracked sleep with a smartphone application and received recommendations based on correlations within their own data. This approach did not use information from other users that might provide helpful insights or motivation for improving someone's sleep.

2.4 User-Focused Recommender Systems

To provide cohort-based recommendations, this work employs a recommender system developed to identify groups of similar users. The approach of making predictions or recommendations about a user based on data from other users is known as collaborative filtering [50]. Collaborative filtering is most often based on 'neighborhood models,' in which the unknown ratings from a user are estimated based on ratings from similar users [50]. Neighborhood models are popular because of their simplicity and the intuitive reasoning behind their recommendations [50]. Another benefit, in relation to sleep, is that given the right parameters, these models can be used to generate a recommendation for new users [50], even before any actual sleep tracking data has been recorded. Our cohort-based approach is a type of a neighborhood model, and thus can be further fine-tuned and improved.

Pu and Chen's user-centric framework for evaluating recommender systems emphasizes the need for minimizing the interaction effort for the user while producing useful and trustworthy recommendations in a transparent way [48]. To address this need, some recommender systems focus on getting users' feedback in order to better understand their preferences and provide more accurate recommendations. In particular, a critique-based recommendation system first suggests a few options based on users' current preferences. Then, the user critiques those suggestions, and then the system generate new ones based on the critiques [18]. Such systems help users build a preference profile, but they rely on the content and are task-specific [33]. Alternatively, studies have shown that users are more satisfied when they are given the chance to directly manipulate the attributes as in [14, 33].

In our study, we chose to focus on collaborative filtering techniques to elicit actionable sleep recommendations based on the data from other users. However, future work could explore how critiquing-based systems can provide recommendations that incorporate users' preferences. We discuss this notion further in Section 7.

2.5 Health Recommender Systems

Health recommender systems have been focused on personal health record systems (PHRS), which centralize each person's electronic health data and allow health professionals to access it [58]. PHRS contain too much expert-oriented data, which leads to information overload for the regular user [66]. Thus, health recommender systems have been developed to provide laymen-friendly information to users to better understand their own data [64, 66, 67]. Even outside of PHRS, recommender systems have been used mainly for information filtering [55]. However, such systems have not been deployed in the domain of sleep.

3 METHOD

3.1 Dataset

Our dataset contained the real-world sleep records collected by the Microsoft Band (MS Band), a wrist-worn fitness tracker, first released in 2014, followed by an updated hardware version in 2015. While there are a total of 40 million sleep records, this study used data gathered between May 2016 and May 2017 (about 1 million users). Overall, MS Band users were from diverse backgrounds, age ranges, gender, occupations, and nationalities. Recent work by Althoff et al. shows that MS Band's sleep data is representative for the general population as the measurements match published sleep estimates [7]. In this study, we used the data only from users who reported being between 18 and 65 years old, were between 50 and 80 inches tall, and weighed under 250 lbs. These criteria are commonly used in sleep and exercise literature [62, 65]. Following similar studies, we also exclude any sleep records with a duration of less than four hours or more than 12 hours [7, 63].

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

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

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

According to sleep literature, intense exercise for 30 minutes three times a week improves sleep [11]. Exercise data including biking, running, and working out is also collected by the MS Band once users tap on the corresponding activity tile. The Band tracks heart rate using an optical sensor. We compute overall intensity of the activity based on the rate of calories burned (using a proprietary algorithm).

3.2 Study Design and Participants

We designed a between-subjects exploratory study to evaluate the effectiveness of cohort-based recommendations and to learn how to make better cohort-based health recommendations in the future. We randomly assigned the participants to one of three conditions. In condition 1, the control "no-recommendations," participants tracked their sleep every night for a month and received no recommendations. In condition 2, participants received a general recommendation halfway through the study. In condition 3, participants also received a recommendation halfway through the study, but it was personalized based on the cohort of similar users. In both conditions with recommendations, participants were sent daily reminders for the last half of the study asking them to follow the same recommendation every day. We describe the algorithm for generating recommendations in Section 4.

