Dear Editor(s),

We are sending you our manuscript entitled “An Observational Study of Job Performance, Sleep Behavior, and Their Indicators in Mobile App Usage” in response to the Big Data Call for Papers.

Establishing the relationship between sleep behavior and job performance has been a challenge in the past due to the difficulty of collecting objective measures in real-world settings. A real-time indicator of a person’s job performance and sleep quality would enable people to increase their awareness of their own abilities and to manage their workloads in a more efficient manner. Unfortunately, objective sleep metrics and job performance metrics can be both difficult and invasive to collect in practice. Our work leverages the emergence of sleep-tracking technologies and the ubiquity of smartphones and the resulting large observational datasets to examine the relationship between sleep behavior, cognition, and psychomotor performance in real-world settings.

In this work, we conducted a 1.6-year-long observational study with 274 participants during their everyday lives. In total, the dataset included 30,618 tracked nights of sleep, 11,140 days of app usage, 289 performance measurements. To the best of our knowledge, this is the largest study to data on relationship between sleep, job performance, and sleep-tracking app usage in real-world setting. The participants tracked their sleep behavior using a mattress sensor and studied their sleep data using an accompanying smartphone app. For a subset of that population including salespeople (N=15) and athletes (N=19), we also tracked job performance through daily sales statistics (118 days) and game-day ratings (171 games). Using the data collected from the study, we investigated the relationship between sleep behavior, job performance, and sleep-tracking app usage. We show that cumulative sleep measures were significantly correlated with job performance metrics. Concretely, when the average salesperson lost one hour of sleep daily for one week, the number of contracts they were able to establish decreased by 9% the next day; when the average athlete lost the same amount of sleep throughout the week, their game grades dropped by 9.5%. We also find that smartphone interaction time correlates with sleep metrics like sleep history and time awake while accounting for potential confounders like circadian rhythms and user-specific baselines in our statistical models. Finally, we show that interaction time is correlated with athletic job performance(ρ=-0.296,p=0.0456).

By taking advantage of ubiquitous technologies like mattress sensors and smartphones, our research addresses the increasing desire within various organizations to evaluate job performance through data. This highlights an interesting opportunity for future assessments of sleep and performance in uncontrolled settings. We believe that this article will be of particular interest to researchers in diverse disciplines, including sleep science, behavioral psychology, public health, and computer science. The article will also attract readers from the general public who are interested in leveraging ubiquitous sleep-tracking technology to establish healthy sleep behaviors and improve their job performance.

We confirm that this manuscript has not been published elsewhere and is not under consideration by another journal. Our study and data analyses were conducted in accordance with the Institutional Review Board at the University of Washington. All authors have approved the manuscript and agree with its submission to the SLEEP journal’s Big Data Call for Papers.

Sincerely,
Chunjong Park

An Observational Study of Job Performance, Sleep Behavior, and Their Indicators in Mobile App Usage

CHUNJONG PARK*†, University of Washington, USA
MORELLE ARIAN*, University of Washington, USA
XIN LIU, University of Washington, USA
LEON SASSON, Rise Science
ALEX MARIAKAKIS, University of Washington, USA
SHWETAK PATEL, University of Washington, USA
TIM ALTHOFF, University of Washington, USA

*Both authors contributed equally to this research.
†Corresponding Author

Authors' addresses: Chunjong Park, cjparkuw@cs.washington.edu, University of Washington, Paul G. Allen Center, 185 E Stevens Way NE, Seattle, Washington, 98195-2350, USA; Morelle Arian, morellea@cs.washington.edu, University of Washington, Paul G. Allen Center, 185 E Stevens Way NE, Seattle, Washington, 98195-2350, USA; Xin Liu, xliu0@cs.washington.edu, University of Washington, Paul G. Allen Center, 185 E Stevens Way NE, Seattle, Washington, 98195-2350, USA; Leon Sasson, leon@risescience.com, Rise Science; Alex Mariakakis, atm15@cs.washington.edu, University of Washington, Paul G. Allen Center, 185 E Stevens Way NE, Seattle, Washington, 98195-2350, USA; Shwetak Patel, shwetak@cs.washington.edu, University of Washington, Paul G. Allen Center, 185 E Stevens Way NE, Seattle, Washington, 98195-2350, USA; Tim Althoff, althoff@cs.washington.edu, University of Washington, Bill Melinda Gates Center, 3800 E Stevens Way NE, Seattle, Washington, 98195-2355, USA.

ABSTRACT

Study Objectives

Although it is commonly conjectured that a good night's sleep is important for job performance, this relationship has historically been hard to quantify due to the difficulty of capturing objective measures in real-world contexts. In this work, we present the findings from an observational study of objectively measured sleep, job performance, and sleep tracking app usage.

Methods

We conduct an observational study in which we tracked the sleep behaviors of 274 participants who used a mattress sensor. For a subset of that population including salespeople (N = 15) and athletes (N = 19), we also track job performance through daily sales statistics and game-day ratings. After analyzing the data to explore the relationship between sleep behaviors and job performance, we investigate the utility of timed smartphone interactions as a proxy cognition indicator.

Results

We show that cumulative sleep measures were significantly correlated with job performance metrics. When the average salesperson lost one hour of sleep daily for one week, the number of contracts they were able to establish decreased by 9% the next day; when the average athlete lost the same amount of sleep throughout the week, their game grades dropped by 9.5%. We also find that smartphone interaction time correlates with sleep history, time awake, and circadian rhythms. Finally, we show that interaction time is correlated with athletic job performance (ρ=-0.296, p=0.0456), but not with salespeople's performance.

Conclusions

We found positive correlations between cumulative sleep behaviors, job performance, and a passively captured smartphone interaction metric.

KEYWORDS

Performance, job performance, interaction speed, sleep behavior, sleep debt, sleep history, mattress sensor

STATEMENT OF SIGNIFICANCE

Having a real-time indicator of a person's job performance would increase self-awareness of a person's own abilities and would enable employers to manage workloads in a more efficient manner. Unfortunately, objective job performance metrics can be both difficult and invasive to collect in practice. Given the emergence of sleep-tracking technologies for continuously measuring sleep without user intervention, our work demonstrates the utility of sleep sensing as a potential indicator of job performance. In addition, we demonstrate that passively sensed, timed smartphone interactions provide a unique opportunity for researchers to assess cognition and productivity in a continuous, unintrusive manner.

INTRODUCTION

Sleep is essential to human function, affecting memory [59], mood [7], energy [9], and alertness [2]. Total sleep deprivation, even for a single day, can affect people's ability to perform simple tasks like reaction time tasks such as the psychomotor vigilance test (PVT) [18, 30] and mental math [5]; the same holds true for studies of chronic sleep restriction [40, 57]. The consequences of sleep deprivation were found to be comparable to the cognitive and motor impairments experienced during alcohol intoxication [61]. Although prior literature has led researchers to suspect that the side effects of poor sleep behaviors can impact real-world job performance, this relationship has remained largely unquantified.

Researchers have often leveraged sleep behavior surveys to assess the relationship between sleep and job performance [10, 20, 31], yet self-reports are often subjective and imprecise in nature [23, 54]. In recent years, sleep tracking has become more commonplace due to the introduction of commercially available sleep-tracking technologies like smartphones, smartwatches, and mattress sensors [35]. Researchers have leveraged these technologies to study the relationship between sleep quality and people's performance at either cognitive tasks [32, 33, 44] or sports [41, 60]. Unfortunately, these studies have examined performance in contrived scenarios rather than everyday life. Many careers involve a complex combination of cognitive and psychomotor tasks, so it is unclear how contrived tasks translate to higher level performance.

In this work, we present the findings from a 1.6-year-long observational study of objectively measured sleep and performance across 274 participants, all of whom tracked their sleep using a mattress sensor and reviewed their data using an accompanying mobile app. A subset of that population (12.4%) worked in two organizations—a bankruptcy law firm consultancy and the National Football League—that have widely accepted metrics for job performance. Since job performance metrics can be difficult to capture in practice, we explore the possibility of using timed interactions with the sleep-tracking app as an indicator for performance. The ways with which users interact with their smartphone provides insights into their psychomotor and cognitive function [4, 24, 43], providing an opportunity for nonintrusive app-based performance measurements. We examine the amount of time a user spends interpreting the information on a screen in the app (*interaction time*) as an instantiation of an app-based performance metric. Lastly, as users engage with their sleep-tracking data through an app, they are likely to learn more about their sleep habits [11, 13, 14]. However, increased awareness of sleep habits does not guarantee an improvement in sleep behavior [39]. The longitudinal nature of our dataset allows us to examine whether engaging with sleep-tracking technology leads to improvement in the context of automatic sleep sensing.

Specifically, we leverage this data to examine the following research questions:

RQ.1 Is sleep behavior correlated with job performance?
RQ.2 Is app-based performance correlated with sleep behavior?
RQ.3 Is app-based performance correlated with job performance?
RQ.4 Is engagement with automatic sleep-tracking technology associated with improved sleep behavior?

