Robust Recognition of Sleep Behavior Using Wearable Sensors

Master's Thesis submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Informatics

presented by
Lidia Alecci

under the supervision of
Prof. Silvia Santini
co-supervised by
Prof. Francesca Gasparini, PhD candidate Shkurta Gashi

July 2021

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Lidia Alecci
Lugano, 27 July 2021

Abstract

Mobile and wearable devices are widely used for tracking and monitoring their users' health. Being non-invasive, on-body tools, they significantly ease the collection of sensor data, which can in turn be used to support applications that monitor health status and help prevent diseases, e.g., by sending an alarm when abnormal vital signs are detected. Recent studies also show that physiological signals collected using wearable devices can be used to infer a variety of human behaviours, emotions, and psychological states, including stress, engagement and academic performance.

The main goal of this thesis is to design and develop an automatic approach to infer sleep duration and quality from physiological data. We focus in particular on the use of electrodermal activity, skin temperature and accelerometer data collected using wrist-worn devices.

Despite the recent advancements in automatic sleep detection using physiological signals, detecting sleep in real-world settings is still very challenging. While there are several reasons for this, detection accuracy and robustness is significantly hampered by the presence of noise and artifacts present in physiological signals, which affect the quality of the collected data and hence the quality of the final results. To cope with this problem, in this thesis, we focus on developing robust models able to operate on physiological signals affected by noise and artifacts.

To achieve the goal of the thesis, we first collected a novel and rich sensor data set during a 1-month long ambulatory study. We collected behavioural data (e.g., phone screen on/off, notifications, amount of environment light, phone screen proximity and background application usage of user's phone) using an Android application installed on user's smartphone, physiological signals (e.g., electrodermal activity, skin temperature, acceleration, blood volume pulse) using empatica e4 wristbands, and self-reports (i.e., about the sleep and wake up time of the user as well as sleep quality) using pen-and-paper diary, smartphone and Google forms from the laptop. The dataset contains data from 16 participants for 30 days.

We implemented dedicated tools to monitor the quality and quantity of the collected data, which enable intervening in case of problems with the data collection – e.g., malfunctioning devices – were detected. We further designed and developed a dashboard to allow explorations and visualizations of the data, both for individual participants and in aggregated forms. After a data cleaning phase, in which we discarded data records with missing answers to the self-reports, we obtained a final dataset of 6557 hours of data in total.

We then designed and developed a machine learning pipeline to detect whether a user is sleeping or is awake – i.e., to discriminate between what we refer to as the "sleep" and "awake" states – as well as to infer sleep quality. The pipeline consists of six steps as follows: data imputation, preprocessing, segmentation, feature extraction, feature interpretation and classification.

Several existing studies demonstrate the relationship between electrodermal activity signal characteristics – such as e.g., peak epochs and storms – and sleep quality. In particular, electrodermal activity peaks during the first quarter of the night are associated to a greater subjective sleep quality. For this reason, part of this work focuses on implementation of a rule-based approach to detect and label peak epochs and storms.

We model the problem of sleep and awake as a binary classification task and the sleep quality as a binary (*high* and *low*), three-class (*high*, *normal*, *low*) and five-class (*very good*, *good*, *normal*, *poor*, *very poor*) classification tasks. Each model was evaluated by using user-independent and user-dependent validation techniques. The purpose of using user independent (i.e., leave-one-subject-out (LOSO)) approach was to understand the performance and generizaiblity of the model to new users. In contrast, user-dependent validation technique was trained with data of the test user by using only his past days, and to have enough data we used as test only sessions from the fourth session on.

To understand the contribution of individual features for the classification result, we used SHAP (SHapley Additive exPlanations). We further investigated different segmentation windows (1, 5, and 10 minutes) and found that a 10-minute window gives the highest results for sleep detection for the majority of the classification tasks. To understand the impact of each sensor and their features in the overall classification performance, we also explored the impact of different sensors and features alone and combined (e.g., electrodermal activity, temperature, accelerometer, electrodermal activity with storms features, temperature and accelerometer).