We recruited participants internally in a large software and services technology company from a pool of beta testers. Due to the remote nature of the study, participants were distributed across the US. Only participants who did not have a self-reported sleep disorder, who could track their sleep for at least 14 nights of the study, who had access to a MS Band (version 1 or 2), and who were not traveling across timezones were allowed to participate in the study. Traveling across timezones affects sleep patterns and causes jetlag, which can interfere with sleep quality the results of the study [53].

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

3.3 Study Procedure

The pre-study questionnaire had questions from the Epworth Sleepiness Scale (ESS) and the Pittsburgh Sleep Quality Index (PSQI), both of which are commonly used in research studies to evaluate sleep quality [17, 35]. The survey also included questions related to sleep quality and lifestyle.

Participants across all conditions were sent a daily email at the same time asking them to rate their sleep quality between 1 and 5, with 5 being their best sleep ever and 1 being their worst sleep ever. Halfway through the study, the participants in the two conditions with recommendations began to receive recommendations via email. Their daily emails also included a question about whether they followed their recommendation on the previous day. Each week, all participants were also asked through a questionnaire if anything unexpected happened that week that might have affected their sleep, as well as how many times they exercised vigorously for at least 30 minutes. All the questions in the questionnaires are summarized in Table 1.

At the end of the study, participants filled out a questionnaire similar to the pre-study one, with questions from the ESS and PSQI, and additional ones about the recommendations and their experiences during the study.

4 COHORT-BASED RECOMMENDATIONS

4.1 Finding Users with Similar Profiles

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

4.1.1 Features for Cohort Selection. We collected the participants' height, weight, and gender. Due to privacy restrictions, we did not collect their age. According to the National Sleep Foundation, body mass index (BMI,

Table 1. Questions asked in each of the four types of questionnaires.

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

calculated based on a person's height and weight) plays a vital role in sleep quality, as obesity can cause undiagnosed sleep-disordered breathing [6]. Previous studies have shown that gender differences also affect sleep [16, 60]. We also asked participants about the average number of days per week they exercised vigorously for 30 minutes or more. Previous research has shown that exercise affects sleep quality [57, 61]. While other external factors such as caffeine consumption can affect sleep, we chose to focus on exercise because it can be tracked with the MS Band. Lastly, we asked participants to rate their overall sleep quality between "very good," "fairly good," "fairly bad," and "very bad." We calculated the average Sleep Recovery Index of each user in the MS Band dataset and mapped the score to a scale of 0 to 3, matching the sleep quality rating from the study participants. Thus, a score of 3 would map to "very good," and 0 would be "very bad." We also calculated the average number of times per week each user exercised vigorously for at least 30 minutes. Thus, the features used to identify a cohort of similar users were: height, weight, gender, number of days per week they exercised vigorously for at least 30 minutes, and their overall sleep quality (as a proxy to the Sleep Recovery Score).

4.1.2 Nearest Neighbor Search. We used the features described above to identify a subset of users who had similar demographic and habit profiles. This was achieved by a nearest neighbor search in k-dimensional space. First, we constructed an anonymous k-d tree from all users using normalized versions of height, weight, gender, frequency of exercise, and average sleep quality. Given the large number of users in our dataset, the height, weight, and BMI of the nearest neighbors that our algorithm returned were almost identical to those of the study participant. Future work can evaluate whether BMI is more effective in identifying appropriate nearest neighbors for specific groups of users, such as those with a high BMI who are particularly susceptible to poor sleep [28].

Next, we used a fast nearest neighbor search library for Approximate Nearest Neighbor Searching (ANN) [46], implemented in an R package called FNN [13], to identify users with similar profiles in the k-d tree. We experimented with a wide range of numbers (2 to 50) of nearest neighbors. We created a simple interface that took as inputs the 5 variables we used for identifying the cohorts, and checked how long it took for our system to return the set of nearest neighbors for each number between 2 and 50. While the MS Band dataset included millions of sleep records, we wanted to use a small enough number that could be attained easily in future studies. In case our cohort-based framework was successful, we wanted to make sure that it was reasonable enough to be used in real-life scenarios where perhaps users input their information in an online system and get a recommendation in real time. We found that $N = 5$ is optimal for real time scenarios (based on performance and quality considerations) and $N = 30$ for offline recommendations in this study. We picked 30 because that is a common sample size used in user studies, and generally leads to the possibility of a more robust statistical analysis for $N \geq 30$.