METHODS

Study Population

The data we use in this paper was collected through a study of 274 users of a mobile app developed by Rise Science[1]. Participants who enrolled into the study were sent a kit consisting of an Emfit QS[2], a sleep-tracking mobile app[3][4], a blindfold, and orange-tinted glasses; the latter two accessories are common interventions for improving sleep. The Emfit QS is a highly sensitive pressure sensor that lies underneath the user's mattress (or their preferred side of the mattress when the bed is shared). The sensor uses ballistocardiography to track heart rate, breathing rate, and movement. In past studies, the Emfit QS has been validated against a standard clinical heart rate monitor and polysomnography equipment [25, 36, 50]. Within the sleep-tracking mobile app, participants can access and visualize their own sleep data, view sleep session summaries, create sleep plans, and learn about the importance of sleep.

[1]https://www.risescience.com/
[2]https://qs.emfit.com/
[3]https://play.google.com/store/apps/details?id=com.risesci.risesciapp
[4]https://apps.apple.com/us/app/rise-science/id1107659850?app=itunes&ign-mpt=uo%3D4

The data collection period started on May 2017 and ended on December 2018, spanning 592 days (1.6 years). Recruitment happened throughout that period, and participants joined and left the study as they pleased. Out of 274 participants who contributed to the dataset, 15 (5.5%) were salespeople at a bankruptcy law firm consultancy and 19 (6.9%) were athletes from the National Football League. Demographic data like age and gender were not collected to maintain participant privacy. Participants did not receive explicit instructions from the research team and were free to follow whatever sleep schedule they chose. Participants were also free to use the Emfit QS and sleep-tracking app at will; if they had to travel while participating in the study, they could choose to either bring the Emfit QS with them or leave it behind. The mobile app sent users notifications, reminders, and recommendations for improving their sleep (e.g., reducing caffeine intake, wearing orange-tinted glasses); users were free to disable these features at any time. Our retrospective data analysis was conducted in accordance with the Institutional Review Board at the University of Washington.

Sleep Behavior Metrics

The Emfit QS reports the following metrics to describe a single night's rest: bedtime, wake time, midpoint, time-in-bed, and total sleep duration. Time-in-bed measures how long a person is in their bed, thus only requiring accurate presence detection. Total sleep duration, on the other hand, estimates how long a person is actually asleep in their bed, thus requiring both accurate presence and sleep detection. Because total sleep duration is susceptible to more sensing errors, we exclude it from the analyses reported in this paper; nevertheless, the two metrics were strongly correlated in our dataset ($\rho = 0.85, p < .001$) and produced comparable results in most cases. Looking beyond a single night's sleep, cumulative metrics across multiple nights can provide further insight into participants' sleep behavior. We use sleep debt [17, 28, 34, 57], the weighted accumulation of sleep loss, as one of those measures. Sleep debt is calculated using the following formula:

$$\sum_{i=1}^{7} -e^{-i/7} * (\text{SleepNeed} - \text{TimeInBed}_i)$$

where i is the number of days in the past. Note that the difference between sleep need and debt is weighted by a decaying exponential with a time constant of 7 days [49], indicating that recent measurements have greater importance. Whenever a participant skips a day of tracking and a time-in-bed value is missing, their average time-in-bed over the past week is imputed. Sleep need is typically estimated in a controlled laboratory study, making it challenging to estimate sleep debt in the wild. Therefore, we estimate sleep need using the approach proposed by Kitamura et al. [34]. Their approach involves using long nights of sleep to predict the difference between sleep need and habitual sleep (i.e., the average time-in-bed over two weeks) for a minimum of four nights. We also introduce a simplified *sleep history* metric that avoids the notion of sleep need, but still captures an aggregate measure of sleep behavior:

$$\frac{1}{\sum_{n=1}^{7} e^{-n/7}} \sum_{i=1}^{7} e^{-i/7} * \text{TimeInBed}_i$$

The calculation of sleep history is normalized such that weights sum to one, making the metric more interpretable as a weighted average of time-in-bed over the past week.

Job Performance Metrics

Through organizational partnerships, we were able to gather job-specific performance metrics for a subset of our study population. We describe these metrics below:

Performance Metrics for Salespeople. The salespeople who participated in our study ($N = 15$) work at a bankruptcy law firm consultancy. Their job entails fielding phone calls from potential clients in need of bankruptcy relief and referring those callers to an attorney. The employees collect a fee upon successfully hiring a client, which is the company's primary revenue source. Employees in this company are evaluated on a variety of metrics related to that revenue stream, such as the amount they collect in fees. However, the distribution of fees is highly variable ($250–$1750) and primarily dependent upon the clients rather than the employees themselves. Therefore, we focus on the *number of hires* the salespeople were able to establish as their job performance metric. Although work hours were generally consistent across the company, we normalized the *number of hires* a salesperson made by the number of hours they worked that day to account for whatever variance remained.

Performance Metrics for Athletes. The athletes who participated in our study ($N = 19$) play in a professional American football league in the United States. We gather job performance metrics for the athletes' performance during weekly games using Pro Football Focus[5] (PFF). PFF evaluates athletes using the following procedure [45]: two expert analysts score every play the athlete is involved in, a third expert resolves disagreements between those experts, an external group of ex-players and coaches verifies the scores, and then the scores are summed together and normalized to a grade between 0-100. Although PFF is not purely quantitative, the experts can account for in-game context that is lost by purely statistical methods (e.g., injuries, matchups). For this reason, PFF has been used in the past literature for assessing performance in football [8, 19, 46].

In American football, each player has their own unique skill set according to their position; on offense, for example, quarterbacks are typically known for their throwing ability and wide receivers are know for their speed and catching ability. The notion of positional specialization makes it difficult to compare athletes across positions in a purely quantitative way, especially since some skills are position-specific. PFF's method of expert analysis and normalization allows for the calculation of an *overall game performance* grade that enables comparisons across positions, overcoming this issue.

Sleep Tracking App Usage Metrics

Users must interact with a sleep-tracking app in order to examine their sleep summaries, so we leverage these interactions as a novel source of data. We take inspiration from Althoff et al. [4] by using *interaction time*—the time between two touch events in the app—as an app-based performance metric. Interaction time is not meant to be a direct replacement of the PVT; instead, it serves as a more general measure of cognition by measuring the user's ability to process information on the app's screen. Interaction speed can be confounded by the type of data and the quantity of data available to the user. To account for these confounds, we restrict our analysis of interaction time to transitions from the home screen (shown in Figure 1) to three endpoints: (1) sleep details view, (2) progress view, or (3) closing the app. We also calculate days of app engagement as a metric related to how often the user engages with the app. This metric serves as an indicator for understanding how much users engage with their sleep through the app.

Data Filtering and Post-Processing

Sleep behavior, job performance, and app usage metrics were collected from separate sources at different intervals. Therefore, post-processing was needed to join and collate them. The data was first processed based on job-specific sampling issues, and then the data was processed based on sleep behavior and app usage data qualities.

Job-Specific Filtering. Job performance data for the salespeople was collected on a daily basis. Therefore, every night of sleep that a salesperson tracked with their Emfit QS was collated with the job performance metric from the next day. Aligning the data streams for the athletes was more difficult since they had games on a weekly basis. The athletes also had to travel to games away from their home stadium, leaving larger gaps in their sleep-tracking data. To accommodate these issues, we aligned the weekly PFF grades with the sleep behavior metrics from the

[5] https://www.pff.com/

most recent tracked night of sleep within the two days before the relevant game day; if no nights were tracked in that span, the game grade from that week was filtered out.

General Post-Processing. The calculation of time-in-bed included naps, which were either automatically annotated if the user's bedtime or wake time fell in the afternoon (12:00-18:00) or manually annotated by the user. Naps appeared in 9.3% of the nightly sleep metrics (62% automatically tagged vs. 38% manually annotated), contributing an additional 1 hour and 13 minutes to time-in-bed on average. Sleep events when the user spent more than 16 hours in bed in a single session were attributed to faulty sensing and removed from the dataset. The remaining nights, along with imputed averages for missing values, were used for calculating sleep debt and sleep history. A full week of sleep data was available for calculating 46.9% of the cumulative sleep metrics, meaning that no imputation was needed for them; three or more nights were only missing in 12.7% of the cumulative sleep metrics. When cumulative sleep metrics were calculated without imputation, the standard deviation of the times within the same week was only 1 hour and 10 minutes; this shows that there was not significant variance within a week, justifying the use of a short-term average. For the analyses related to app-based performance, interaction events that were shorter than 0.45 seconds (2.5$^{\text{th}}$-percentile) were excluded since were likely accidental or automatically generated by the app itself, and events longer than 54.83 seconds (97.5$^{\text{th}}$-percentile) were excluded since they were likely indicative of the user engaging in another activity. Summary statistics of the resulting dataset after filtering are shown Table 2. The large standard deviations in the various metrics are due to the logistics of our study. Participants were recruited throughout the 1.6-year-long period, so some people had many more opportunities to use the sleep-tracking tools than others.

Correlational Analyses

Using D'Agostino's K^2 test [12], we determined that the job performance metrics in our dataset were non-normally distributed (number of hires: K^2=21.37, p=2.3×10^{-5}; game grades: K^2=14.87, p=5.9×10^{-4}). The same holds true for app-based performance (K^2=5177, p<1.0×10^{-20}) and app event count (K^2=71.60, p=2.8×10^{-16}). Therefore, we use Spearman's Rank Correlation (rho) across all correlational analyses throughout this paper.