4.1.3 Selecting the Recommendation. Once we had identified each participant's cohort of nearest neighbors, we determined which recommendation would be most appropriate. To do so, we calculated the participant's median

Table 2. Template of the recommendation text in each of the four categories.

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

value for each dependent variable, and looked for its quartile rank among the values of the nearest neighbors (a low rank meant that this participant was doing worse than most of his/her neighbors). The dependent variables were (1) time to fall asleep, (2) number of wakeups per hour, (3) Sleep Recovery Index, and (4) sleep time.

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

Once the dependent variable was selected, we identified which of the following was affecting it the most: sleep duration (the time spent in bed, but not necessarily asleep), consistency, exercise frequency, or heart rate. There was a recommendation for each one of these options, as shown in Table 2. The user received the recommendation for the option that had the highest effect towards improving the target variable. For example, continuing the example from above, if a participant is doing the poorly on the time to fall asleep metric compared to their cohort, we identified what factor is affecting the time to fall asleep of the cohort users the most. If it turned out to be their bedtime consistency, we would send the participant the recommendation from the ‘consistency’ category.

In order to avoid confounding the effects of multiple recommendations, we only recommended one behavior change to each participant. Halfway through the study, the participants received the first email with their recommendation. Then, every day for the remaining two weeks, they received a reminder of the recommendation. Previous studies show that such periodic prompts increase adherence [30]. Similar to previous studies [20, 27, 32], the intervention consisted of a single behavior change, as it takes time for the change to have an effect on sleep and to avoid confounding effect of multiple interventions. While the focus of our study was not on behavior change, we wanted to design a study that maximized the benefit to the users in case the recommendation was helpful and they followed its advice.

4.2 Text of the Recommendations

Previous work has shown that people prefer recommendations containing more details about their own sleep [27]. In this study, we selected four main recommendation categories based on the MS Band data. Table 2 shows the text of each category (both recommendation conditions used the same template). ’Winding down’ refers to doing something to relax before bed to let your body recognize that it is time to slow down.

Table 3. The average number of nights tracked per condition, and average percentage difference in sleep time between before and after the recommendation period per condition. The sleep time of the cohort-based recommendations condition increased the most.

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

5 QUANTITATIVE SUMMARY

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

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

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

5.1 Microsoft Band Data

Table 3 shows the number of nights tracked per condition. For the conditions with recommendations, we employed a similar analysis procedure to SleepCoacher [27]: only the days that the participants said they followed the recommendation were used for analysis, which explains the drop in the number of data points. It is worthwhile to note that participants may have had troublesome days, making them unable to follow the recommendation, and that stress would have affected their sleep. However, since our goal was to evaluate the effectiveness of the cohort-based recommendations in comparison to that of the general ones, we are keeping the analysis as consistent as possible, and excluding the days when they did not follow the recommendation in both conditions.

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

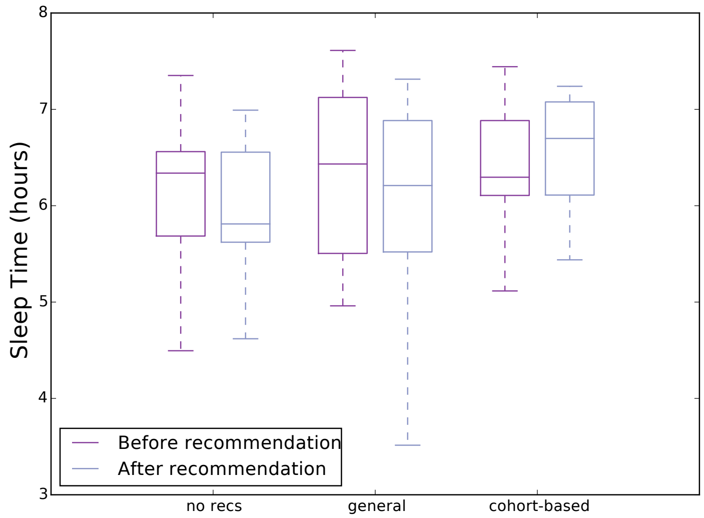

Fig. 1. Sleep time amounts per condition before and after the recommendation. While there was high variance, the three groups were not significantly different before the recommendation. However, after the recommendation, cohort-based recommendations resulted in longer sleep times.

Figure 1 shows the average sleep time for each condition before and after the recommendation. The condition without any recommendations has the shortest sleep time, whereas the cohort-based recommendations one has the longest sleep time. Furthermore, the highest percentage of people who increased their sleep time is in the cohort-based recommendations condition (63%), along with the highest average percentage of improvement (4.2%). In comparison, only 55% of people in the general condition and 53% of people in the no recommendations condition increased their sleep time.

5.2 PSQI and ESS Sleep Measures

We collected the PSQI and ESS scores of participants both before and after the study. Initially, the scores between the three conditions were not significantly different. A higher score on the PSQI (out of 21 points) and ESS (out of 24 points) scales is interpreted as worse sleep, and thus a decrease reflects improvement when looking at the change in these metrics. The PSQI scores of the no recommendations condition increased by 2.25 points on average, compare to 1 point for general condition, and only 0.42 points for the personalized condition. We summarize the changes of PSQI scores in Figure 2. However, both questionnaires are based on self-reported sleep quality factors such as duration and latency. Previous studies have indicated that people report higher awareness of their habits when they start self-tracking [22, 27, 41]. Therefore, these results might indicate that participants were more conscious of their sleep patterns when filling out the end-of-study questionnaire, causing their worsened sleep scores.

6 QUALITATIVE FINDINGS

In addition to the data collected by the MS Band and the ESS & PSQI questionnaires, participants provided written responses in the post-study questionnaire. We adopted an inductive approach to analyze the responses by performing a thematic analysis [15] on the data. Three researchers independently coded the written responses. Then, the coding schemes were discussed until a consensus was reached on the overall themes and sub-themes (Figure 3). The major themes were focused on whether participants learned something from being a part of the

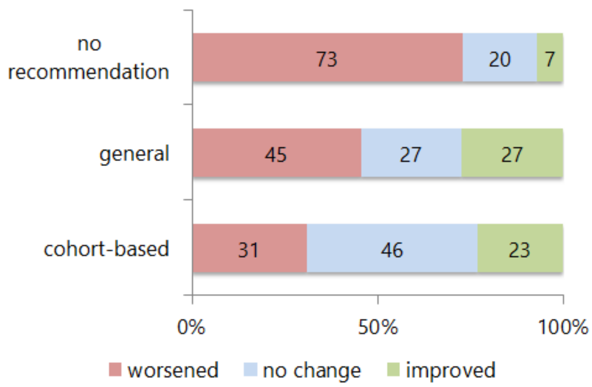

Fig. 2. The percentage of participants whose PSQI score changed in each direction per condition. The PSQI scores of the no-recommendations condition worsened the most.

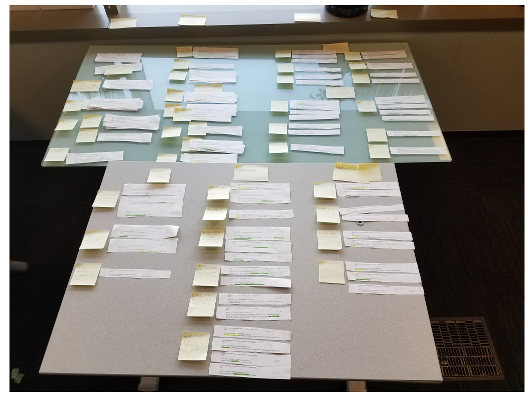

Fig. 3. Three of the authors performed a thematic analysis on the final survey data, which resulted in a few major themes discussed in the Qualitative Findings section.

study, whether they thought the recommendations were helpful, and their main reasons for following or not following the recommendations. The participants who adhered to the recommendations the most were the ones that increased their sleep time the most.