Research Questions and Methods

Understanding the relationships between app usage, sleep behavior, and job performance metrics could help improve real-world productivity. Concretely, we aim to answer these four research questions and describe how we analyze the data to investigate each of these questions below (Figure 2):

RQ.1 Is sleep behavior correlated with job performance?
RQ.2 Is app-based performance correlated with sleep behavior?
RQ.3 Is app-based performance correlated with job performance?
RQ.4 Is engagement with automatic sleep-tracking technology associated with improved sleep behavior?

RQ.1: The Relationship Between Sleep Behavior and Job Performance. Our first research question examines whether better sleep improves job performance [40, 41]. Because we require job performance data to investigate this relationship, this analysis is limited to the 19 professional athletes and the 15 salespeople in our dataset. We calculate correlation coefficients between the job performance metrics and three sleep behavior metrics: (1) time-in-bed, sleep debt, and sleep history. Sleep metrics can also vary across individuals due to genetic predisposition and other factors [3, 53], so we repeat the correlation analysis using standardized sleep behavior metrics according the Z-score within each individual's data. Participants who did not track at least 5 nights of sleep were excluded from this analysis to ensure that the data was representative of their typical sleep behavior.

RQ.2: The Relationship Between App-Based Performance and Sleep Behavior. Using the PVT, sleep researchers have demonstrated that psychomotor and cognitive function improve with better sleep behavior [47, 48]. Separately, computing researchers have shown that the timing between interaction events in a desktop or smartphone can be an indicator of psychomotor and cognitive function [4, 58]. Our second research question aims to join these two bodies of literature and establish that an app-based performance metric—interaction time, in our case—can be used to assess sleep behavior. We calculate the correlation coefficients between all of the available sleep metrics and interaction time; since job performance data is not needed to investigate this question, we conduct this analysis on all 274 participants in our dataset.

It is well established that psychomotor and cognitive function vary throughout the day due to circadian rhythms homeostatic sleep drive, and sleep inertia, collectively forming the three-process model of sleep [2, 4, 22, 42]. Any performance indicator should therefore be sensitive to variations of time and sleep. To examine whether this is the case for our app-based performance metric, we evaluate the relationship between interaction time and four different measures: time of day, time since wake-up, sleep debt, and sleep history. We use a generalized additive model proposed in prior work [4] and extend it with random effects intercepts for each user. Modeling users with random effects not only accommodates user-specific performance baselines, but also accounts for device-specific effects like the rendering capabilities of the user's smartphone.

RQ.3: The Relationship Between App-Based Performance and Job Performance. We hypothesize that app-based performance provides a low-level, *in-situ* measurement of psychomotor and cognitive performance that relates to high-level performance in the workplace. However, we cannot presume that this statement is true even if we identify statistically significant correlations for the previous two research questions, because correlations are not necessarily transitive. In other words, if sleep behavior is positively correlated with both job performance and app-based performance, we cannot assume that job performance and app-based performance are positively correlated. We therefore evaluate the correlation between these two data sources in a separate analysis within the subpopulations of athletes ($N = 19$) and salespeople ($N = 15$). We also fit least squares models between app interaction time and job performance metrics to determine effect sizes.

RQ.4: The Relationship Between App Engagement and Sleep Behavior. Sleep-tracking technology is designed to make users more aware of their sleep habits in order to encourage behavior change [11, 13, 14]; however, prior work has questioned whether such awareness translates to improved sleep behavior [39]. Assuming that engaging with sleep-tracking technology leads to increased awareness of sleep behavior, our longitudinal dataset allows us to explore the link between awareness and improvement in the context of an automatic sleep sensor and its accompanying sleep-tracking app. We begin this analysis by calculating the correlation coefficients between all possible combinations of app engagement and sleep behavior metrics for all 274 participants in our dataset. We also examine if app engagement affects sleep behavior consistency (e.g., going to bed at the same time every day) by considering the standard deviations of the sleep metrics.

Correlations between app engagement and sleep could simply be a result of selection effects, where particular users may happen to be simultaneously sleeping well and highly engaged in the app, rather than having a causal effect between them. Therefore, we leverage the longitudinal nature of our data to study whether higher app engagement was associated with improvements in sleep behaviors over time. We calculate changes in sleep behavior by comparing the average time-in-bed from the first D days to the last D days in a 5-week period; we call this difference *sleep improvement*. In our analysis, we vary D from four to seven days. The larger the value of D, the larger the sample size for calculating sleep behavior; however, larger values of D will also attenuate the effects of interest as users may be improving their sleep behaviors during the first D days. For each D, we filter out users with less than $D/2$ days of sleep tracking to ensure there is enough data to accurately estimate their sleep behavior. We compare sleep behavior improvement with the number of app engagement days between the first and last D days.

RESULTS

We also use these statistical analyses to explore the possibility of leveraging passively captured app interaction data as a performance indicator.

RQ.1: The Relationship Between Sleep Behavior and Job Performance

The analyses for this research question cover 19 athletes and 15 salespeople who, respectively, contributed data from 171 and 118 nights of sleep with corresponding job performance metrics. The correlation coefficients between the sleep behavior and job performance metrics in our dataset are presented in Table 3. The analysis reveals positive, statistically significant correlations in some, but not all, cases. For the salespeople, sleep debt was positively correlated with the number of hires they made (ρ=0.218, p=0.022). For the athletes, normalized sleep history (ρ=0.179, p=0.020) and sleep debt (ρ=0.166, p=0.031) were both positively correlated with game performance. Fewer correlations were found for the salespeople than the athletes, which could be due to the nature of their jobs. The athletes rely on millisecond-scale reaction times during their games, whereas salespeople do not need to operate at such a rapid pace. These results could imply that careers focused on physical and psychomotor skills may be more strongly affected by sleep behaviors than careers that focus primarily on cognition. The fact that multiple correlations emerged between cumulative sleep behavior metrics and job performance, combined with the lack of such correlations from single-day metrics, suggests that sleep over an extended period has a stronger impact on a person's job performance than a single night of sleep. Additionally, the general increase in correlation coefficients after the sleep behavior metrics were normalized within users supports the notion that sleep needs and behaviors vary between individuals.

We further analyze the statistically significant correlations by measuring their effect sizes. One hour of sleep loss the night before by the average salesperson resulted in 1.9% decrease in the number of hires they were able to make. The average salesperson made 3.8 hires per workday and collected $936 in fees per hire. Therefore, a 1.9% decrease translates to a $67 loss per day. The average athlete experienced a 2.0% drop (1.3 points) in their game grade when they lost one hour of sleep the night before. Although these performance decreases may appear small, they can accumulate over time or across multiple people on the same team. In fact, sleep debt implies that a deficit can be spread over multiple days, so one hour of sleep loss the night before is equivalent to 2.4 hours of sleep loss a week before or 0.2 hours of sleep loss every day for a week. A more severe, but not uncommon scenario of losing an hour of sleep every day for a week is equivalent to losing 4.75 hours of sleep yesterday or 11.2 hours of sleep one week ago. On average, this loss in sleep debt causes a 9.5% (6.2 points) reduction in game performance, and a 9% ($317) reduction in hires for salespeople.

RQ.2: The Relationship Between App Interaction Time and Sleep Behavior

The analyses for this research question cover all 274 participants in our dataset. Our participants logged 7,195 nights of sleep that were paired with at least one app interaction event during the same day. Table 4 summarizes the correlation coefficients between our app-based performance metric (interaction time) and sleep behavior metrics for all 274 participants. Time-in-bed (ρ=-0.154, p<0.05), sleep history (ρ=-0.0549, p<0.05), and sleep debt (ρ=-0.0948, p<0.05) had negative correlations with interaction time; in other words, participants with better sleep behaviors had faster interaction times. As before, the cumulative sleep behavior metrics exhibited stronger correlations than total time-in-bed; however, app-based performance correlated better with non-normalized sleep behavior metrics. This result suggests that the minimal complexity of the app interaction task engendered less variance across individuals. Users who accumulated one less hour of sleep debt were 0.2 seconds faster each time they moved between app screens.

Since we found statistically significant correlations between the cumulative sleep behavior metrics and job performance, we used sleep history and sleep debt in our generalized additive model. Figure 5 shows the variation of app interaction time as a function of time of day, time since wake-up, and the aforementioned metrics. We find that interaction times are slowest at night hours, when users are most likely to be asleep, and fastest between 3-6 PM; the difference between those extremes is approximately 1.5 seconds. Note that the curves between interaction time against time of day (Figure 5, left column) generally align with circadian rhythm processes as measured through controlled sleep studies [2, 6, 16]. In the first hour of wake time, users tend to have slower interaction times. This result concurs with the chronobiological process of sleep inertia [2], which dictates that users are slower right after waking up. Interaction time decreases in the first six hours after wake-up and then begins to increase again, consistent with the chronobiological process of homeostatic sleep drive [6] as well as previous work examining click speeds in search engines [4]. On average, interaction time increases by 0.4 seconds when sleep history improves from 6 to 8 hours, and interaction time increases by 0.5 seconds beyond the threshold of -5 sleep debt hours.