While the quantitative data analysis focused only on the participants who followed the recommendations for at least one day, the qualitative analysis also explored the answers of those who did not follow the recommendations at all to gain a deeper understanding of why participants chose to follow or not follow the suggestions. Due to the qualitative nature of the data, we did not seek measurable differences between the groups, so here we describe findings from both conditions—the most helpful aspects of the recommendations as well as the aspects that need further improvement. We include some nuances of the participants' experiences to highlight the complexity of adhering to recommendations. We further elaborate on these findings in the context of cohort-based recommendations in Section 7.

6.1 Helpful Aspects of the Recommendations

Similar to previous studies such as ShutEye [12] and SleepTight [22], we found sleep-tracking systems: (1) serve as reminders for well-known healthy sleep behaviors and increase people's awareness about their current sleep habits, and (2) help users identify patterns in the effects of various behaviors on their sleep. However, we also identified a type of self-reflection insight due to cohort formation: participants in our study were comparing themselves to the 'people' mentioned in the text of the recommendations.

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

Five other participants' reason for following the recommendations was that sleep was important to them and they wanted to try the suggested behavior change to see if it would "make a difference." However, receiving the extra nudge from the study was what triggered the increase in *"attention and focus,"* as P9 pointed out. Overall, participants found the recommendation helpful because it made them more conscious about their current sleep habits in general, which is a common finding to previous sleep studies such as ShutEye [12] and SleepCoacher [27].

P29, who received a cohort-based recommendation, said that "I consciously stepped away from screens before bedtime and started doing meditation to relax." P7, who received a cohort-based recommendation for consistency, pointed out that it made him "more conscious about getting sleep—trying to adjust my bedtime."

6.1.2 Emphasizing Impact of Various Factors on Sleep. On the one hand, two participants distinctly noticed that they slept better when they were following the recommendations. P11, for example, said that *"I found that if I did go to bed as suggested, I felt more rested on the next day and getting up was easier."* On the other hand, some also inferred causal effects about interferences to sleep, such as P15, who received a general recommendation about consistency, actually learned that *"temperature fluctuation definitely affected my sleep."* Another participant P9, whose recommendation was to wind down before bed, noticed that *"meditating makes a huge difference, but little else seems to."* Overall, five participants specifically pointed out that their sleep habits were "wrong" and should do something to change that, which matches the "trend" reflection type from Choe et al. [23]. P21, for example, found that *"I lack a lot of sleep, and need to to go bed earlier."* Thus, participants noticed patterns in how the recommendations affected their sleep.

6.1.3 Using Social Comparison as Behavior Change Motivation. The recommendations stated statistics about other people, which led some participants to compare themselves to them. Thus, some participants found out that they were not doing so well compared to others. P33 had gained Choe et al.'s "comparison" type insight [23] that *"I am getting less sleep than the body of users I am being compared against."* We consider implications of this finding in Section 7.

6.2 Lessons Learned About the Shortcomings of the Recommendations

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

6.2.1 Prior Commitments Made It Difficult to Fit the Recommendation in Daily Schedule. The most common reason for not following the suggestion, reported by 12 of the participants, was because it was too difficult to adjust their schedules. 9 of them followed it only three times or less per week. They had social and work commitments that prevented them from going to bed early enough to get more sleep or from waking up later. Unfortunately, according to sleep literature [4], a consistent bedtime and waketime schedule will have the best effects only when followed every day, including on weekends.

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

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

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

6.2.3 The Recommendation Did Not Seem Trustworthy nor Encouraging. Three participants expressed mistrust in the recommendation, caused by one of the following: (1) they doubted it was based on the correct metric, (2) the phrasing was unconvincing, or (3) it did not match their preconceptions about what good sleep, in particular, about amount of sleep. P18 pointed out that she thought the time to fall asleep was not measured properly. P14, who received a cohort-based recommendation for sleep duration, actually pointed out that "it didn't seem like a real person looked at that statement." P10 received a similar recommendation, but was disappointed because "there wasn't any encouragement like FitBit to actually try to change my sleep." P27, on the other hand, did not trust the recommendation because the suggested hours of sleep seemed inaccurate. She strongly believed that *"people need 7.5 hours of sleep,"* so it did not seem logical that the recommendation was for a different amount.