RQ.3: The Relationship Between App Interaction Time and Job Performance

The analyses for this research question cover the 19 athletes and 15 salespeople who, respectively, interacted with their sleep-tracking app on 46 and 122 days with corresponding job performance metrics. Figure 6 shows real-world job performance against interaction time for those participants. Interaction time was not found to be significantly correlated with the number of hires the salespeople made (ρ=-0.0752, p=0.411). A significant correlation was found between interaction time and the athletes' game grade (ρ=-0.296, p=0.0455). The effect size shows that athletes who were 10 seconds faster in their interaction time had an average of 5 more points in game grades. Our app interaction metric is partly related to reaction time, so the discrepancy between athletes and salespeople in this analysis may be because the athletes' day-to-day activities require rapid, precise reactions; the salespeople's activities, on the other hand, are typically more forgiving in the sense of psychomotor function. Another explanation could be that PFF includes contextual information, such as whether the opponent presented a favorable matchup during a game; the number of hires a salesperson can make in a given day is more dependent upon external factors (e.g., customer needs, health of the economy).

RQ.4: The Relationship Between App Engagement and Sleep Behavior

Figure 7 shows the correlation and regression analysis between the number of app engagement days and sleep behavior metrics for the 274 participants in our dataset. The top row of the figure shows that participants who used the app more frequently slept longer, earlier, and accumulated less sleep debt on average. The bottom row of the figure shows that more frequent app usage also resulted in less variance in sleep behavior across all metrics, including more consistent bedtimes, wake-up times, and time-in-bed. All of these relationships except those related to average wake time, consistency of wake time, and sleep history were statistically significant.

Participants who engaged with the app for 35 days on average slept 42 minutes longer, had 3.5 hours less sleep debt, and went to bed 48 minutes earlier, compared to less engaged users who only engaged with the app for 7 days. Participants who engaged with the app more often also exhibited more consistent bedtimes, with the standard deviation dropping from 1.4 hours to 1.2 hours. We note that the wake-up time was more consistent across users than bedtime; this can be attributed to the fact that people typically need to wake up at a particular time for work, but there is no corresponding societal pressure for going to bed.

To analyze whether high app engagement is associated with *improvement* in sleep behaviors, we examine changes in sleep behavior within each user by comparing the average time-in-bed from the first D days to the last D days in a 5-week period (Figure 8). We find that higher app engagement was associated with increases in average time-in-bed between the first and last week of this five week period. This supports the assertion that the previously described effects in Figure 7 are not entirely explained by selection effects. Higher app engagement was consistently associated with an improvement in average time-in-bed, with an average increase of about

15 minutes per night. However, sleep improvement decreased as D increased. This phenomenon was expected since users likely saw significant improvement within the first few days of app engagement due to the novelty effect; as D increases, that short-term benefit becomes part of the baseline and makes the long-term benefit of app engagement less obvious.

DISCUSSION

Establishing the relationship between sleep behavior and job performance has been a challenge in the past due to the difficulty in collecting objective measures in real-world settings. By taking advantage of ubiquitous sleep-tracking technology and the increasing desire within companies to evaluate job performance through data, our research signifies a first step towards understanding this relationship. We also demonstrate that an app-based performance metric is correlated with sleep behaviors and time of day in a way that is consistent with sleep biology, and that it is correlated with job performance metrics as well. This highlights an interesting opportunity for future assessments of sleep and performance in uncontrolled settings. Finally, we demonstrate that increased engagement with automatic sleep-tracking technology is associated with improved sleep behavior. Below, we describe the implications and limitations of our work.

Opportunities for Passive Sensing

The PVT has been used to measure psychomotor and cognitive function in the wild [1]; however, the PVT can be disruptive if deployed at inopportune moments. Other prior work has required participants to adhere to a strict sleep schedule in order to measure the effects of sleep on behavior. In our work, we found that our instantiation of app-based performance (interaction time) was correlated with both better sleep behavior and athletic job performance, suggesting the potential power of a passive, nonintrusive performance indicator. Passive sensing additionally enables continuous collection of data in a wide variety of settings and contexts. Leveraging the use of ubiquitous technology to collect relevant benchmarks can also help produce large datasets or enable the study of populations that have traditionally been difficult to recruit to controlled studies.

We restricted our correlation analysis of app-based performance to specific, common, repeated, and comparable interactions within the sleep-tracking app that started from the same home screen and involved single touches; however, not all interactions are created equal, nor does the interaction time tell the whole story about how the user is engaging with the app's content. One may argue that some screens require more time to process than others, and longer processing times may indicate that the user is engaging more with the displayed information. Understanding the factors between app interaction time and on-screen content more broadly could be explored further to provide more frequent and representative measurements of interaction time. Comparable performance metrics to interaction time have also been elicited through other interactions, like typing and web browsing [4, 55, 58]. Responses to sleep alarms and notifications could also provide more natural opportunities for capturing app-based performance in the future.

Recommendations for Sleep-Tracking App Design

One design recommendation that we propose for sleep-tracking apps involves personalized views of sleep metrics. Many researchers have noted that sleep behaviors are unique according to genetic predisposition and chronotyping [3, 53]. Throughout our analyses, there were cases when normalizing sleep behavior metrics according to each user's history produced statistically significant correlations, but the same was not true for the raw data. Presenting raw values in combination with data that is scaled relative to the individual could provide useful insights to users in the future. Because sleep quality is subjective and not well-defined [27, 51], future apps could also allow users to explore what sleep metrics matter to their perceived sleep quality. In fact, we posit that job performance may be influenced by a person's perception of their own sleep quality, so our research may inform ways of exploring this matter in the future.

Finally, lapses in sleep tracking and the resulting lack of data are an important consequence of real-world data collection that should be addressed. Our dataset exhibited an extreme case of this issue since athletes can be away from home for at least 3-4 days at a time; nevertheless, travel is a regular occurrence for many people. The cumulative sleep metrics in our dataset, sleep history and sleep debt, were most informative in our analyses related to sleep behavior. We used the average time-in-bed of nearby nights for imputation when a participant skipped a night of sleep tracking. In future work, an alternative sensing approach (e.g., smartphone, smartwatch [35]) could be used for imputation. Finding ways of combining and resolving metrics across these different data sources could remedy data gaps.

Additional Context Information

With the exception of PFF's game grades for NFL athletes, all of our data streams lacked some amount of contextual information. PFF game grades incorporate context because they are assigned by experts who watch the games and understand the athletes' match. The performance of salespeople, on the other hand, depends on the demand of their goods and services and potentially other constraints; unfortunately, that information was unavailable to us. Job performance in a broader sense is also a function of experience and division of labor. Such information could be captured by worker profiles and more refined tracking in future work.

Sleep is known to be affected by a wide variety of factors: age [15, 21, 62], ambient light [37], caffeine intake [38], and diet [26], to name a few. The effect of travel between time zones (2–3 hour difference) has not been shown to significantly impact sleep [52], but an effect has been demonstrated on athletic performance [29]. Meanwhile, app usage can be affected by the user's interest in other apps on their smartphone and their overall workload. Measuring these factors through sensors and accounting for their effects in statistical analyses could improve evidence of links between sleep behavior, job performance, and app usage.

Limitations

Our dataset included participants from a bankruptcy law firm consultancy and the NFL, which allowed us to compare two populations with distinct job demands whose job performance can be quantified effectively. In both cases, we were able to identify sleep behavior metrics that correlated with job performance; however, the correlations manifested in different sleep behavior metrics (e.g., sleep debt and hiring rate for salespeople, personalized sleep history and game grades for athletes). Beyond the discrepancy between the two groups' job demands, the differences in results can also be attributed to idiosyncrasies within the job performance metrics themselves. For the salespeople, the number of hires an employee is able to make may depend on the state of the economy and the rate of bankruptcy in the country. For the athletes, the subjective nature of the expert's grades can manifest in anchoring effects towards common values [56]. We use rank-based correlation methods and per-person normalization to account for some of these idiosyncrasies, but future work should explore and compare alternative sources of job performance data. Furthermore, an exciting avenue of research may entail the creation of a job performance metric that generalizes across different careers.

Although salespeople and athletes have very different job demands, they do not cover the entire spectrum of careers. Each profession has its own demands and may not overlap with either of the ones that were included in our study. There was also an element of selection bias in our participant pool; the people who enrolled in our observational study may have been more excited to track their sleep and interact with the app than the average person, producing inflated app engagement measurements. Similarly, the observational and correlational nature of our data preclude us from making causal inferences. Learning about how our findings may generalize to other populations remains an area of future work.

CONCLUSION

Many people recognize that improving sleep behavior benefits job performance, but the precise relationship between the two has been difficult to capture and quantify in the past. Our study advances the literature in this space by providing a correlational analysis between objectively measured sleep behavior metrics from a mattress sensor and job performance metrics from a bankruptcy law firm and the NFL. Our findings suggest that establishing good sleep behaviors over extended periods is more important to job performance than simply getting a good night's sleep one day prior. We also found evidence that passively captured app interaction metrics can serve as a useful indicator for some job performance and sleep measures, thereby highlighting another mechanism through which researchers can collect relevant psychomotor and cognitive performance measures. Lastly, we found that increased engagement with automatic sleep-tracking technology is associated with improvements in sleep behavior over time. It is our hope that our work inspires researchers to examine *in-situ* sleep behaviors and performance measures across diverse contexts to further develop our understanding of human performance.

FUNDING

This research was supported by NSF grant IIS-1901386, an Adobe Data Science Research Award, and the Allen Institute for Artificial Intelligence.

Conflict of interest statement. Leon Sasson is the co-founder and CTO of Rise Science.

REFERENCES

[1] Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, Matthew Kay, Julie A Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone. In *Proc. UbiComp '16*. ACM Press, New York, New York, USA, 178–189. https://doi.org/10.1145/2971648.2971712

[2] Torbjörn Åkerstedt and Simon Folkard. 1997. The three-process model of alertness and its extension to performance, sleep latency, and sleep length. *Chronobiology International* 14, 2 (jan 1997), 115–123. https://doi.org/10.3109/07420529709001149

[3] Karla V Allebrandt, Maris Teder-Laving, Mahmut Akyol, Irene Pichler, Bertram Müller-Myhsok, Peter Pramstaller, Martha Merrow, Thomas Meitinger, Andreas Metspalu, and Till Roenneberg. 2010. CLOCK Gene Variants Associate with Sleep Duration in Two Independent Populations. *Biological Psychiatry* 67, 11 (jun 2010), 1040–1047. https://doi.org/10.1016/j.biopsych.2009.12.026

[4] Tim Althoff, Eric Horvitz, Ryen W White, and Jamie Zeitzer. 2017. Harnessing the Web for Population-Scale Physiological Sensing. In *Proc. WWW '17*. 113–122. https://doi.org/10.1145/3038912.3052637

[5] Joseph Baranski and Ross Pigeau. 1997. Self-monitoring cognitive performance during sleep deprivation: effects of modafinil, d-amphetamine and placebo. *Journal of Sleep Research* 6, 2 (jun 1997), 84–91. https://doi.org/10.1111/j.1365-2869.1997.00032.x

[6] Alexander A Borbély, Serge Daan, Anna Wirz-Justice, and Tom Deboer. 2016. The two-process model of sleep regulation: A reappraisal. *Journal of Sleep Research* 25, 2 (2016), 131–143. https://doi.org/10.1111/jsr.12371

[7] David H Brendel, CF Reynolds, JR Jennings, CC Hoch, TH Monk, SR Berman, F. T. Hall, D. J. Buysse, and D. J. Kupfer. 1990. Sleep Stage Physiology, Mood, and Vigilance Responses to Total Sleep Deprivation in Healthy 80-Year-Olds and 20-Year-Olds. *Psychophysiology* 27, 6 (nov 1990), 677–685. https://doi.org/10.1111/j.1469-8986.1990.tb03193.x

[8] Nikhil Byanna and Diego Klabjan. 2016. Evaluating the Performance of Offensive Linemen in the NFL. *arXiv preprint arXiv:1603.07593* (2016). https://arxiv.org/abs/1603.07593

[9] Ronald D Chervin. 2000. Sleepiness, fatigue, tiredness, and lack of energy in obstructive sleep apnea. *Chest* 118, 2 (2000), 372–379. https://doi.org/10.1378/chest.118.2.372

[10] Yong Won Cho, Jun Seok Lee, and Keun Tae Kim. 2019. 0168 Sleep and Academic Performance in Korean High School Students. *Sleep* 42, Supplement_1 (apr 2019), A69–A69. https://doi.org/10.1093/sleep/zsz067.167

[11] Eun Kyoung Choe, Sunny Consolvo, Nathaniel F Watson, and Julie A Kientz. 2011. Opportunities for computing technologies to support healthy sleep behaviors. In *Proc. CHI '11*. 3053–3062. https://doi.org/10.1145/1978942.1979395

[12] Ralph B D'Agostino. 1971. An omnibus test of normality for moderate and large size samples. *Biometrika* 58, 2 (1971), 341–348. https://academic.oup.com/biomet/article-abstract/58/2/341/263666

[13] Nediyana Daskalova, Bongshin Lee, Jeff Huang, Chester Ni, and Jessica Lundin. 2018. Investigating the Effectiveness of Cohort-Based Sleep Recommendations. *Proc. IMWUT '18* 2, 3 (2018), 1–19. https://doi.org/10.1145/3264911

[14] Nediyana Daskalova, Danaë Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. 2016. SleepCoacher: A Personalized Automated Self-Experimentation System for Sleep Recommendations. In *Proc. UIST '16*. 347–358. https:

//doi.org/10.1145/2984511.2984534

[15] Derk Jan Dijk and Jeanne F Duffy. 1999. Circadian regulation of human sleep and age-related changes in its timing, consolidation and EEG characteristics. , 130–140 pages. https://doi.org/10.3109/07853899908998789

[16] Derk Jan Dijk, Jeanne F Duffy, and Charles A Czeisler. 1992. Circadian and sleep/wake dependent aspects of subjective alertness and cognitive performance. *Journal of Sleep Research* 1, 2 (1992), 112–117. https://doi.org/10.1111/j.1365-2869.1992.tb00021.x

[17] David F Dinges, Frances Pack, Katherine Williams, Kelly A Gillen, John H Powell, Goeffrey E Ott, Caitlin Aptowicz, and Allen I Pack. 1997. *Cumulative Sleepiness, Mood Disturbance, and Psychomotor Vigilance Performance Decrements During a Week of Sleep Restricted to 4-5 Hours per Night.* Technical Report 4. 267–277 pages. https://academic.oup.com/sleep/article-abstract/20/4/267/2732104

[18] David F Dinges and John W Powell. 1985. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers* 17, 6 (1985), 652–655. https://doi.org/10.3758/BF03200977

[19] Christopher C Dodson, Eric S Secrist, Suneel B Bhat, Daniel P Woods, and Peter F Deluca. 2016. Anterior Cruciate Ligament Injuries in National Football League Athletes From 2010 to 2013: A Descriptive Epidemiology Study. *Orthopaedic Journal of Sports Medicine* 4, 3 (mar 2016). https://doi.org/10.1177/2325967116631949

[20] Yuriko Doi, Masumi Minowa, and Toshiro Tango. 2003. Impact and correlates of poor sleep quality in Japanese white-collar employees. , 467–471 pages. https://doi.org/10.1093/sleep/26.4.467

[21] Irwin Feinberg. 1974. Changes in sleep cycle patterns with age. *Journal of Psychiatric Research* 10, 3-4 (1974), 283–306. https://doi.org/10.1016/0022-3956(74)90011-9

[22] Namni Goel, Mathias Basner, Hengyi Rao, and David F Dinges. 2013. Circadian rhythms, sleep deprivation, and human performance. In *Progress in Molecular Biology and Translational Science*. Elsevier, 155–190. https://doi.org/10.1016/B978-0-12-396971-2.00007-5

[23] Scott A Golder and Michael W Macy. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333, 6051 (2011), 1878–1881. https://doi.org/10.1126/science.1202775

[24] Mitchell L Gordon, Leon Gatys, Carlos Guestrin, Jeffrey P Bigham, Andrew Trister, and Kayur Patel. 2019. App Usage Predicts Cognitive Ability in Older Adults. In *Proc. CHI '19*. 168. https://doi.org/10.1145/3290605.3300398

[25] G. Guerrero-Mora, Palacios Elvia, A. M. Bianchi, J. Kortelainen, M. Tenhunen, S. L. Himanen, M. O. Mendez, E. Arce-Santana, and O. Gutierrez-Navarro. 2012. Sleep-wake detection based on respiratory signal acquired through a Pressure Bed Sensor. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. 3452–3455. https://doi.org/10.1109/EMBC.2012.6346708

[26] Shona L Halson. 2008. Nutrition, sleep and recovery. *European Journal of Sport Science* 8, 2 (mar 2008), 119–126. https://doi.org/10.1080/17461390801954794

[27] Allison G Harvey, Kathleen Stinson, Katriina L Whitaker, Damian Moskovitz, and Harvinder Virk. 2008. The Subjective Meaning of Sleep Quality: A Comparison of Individuals with and without Insomnia. *Sleep* 31, 3 (mar 2008), 383–393. https://doi.org/10.1093/sleep/31.3.383

[28] Jim Horne. 2004. Is there a sleep debt? *Sleep* 27, 6 (sep 2004), 1047–1049. http://www.ncbi.nlm.nih.gov/pubmed/15532195

[29] Richard Jehue, David Street, and Robert Huizenga. 1993. Effect of time zone and game time changes on team performance: National Football League. *Medicine and Science in Sports and Exercise* 25, 1 (jan 1993), 127–131. https://doi.org/10.1249/00005768-199301000-00017

[30] Megan E Jewett, Derk Jan Dijk, Richard E Kronauer, and David F Dinges. 1999. Dose-response relationship between sleep duration and human psychomotor vigilance and subjective alertness. *Sleep* 22, 2 (1999), 171–179. https://doi.org/10.1093/sleep/22.2.171

[31] Ronald C Kessler, Patricia A Berglund, Catherine Coulouvrat, Goeran Hajak, Thomas Roth, Victoria Shahly, Alicia C Shillington, Judith J Stephenson, and James K Walsh. 2011. Insomnia and the Performance of US Workers: Results from the America Insomnia Survey. *Sleep* 34, 9 (sep 2011), 1161–1171. https://doi.org/10.5665/sleep.1230

[32] William DS Killgore, Thomas J Balkin, and Nancy J Wesensten. 2006. Impaired decision making following 49 h of sleep deprivation. *Journal of Sleep Research* 15, 1 (mar 2006), 7–13. https://doi.org/10.1111/j.1365-2869.2006.00487.x