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

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

Alternatively, two participants did not find the recommendation helpful because they were hoping to improve a different aspect of their sleep that they were having trouble with. P20, for example, who was recommended to wind down before bed said that *"I have trouble staying asleep and getting quality sleep, and I have no trouble falling asleep. The recommendation was too generic and doesn't actually apply to my sleeping habits."*

6.2.5 The Recommendation Did Not Lead to Immediate Improvement. Finally, two participants pointed out that they did not think the recommendation was helpful because they felt better on the days that they did not follow it. Specifically, both of those participants were referring to the fact that they would rather wake up naturally than use an alarm clock for a preset wake up time. Furthermore, another participant pointed out that"the exercise does affect sleep but other things could affect it way more." This quote also represents similar feedback from 3 other participants who did not think that what the recommendation was suggesting to change was really the cause of their sleep issues.

7 DISCUSSION

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

7.1 Selecting a Cohort of Similar Users

Collaborative filtering recommender systems use a variety of algorithms to recommend items to a user that other similar users have liked [49]. As in our approach, the similar users are clustered together based on relevant attributes. Usually, that means that for each user, a set of nearest neighbors is found with whose past ratings there is the strongest correlation [9]. However, while those recommender systems are based on what other users have liked, our method is based on what other users with better sleep quality have done. We chose to base cohorts on the simple model of demographic information (BMI and gender), self-perceived sleep quality, and self-reported exercise level. While our approach used only a nearest neighbors classifier, previous studies have shown that clustering can be added as a pre-processing step to increase efficiency[68].

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

As stated in Section 6.2.1, twelve of the twenty-four participants who received any kind of recommendation brought up the issue that the suggested behavior change was not attainable due to constraints in their schedule. Nevertheless, a part of them attempted to incorporate it in their lives as much as possible. The participants who followed their recommendations the most were the ones that improved their sleep the most. However, for the majority of participants, the suggestion was not actionable enough, so they did not follow it or benefit from it. Therefore, based on the qualitative findings from our study, the main deterrent to improving sleep was the practicality of the recommendations. Thus, in this section, we focus on ways we can improve the framework to generate actionable and impactful recommendations, inspired by social and behavioral theory.

To provide more actionable recommendations, a future system can ask the user for what limitations already exist in their daily routine and build a cohort based on people with similar restrictions. Furthermore, the system could also ask users for what activities (e.g., exercise and relaxation) they engage in to build on existing habits, rather than attempt to introduce a completely new one. That way, the cohorts would be based more realistically on people whose daily lives include similar opportunities for improvement and challenges. Finally, the system can also ask the user what they want to improve to make sure it finds a cohort that worked towards a similar goal, addressing participant's feedback from Section 6.2.4.

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

Additionally, while the "people like you" in recommender systems like Netflix are usually hidden [69], we found that participants in our study generally wanted to know more about who those people are. Therefore, perhaps in the health domain the framework for cohort-based recommendations needs to provide details about how the cohort of nearest neighbors is similar to the given user. It is important to be able to balance a sense of cohesion in a cohort while preserving individual privacy.

Selecting the right cohort that the user identifies with is critical for inspiring behavior change. According to social cognitive theory [10], self-efficacy is the belief in one's ability to perform certain behaviors in a given context, and that they will have an effect on one's life. We can use this notion to show users that people with similar life constraints and schedules are still changing their behavior in some way and improving their sleep. This could potentially motivate the user to also implement the changes and, in turn, increase their self-efficacy.

7.2 Phrasing of the Cohort-Based Recommendations

Since the only recommendations in sleep research studies so far have been either based just on the individual user [27] or on the general sleep hygiene guidelines, there was no clear direction on how to phrase the cohort-based recommendations. We chose to frame the recommendations as general advice, followed by specific averages for the cohort and the individual user. While the focus in this study was not on how to best phrase the recommendations, participants provided us with valuable feedback on how to improve them in the future. Below, we focus on six main improvements based on our findings: (1) increase trustworthiness of the recommendation, (2) include information about how helpful and worthy this recommendation is, incorporating statistics from people in the cohort who have attempted it previously, (3) start with small actionable steps, (4) explore less known sleep recommendations to introduce novelty, (5) phrase them differently for collectivist vs individualist societies, and (6) be clear about how the cohort is similar to user.