[33] William DS Killgore, Ellen T Kahn-Greene, Erica L Lipizzi, Rachel A Newman, Gary H Kamimori, and Thomas J Balkin. 2008. Sleep deprivation reduces perceived emotional intelligence and constructive thinking skills. *Sleep Medicine* 9, 5 (jul 2008), 517–526. https://doi.org/10.1016/j.sleep.2007.07.003

[34] Shingo Kitamura, Yasuko Katayose, Kyoko Nakazaki, Yuki Motomura, Kentaro Oba, Ruri Katsunuma, Yuri Terasawa, Minori Enomoto, Yoshiya Moriguchi, Akiko Hida, and Kazuo Mishima. 2016. Estimating individual optimal sleep duration and potential sleep debt. *Scientific Reports* 6, 1 (dec 2016), 35812. https://doi.org/10.1038/srep35812

[35] 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* 11, 12 (2015), 1455–1461. https://doi.org/10.5664/jcsm.5288

[36] Juha M. Kortelainen, Martin O. Mendez, Anna Maria Bianchi, Matteo Matteucci, and Sergio Cerutti. 2010. Sleep staging based on signals acquired through bed sensor. *IEEE Transactions on Information Technology in Biomedicine* 14, 3 (may 2010), 776–785. https://doi.org/10.1109/TITB.2010.2044797

[37] Tomoaki Kozaki, Shingo Kitamura, Yuichi Higashihara, Keita Ishibashi, Hiroki Noguchi, and Akira Yasukouchi. 2005. Effect of Color Temperature of Light Sources on Slow-wave Sleep. *Journal of Physiological Anthropology and Applied Human Science* 24, 2 (2005), 183–186. https://doi.org/10.2114/jpa.24.183

[38] Hans Peter Landolt, Derk-Jan Dijk, Stephanie E Gaus, and Alexander A Borbély. 1995. Caffeine reduces low-frequency delta activity in the human sleep EEG. *Neuropsychopharmacology* 12, 3 (1995), 229–238. https://www.sciencedirect.com/science/article/pii/0893133X9400079F

[39] Zilu Liang and Bernd Ploderer. 2016. Sleep tracking in the real world: A qualitative study into barriers for improving sleep. In *Proc. OzCHI '16*. 537–541. https://doi.org/10.1145/3010915.3010988

[40] June C Lo, Ju Lynn Ong, Ruth LF Leong, Joshua J Gooley, and Michael WL Chee. 2016. Cognitive Performance, Sleepiness, and Mood in Partially Sleep Deprived Adolescents: The Need for Sleep Study. *Sleep* 39, 3 (2016), 687–698. https://doi.org/10.5665/sleep.5552

[41] Cheri D Mah, Kenneth E Mah, Eric J Kezirian, and William C Dement. 2011. The Effects of Sleep Extension on the Athletic Performance of Collegiate Basketball Players. *Sleep* 34, 7 (jun 2011), 943–950. https://doi.org/10.5665/sleep.1132

[42] Robert L Matchock and J Toby Mordkoff. 2009. Chronotype and time-of-day influences on the alerting, orienting, and executive components of attention. *Experimental brain research* 192, 2 (jan 2009), 189–98. https://doi.org/10.1007/s00221-008-1567-6

[43] Elizabeth L Murnane, Saeed Abdullah, Mark Matthews, Matthew Kay, Julie A Kientz, Tanzeem Choudhury, Geri Gay, and Dan Cosley. 2016. Mobile manifestations of alertness. In *Proc. MobileHCI '16*. ACM Press, New York, New York, USA, 465–477. https://doi.org/10.1145/2935334.2935383

[44] June J Pilcher and Allen I Huffcutt. 1996. Effects of sleep deprivation on performance. *Sleep* 19, 4 (1996), 318–326. https://doi.org/10.2466/pr0.1975.37.2.479

[45] Pro Football Focus. 2017. How We Grade. , 4 pages. https://web.archive.org/web/20170526235935/https://www.profootballfocus.com/about/how-we-grade/

[46] Matthew T Provencher, James P Bradley, Jorge Chahla, Anthony Sanchez, Brendin R. Beaulieu-Jones, Justin W. Arner, Nicholas I. Kennedy, George Sanchez, Mitchell I. Kennedy, Gilbert Moatshe, Mark E. Cinque, and Robert F. LaPrade. 2018. A History of Anterior Cruciate Ligament Reconstruction at the National Football League Combine Results in Inferior Early National Football League Career Participation. *Arthroscopy - Journal of Arthroscopic and Related Surgery* 34, 8 (2018), 2446–2453. https://doi.org/10.1016/j.arthro.2018.03.018

[47] Pooja Rajdev, David Thorsley, Srinivasan Rajaraman, Tracy L Rupp, Nancy J Wesensten, Thomas J Balkin, and Jaques Riefman. 2013. A unified mathematical model to quantify performance impairment for both chronic sleep restriction and total sleep deprivation. *Journal of theoretical biology* 331 (2013), 66–77. https://www.sciencedirect.com/science/article/pii/S0022519313001811

[48] Sridhar Ramakrishnan, Srinivas Laxminarayan, David Thorsley, Nancy J Wesensten, Thomas J Balkin, and Jaques Reifman. 2012. Individualized performance prediction during total sleep deprivation: Accounting for trait vulnerability to sleep loss. In *Proc. EMBS '12*. 5574–5577. https://doi.org/10.1109/EMBC.2012.6347257

[49] Sridhar Ramakrishnan, Nancy J Wesensten, Thomas J Balkin, and Jaques Reifman. 2016. A Unified Model of Performance: Validation of its Predictions across Different Sleep/Wake Schedules. *Sleep* 39, 1 (2016), 249–262. https://doi.org/10.5665/sleep.5358

[50] Jukka Ranta, Timo Aittokoski, Mirja Tenhunen, and Mikko Alasaukko-Oja. 2019. EMFIT QS heart rate and respiration rate validation. *Biomedical Physics and Engineering Express* 5, 2 (2019), 25016. https://doi.org/10.1088/2057-1976/aafbc8

[51] Ruth Ravichandran, Sang Wha Sien, Shwetak N Patel, Julie A Kientz, and Laura R Pina. 2017. Making sense of sleep sensors: How sleep sensing technologies support and undermine sleep health. In *Proc. CHI '17*. Association for Computing Machinery, 6864–6875. https://doi.org/10.1145/3025453.3025557

[52] Louise K Richmond, Brian Dawson, Glenn Stewart, Stuart Cormack, David R Hillman, and Peter R Eastwood. 2007. The effect of interstate travel on the sleep patterns and performance of elite Australian Rules footballers. *Journal of Science and Medicine in Sport* 10, 4 (jun 2007), 252–258. https://doi.org/10.1016/j.jsams.2007.03.002

[53] Till Roenneberg, Anna Wirz-Justice, and Martha Merrow. 2003. Life between clocks: Daily temporal patterns of human chronotypes. *Journal of Biological Rhythms* 18, 1 (feb 2003), 80–90. https://doi.org/10.1177/0748730402239679

[54] Daniel A Sternberg, Kacey Ballard, Joseph L Hardy, Benjamin Katz, P Murali Doraiswamy, and Michael Scanlon. 2013. The largest human cognitive performance dataset reveals insights into the effects of lifestyle factors and aging. *Frontiers in Human Neuroscience* 7 (2013), 292. https://doi.org/10.3389/fnhum.2013.00292

[55] Martin Thirkettle, Jennifer Lewis, Darren Langdridge, and Graham Pike. 2018. A Mobile App Delivering a Gamified Battery of Cognitive Tests Designed for Repeated Play (OU Brainwave): App Design and Cohort Study. *JMIR Serious Games* 6, 4 (2018), e10519. https://doi.org/10.2196/10519

[56] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1017/cbo9780511809477.002

[57] Hans PA Van Dongen, Naomi L Rogers, and David F Dinges. 2003. Sleep debt: Theoretical and empirical issues. https://doi.org/10.1046/j.1446-9235.2003.00006.x

[58] Lisa M Vizer, Lina Zhou, and Andrew Sears. 2009. Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human Computer Studies* 67, 10 (oct 2009), 870–886. https://doi.org/10.1016/j.ijhcs.2009.07.005

[59] Matthew P Walker and Robert Stickgold. 2005. Sleep, Memory, and Plasticity. *Annual Review of Psychology* 57, 1 (jan 2005), 139–166. https://doi.org/10.1146/annurev.psych.56.091103.070307

[60] Andrew M Watson. 2017. Sleep and Athletic Performance. *Current Sports Medicine Reports* 16, 6 (2017), 413–418. https://doi.org/10.1249/JSR.0000000000000418

[61] Ann M Williamson and Anne-Marie Feyer. 2000. Moderate sleep deprivation produces impairments in cognitive and motor performance equivalent to legally prescribed levels of alcohol intoxication. *Occupational and Environmental Medicine* 57, 10 (2000), 649–655. https://oem.bmj.com/content/57/10/649.short

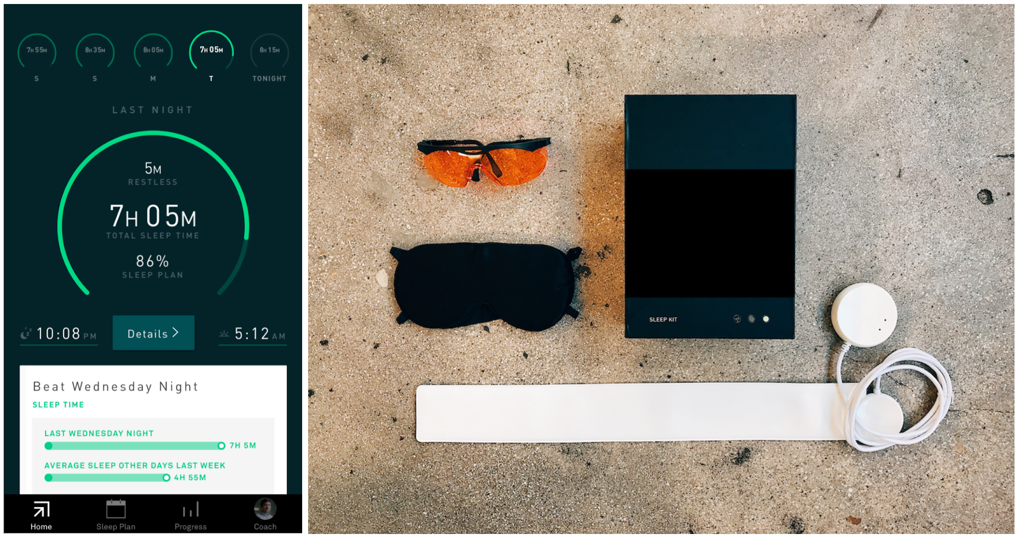
[62] In Young Yoon, Daniel F Kripke, Jeffrey A Elliott, Shawn D Youngstedt, Katharine M Rex, and Richard L Hauger. 2003. Age-related changes of circadian rhythms and sleep-wake cycles. *Journal of the American Geriatrics Society* 51, 8 (aug 2003), 1085–1091. https://doi.org/10.1046/j.1532-5415.2003.51356.x

LIST OF FIGURES

- Figure 1: (left) The home screen of a sleep-tracking app. (right) The kit that participants received upon enrolling in the study, including: an Emfit QS, a sleep-tracking mobile app, a blindfold, and orange-tinted glasses.
- Figure 2: A diagram that summarizes the research questions explored in this work.
- Figure 3: Regression plots showing the effect sizes for the statistically significant results from Table 3. For all similar figures in this paper, data is binned into discrete intervals, with the estimated mean and standard error shown in blue. Intervals are chosen such that samples are evenly distributed throughout the bins. Orange lines represent the best linear regression fit to the *raw data* (not the interval aggregates) along with 95% confidence intervals.
- Figure 4: Regression plots showing the effect sizes for the statistically significant results from Table 4.
- Figure 6: Regression plots showing the effect sizes between job and app-based performance.
- Figure 6: Regression plots showing the effect sizes between job and app-based performance.
- Figure 7: Regression plots showing the effect sizes between days of app engagement and (top row) sleep behavior metrics and (bottom row) variation in sleep behaviors.
- Figure 8: Sleep improvement over a five-week period vs. app engagement (excluding first and last D days). N indicates the number of participants who tracked their sleep for the first and last D days during the five-week period. Standard errors are shown.

LIST OF TABLES

- Table 1: A summary of the metrics we collect in our dataset through three data streams: (1) sleep metrics through the Emfit QS, (2) job performance through the participants' employers, and (3) app usage through a sleep tracking mobile app.
- Table 2: Summary statistics for data from different participant cohorts after filtering (see section on Data Filtering and Post-Processing).
- Table 3: Spearman correlation coefficients between sleep behavior and job performance. P-values are provided in parentheses; results with p-value < 0.05 are shown in bold.
- Table 4: Spearman correlation coefficients between sleep behavior and app-based performance. P-values are provided in parentheses; results with p-value < 0.05 are shown in bold.



(a) (b)

Fig. 1. (left) The home screen of a sleep-tracking app. (right) The kit that participants received upon enrolling in the study, including: an Emfit QS, a sleep-tracking mobile app, a blindfold, and orange-tinted glasses.

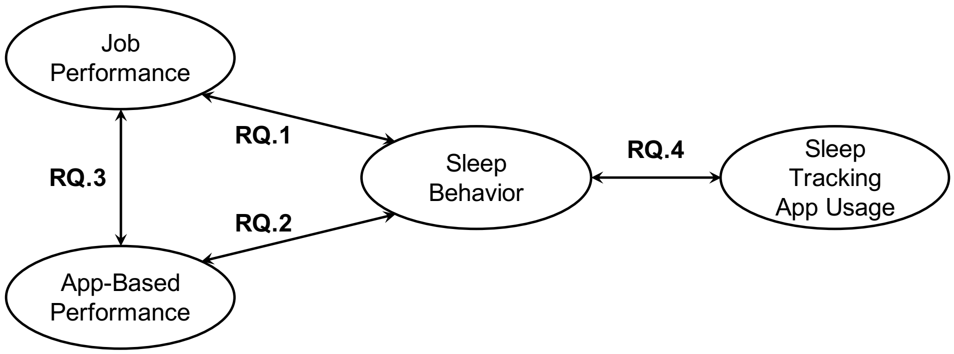


Fig. 2. A diagram that summarizes the research questions explored in this work.

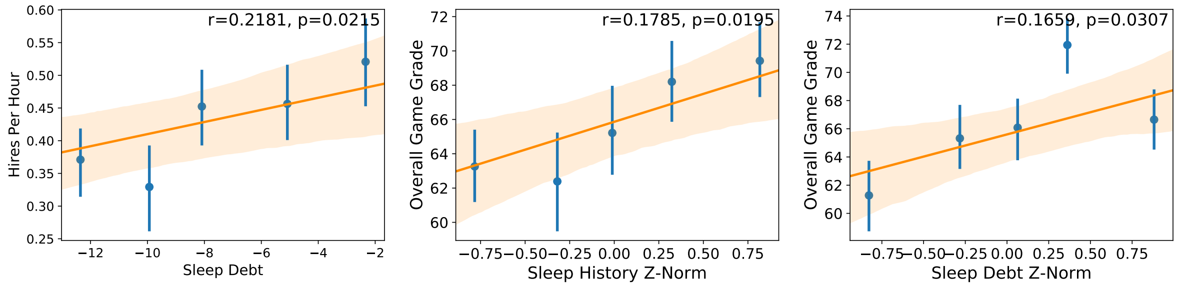


Fig. 3. Regression plots showing the effect sizes for the statistically significant results from Table 3. For all similar figures in this paper, data is binned into discrete intervals, with the estimated mean and standard error shown in blue. Intervals are chosen such that samples are evenly distributed throughout the bins. Orange lines represent the best linear regression fit to the *raw data* (not the interval aggregates) along with 95% confidence intervals.

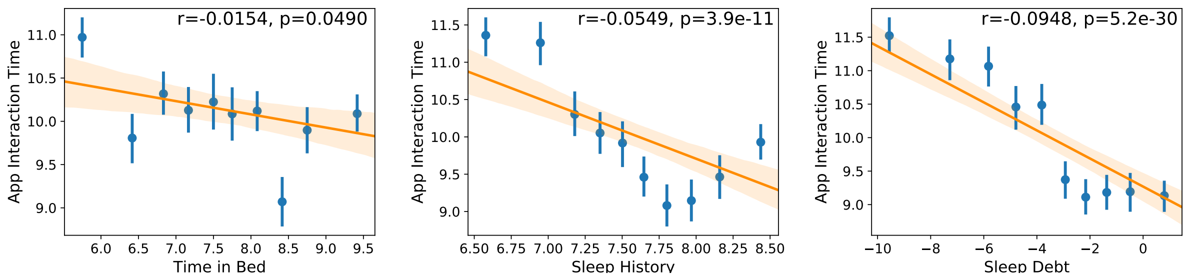


Fig. 4. Regression plots showing the effect sizes for the statistically significant results from Table 4.

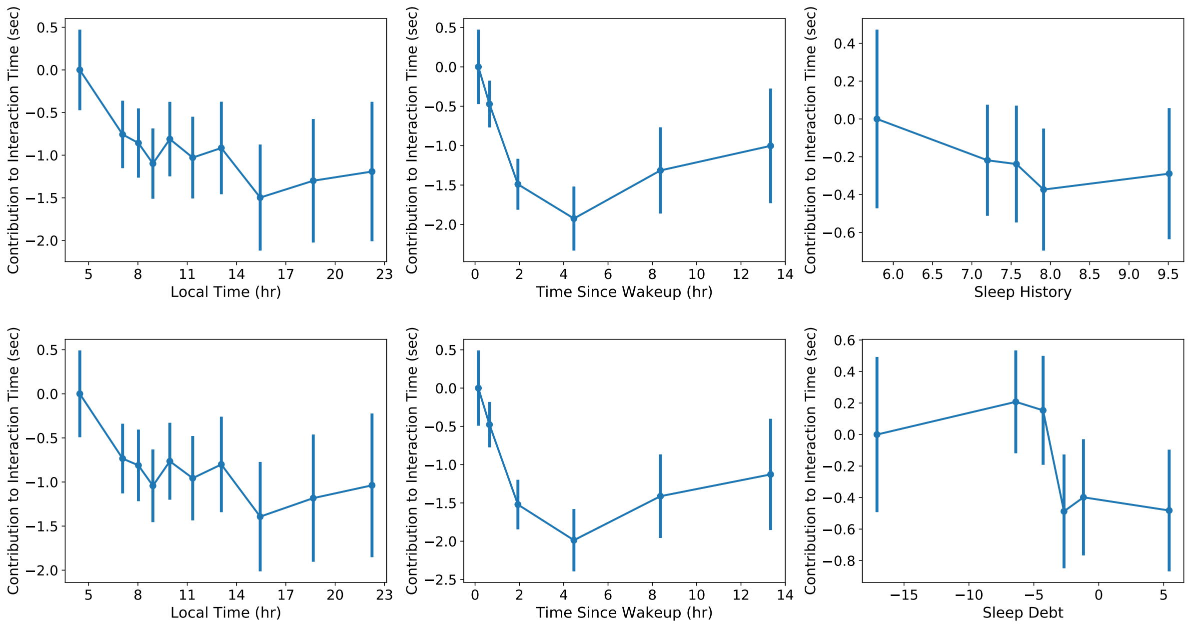


Fig. 5. (Top row) The generalized additive model of interaction time against the local time in the user's time zone, time since wake-up, and sleep history. (Bottom row) A similar generalized additive model using sleep debt instead of sleep history. Standard errors are shown in both cases. Both models included random intercepts for each user.