The trustworthiness of the recommendation was brought up in a few ways from different participants in Section 6.2.3. One important tension that we identified was between general sleep hygiene guidelines and personalize recommendations. For example, P27 strongly believed in the general advice that "people need 7.5 hours of sleep," so when she received her specific cohort-based recommendations for a different amount, she deemed it untrustworthy. Thus, the discrepancy between the well-known tips and the new recommendations causes skepticism. One way to mitigate that would be to include the source of the suggestion, as discussed in ShutEye [12]. Similarly, cohort-based recommendations might benefit from including short snippets with facts from sleep literature, presented in layman's terms to educate the participant.

Furthermore, recommendations might be more trustworthy if it is obvious that they helped similar users. As illustrated in Section 6.2.2, participants have a preconception of what is considered an impactful and worthy outcome for which to change their behavior. Five participants reported that they ignored the recommendations because they considered them unworthy of the effort required for the behavior change. Further research is needed to identify where that boundary lies. The trustworthiness of the recommendations can be increased by being more transparent about how each metric is calculated. This is also related to participants' desire for more specific details. However, this brings up another interesting tension: what is the right balance between including all the details that we have to make the recommendation believable and at the same time making sure the user is not being overloaded with information.

The third improvement we suggest is grounded in behavior change literature: the phrasing of the intervention affects the way the patient understands it [52]. The Nudge Theory, for example, emphasizes that the intervention must be simple and easy to follow, such as presenting fruit at an eye level and fried chips on a top shelf [59]. Similarly, the recommendations in the domain of health have to be as clear as possible. According to Fogg's Tiny Habits, in addition to being clear, they also need to start with the smallest actionable step possible [5]. Exercise seems to be the most difficult recommendation to follow, as it involves introducing a new habit in the cases when participants do not usually workout. The participants in our study that improved their sleep the most were the ones that followed their recommendation most often. This leads to an interesting question of whether following the recommendation on just a few nights leads to enough of an effect and if the frequency can potentially be built up from there. If it is not enough, then the question is whether the behavior change is worth pursuing at all.

Section 6.1.1 described that participants liked receiving recommendations because they served as a reminder of a health behavior they already knew they should be engaging in. It is not surprising that participants were already aware of these suggestions since we specifically chose recommendations that were part of the general sleep hygiene guidelines. However, not all general guidelines will work for everyone, so another improvement in could be to explore whether less well-known guidelines based on the specific behaviors of the cohort are appealing enough to be followed.

An interesting study by Cialdini et al. assessed the impact of two social influence principles: 1) commitment/consistency and 2) social proof on participants' decisions. They found Americans and Poles are impacted differently by the two principles, with Americans being more impacted by the commitment/consistency principle [24]. This discrepancy implies that behavior change recommendations might need to be phrased differently according to the user's background. Specifically, cohort-based recommendations that rely on parallels between the individual and other similar users might be most effective in collectivist societies.

Lastly, based on participant feedback, the recommendations would also be more dependable if they knew how the other people in their cohort were similar to them (as described in Section 7.1).

7.3 Social Comparison and Interaction

According to the theory of social comparison, people compare themselves to others with similar opinions or abilities [29]. In the context of our study, this emphasizes the need of a participant to know who the other people in their cohort are in order to know how closely related they are to them. Some participants in Section 6.1.3 specifically mentioned that they compared themselves to the cohort. Past applications such as Shakra [8] and Houston [25] helped users compare their fitness data to that of their friends, but the cohorts in our study were strictly strangers. Future work could explore whether cohorts based on people we know give a basis for better sleep recommendations.

Previous research has shown that sharing fitness information with friends was helpful [8], that online social networks may be effective in behavior change interventions [43], and that human interaction was successful in promoting increased physical activity among middle-aged and elderly people [37]. The benefit of cohort-based recommendations specifically are that participants can be following the same recommendations, or can even be a part of the same larger cohort for further support. Additionally, users might benefit from being able to interact with their cohort: both for giving reminders to each other and for inspiration for behavior change.

7.4 Improving the Recommendation Generation

In this study, we used an algorithm to generate recommendations based on nearest neighbors, and we identified design hypotheses about what could potentially improve the overall framework. One suggested change is to use only recently generated data from current active users. In the case that such information is not available, we could also use a similar time frame and location from previous years. This could ensure similar weather trends to give more accurate recommendations as seasonal variations can have great impact on people's sleep. Another possible improvement is to adjust the number of nearest neighbors. We picked a number to match studies conducted without access to such a large dataset, but future studies can explore the effect of a cohort size.

Finally, since two participants in Section 6.2.4 received a recommendation for a variable different from what they were hoping for, future systems could ask users what they would like to focus on and provide intervention suggestions specifically for that variable. This design hypothesis is in agreement with goal-setting theory [42], according improvement is highest when the user sets the goals. Thus, the system could identify the people in this user's cohort that are doing better on the target variable, and suggest a recommendation based on what they are doing differently from the user.

7.5 Limitations

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

8 CONCLUSION

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

REFERENCES

[1] 2014. Lack of Sleep is Affecting Americans, Finds the National Sleep Foundation. (2014). https://sleepfoundation.org/media-center/press-release/lack-sleep-affecting-americans-finds-the-national-sleep-foundation

[2] 2014. Polysomnography (sleep study). (Dec 2014). http://www.mayoclinic.org/tests-procedures/polysomnography/basics/definition/prc-20013229

[3] 2016. Tips for Better Sleep. (Jul 2016). https://www.cdc.gov/sleep/about_sleep/sleep_hygiene.html

[4] 2017. Sleep Hygiene Tips - Research & Treatments - American Sleep Assoc. (2017). https://www.sleepassociation.org/patients-general-public/insomnia/sleep-hygiene-tips/

[5] 2017. Tiny Habits. (2017). http://tinyhabits.com/

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

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

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

[9] Marko Balabanović and Yoav Shoham. 1997. Fab: content-based, collaborative recommendation. *Commun. ACM* 40, 3 (1997), 66–72.

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

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

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

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

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

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

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

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

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

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

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

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

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

[23] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding Self-Reflection: How People Reflect on Personal Data Through Visual Data Exploration. https://www.microsoft.com/en-us/research/publication/understanding-self-reflection-people-reflect/

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

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

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

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

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

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

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

[31] Tian Hao, Guoliang Xing, and Gang Zhou. 2013. iSleep: unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, 4.

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

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

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

[35] Murray W Johns. 1991. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *sleep* 14, 6 (1991), 540–545.

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

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

[38] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3063–3072.

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

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

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

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

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

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

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

[46] David M. Mount and Sunil Arya. 2010. ANN. (2010). http://www.cs.umd.edu/~mount/ANN/

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

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

[49] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–58.

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

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

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

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

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

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

[56] Alireza Sahami Shirazi, James Clawson, Yashar Hassanpour, Mohammad J Tourian, Albrecht Schmidt, Ed H Chi, Marko Borazio, and Kristof Van Laerhoven. 2013. Already up? using mobile phones to track & share sleep behavior. *International Journal of Human-Computer Studies* 71, 9 (2013), 878–888.

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

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

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

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

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

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

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

[64] Shu-Lin Wang, Young Long Chen, Alex Mu-Hsing Kuo, Hung-Ming Chen, and Yi Shiang Shiu. 2016. Design and evaluation of a cloud-based Mobile Health Information Recommendation system on wireless sensor networks. *Computers & Electrical Engineering* 49 (2016), 221–235.

[65] Douglas L Weeks, Stacie Borrousch, Andrew Bowen, Lisa Hepler, Mary Osterfoss, Andrea Sandau, and Frank Slevin. 2005. The influence of age and gender of an exercise model on self-efficacy and quality of therapeutic exercise performance in the elderly. *Physiotherapy theory and practice* 21, 3 (2005), 137–146.

[66] Martin Wiesner and Daniel Pfeifer. 2014. Health recommender systems: concepts, requirements, technical basics and challenges. *International journal of environmental research and public health* 11, 3 (2014), 2580–2607.

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

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

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

Received February 2018; revised May 2018; accepted September 2018