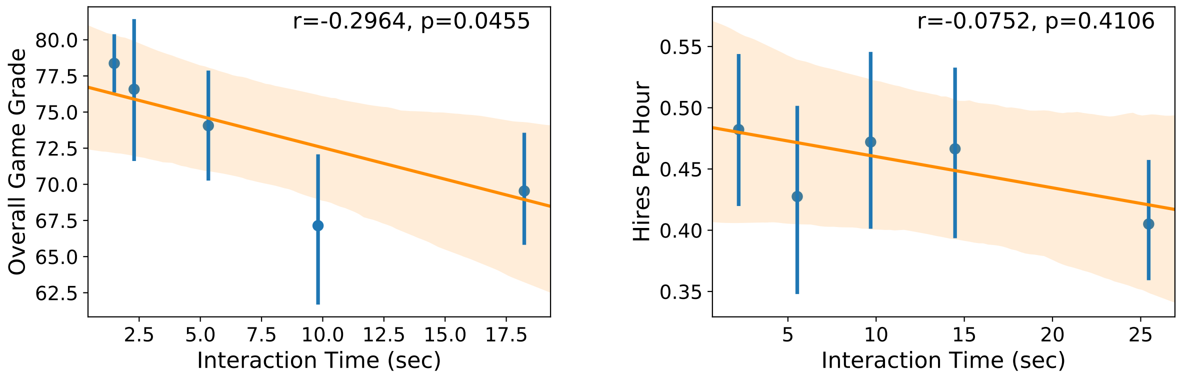


Fig. 6. Regression plots showing the effect sizes between job and app-based performance.

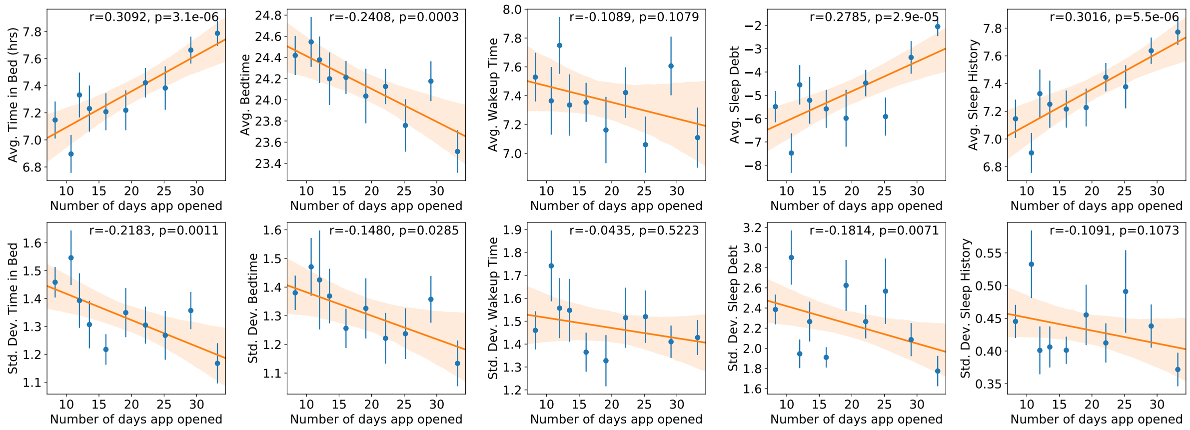


Fig. 7. Regression plots showing the effect sizes between days of app engagement and (top row) sleep behavior metrics and (bottom row) variation in sleep behaviors.

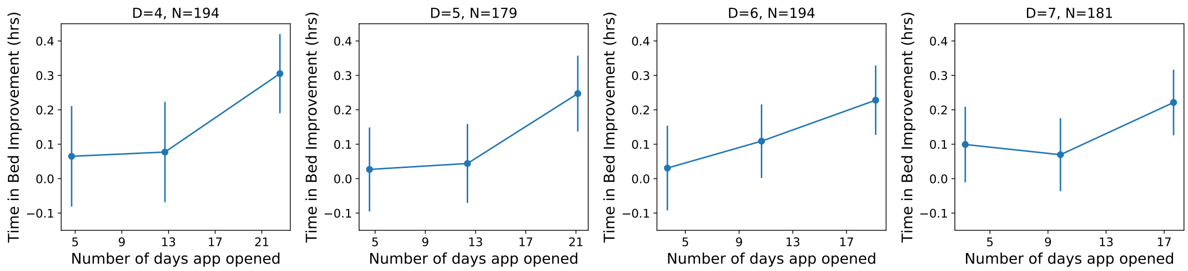


Fig. 8. Sleep improvement over a five-week period vs. app engagement (excluding first and last D days). N indicates the number of participants who tracked their sleep for the first and last D days during the five-week period. Standard errors are shown.

	Metric	**Description**
Sleep	Bedtime	Time at which the user got into their bed
	Wake time	Time at which the user got out of their bed
	Midpoint	Midpoint between start and end time
	Time-in-bed	The total time the user spent in bed during a single day including nighttime sleep and naps, regardless of whether they were sleeping
	Sleep debt	Weighted average of difference between sleep need and time-in-bed
	Sleep history	Weighted average of time-in-bed
Job Performance	Number of hires (salespeople)	Number of contracts made after consulting, normalized by the number of hours they work
	Game grade (athletes)	Score of a player's game performance out of 100 assigned through three independent experts
App Usage	Interaction time	Time between opening home screen of app to another screen by user's touch input
	Days of app engagement	Number of days the user opened the app

Table 1. A summary of the metrics we collect in our dataset through three data streams: (1) sleep metrics through the Emfit QS, (2) job performance through the participants' employers, and (3) app usage through a sleep tracking mobile app.

Statistic	All Users	Salespeople	Athletes
Number of participants	274	15	19
Total unique days with both sleep-tracking and job performance measurements	289	118	171
Total unique days with both app interaction and job performance measurements	168	122	46
Total nights of sleep tracked with app-based performance measure	7,195	234	418
Total nights of sleep tracked	30,618	834	2,687
Total number of transitions between screens	16,336	679	909
Total number of times app was opened	11,140	425	691
Nights of sleep tracked per user (avg ± std)	109.2 ± 91.81	46.33 ± 37.45	133.1 ± 89.92
Time-in-bed in hours (avg ± std)	7.338 ± 1.628	7.283 ± 2.020	7.308 ± 1.920
Days of app use per user (avg ± std)	43.68 ± 46.48	28.25 ± 21.06	40.65 ± 50.43

Table 2. Summary statistics for data from different participant cohorts after filtering (see section on Data Filtering and Post-Processing).

		Sleep Metrics					
		Raw Metrics			**Per-Person Z-Normalized Metrics**		
		Time-in-Bed	Sleep Debt	Sleep History	Time-in-Bed	Sleep Debt	Sleep History
Job Performance Metrics	NFL Player Game Grades (N = 19)	-0.024 (p=0.751)	-0.095 (p=0.218)	-0.029 (p=0.711)	0.086 (p=0.263)	**0.166 (p=0.031)**	**0.179 (p=0.020)**
	Salespeople Hires per Day (N = 15)	-0.067 (p=0.469)	**0.218 (p=0.022)**	0.039 (p=0.690)	-0.102 (p=0.283)	0.164 (p=0.088)	-0.047 (p=0.634)

Table 3. Spearman correlation coefficients between sleep behavior and job performance. P-values are provided in parentheses; results with p-value < 0.05 are shown in bold.

	Raw Sleep Data			**Per Person Z-Normalization of Sleep Data**		
	Time-in-Bed	Sleep History	Sleep Debt	Time-in-Bed	Sleep History	Sleep Debt
Interaction Time	**-0.015** **(p=0.049)**	**-0.055** **(p=3.9×10^{-11})**	**-0.095** **(p=5.2×10^{-30})**	0.006 (p=0.483)	-0.012 (p=0.140)	-0.010 (p=0.230)

Table 4. Spearman correlation coefficients between sleep behavior and app-based performance. P-values are provided in parentheses; results with p-value < 0.05 are shown in bold.