Our results indicate that is feasible to estimate sleep quality and duration with a reasonable degree of accuracy, even if quality based on more than two different classes still remains a difficult problem. In particular using electrodermal activity, accelerometer and skin temperature sensors a segmentation of 10 minutes, we are able to achieve an accuracy above 90% for distinguishing sleep and awake, with a user-independent model, which is 40 percentage points increment from the most frequent baseline classifier. Our results show that our model can distinguish between the high and low sleep quality using user-dependent model and a segmentation of 1 minute with a balanced accuracy of 63%, which is 13 percentage points higher than the performance of baseline classifiers. These results are achieved using only the features of the skin temperature sensor.

Overall, the results of this thesis show that even by using electrodermal activity, skin temperature and acceleration data only, a sleep detection system can achieve a reasonable level of accuracy. This opens up opportunities to develop personal informatics systems for encouraging healthy sleep routines, to prevent and treat sleep disorders, and to investigate how sleep is correlated to other diseases.

Contents

Figures

Tables

Chapter 1

Introduction

The past decades have seen an increased interest in wearable technology and, as long as the producers keeping pace with releasing new sensors and improving existing ones, this market will only continue to expand. The GlobalData Report of 2020 estimated that this sector will be worth $54 billion by 2023 [GlobalData, 2019].

This growth enables the possibility of investigating and tracking human behavior with the aim of supporting humans in daily life.

Wearable applications are many, starting with the fitness tracker [Imani et al., 2016] for mobile health management [Dunn et al., 2018], passing through a simple extension of smartphones (read the notifications on the SmartWatch) or education purpose [Sparacino, 2021; Tanenbaum et al., 2015].

For instance, Wang et al. [2017] used ubiquitous computing to assess academic performance, mental well-being and behavioral trends of students using a smartphone sensing app. Instead, Sano et al. [2015] used wearable sensors and mobile phones to recognize and found relation between: self-reported sleep quality, mental health condition, stress and academic performance. The reason why they choose to use also wearable is that physiological responses are strongly related to some emotions (e.g. anger, anxiety, fear, sadness, joy, surprise) [Kreibig, 2010], since physiological parameters are physical manifestations of autonomic nervous system (ANS) responses. For example, heart rate increases when a subject feels angry and heart rate variability decreases when a subject is anxious.

Several researchers have shown that sleep quality has an impact on both, physical and mental health. For instance, having insufficient amount of sleep would affect people's focus, concentration, and memory during the day [Lim and Dinges, 2008]. Additionally, the physical exhaustion results in headache and dizziness. The key fact is that bad sleep behavior does not only affect humans in the short term, but also in long-term [Knutson et al., 2007; Zhao et al., 2013; Miller and Cappuccio, 2007]. Indeed, clinical studies show that there is a strong correlation between bad sleeping habits and health issues as: obesity, diabetes [Knutson et al., 2007; Buxton and Marcelli, 2010], weakened immune system [Miller and Cappuccio, 2007], Parkinson [Kay et al., 2018] and cancer [Zhao et al., 2013].

Furthermore, sleep duration and regularity have been linked to depression [Wang et al., 2017], stress [Wang et al., 2017],academic grades [Phillips et al., 2017] and personality [Soehner et al., 2007]. In particular, Phillips et al. [2017] and Sano et al. [2015] found a positive correlation between sleep regularity and academic grades and Wang et al. [2017] showed that students who sleep less are more likely to experience depressive symptoms than students that sleep more.

Soehner et al. [2007] observed that higher neuroticism was associated with poorer sleep and some personality aspects have an impact on timing or sleep quality, even if no significant correlation was found between personality and sleep duration in general. Furthermore, according to Baran and Chervin [2009], almost ∼30% of adults suffer from sleep disorders, and most of them are not diagnosed, so having non-intrusive ways to diagnose these problems can improve personal well-being.

A recent study, conducted by Philips [Philips, 2021] during the COVID-19 pandemic, has shown that people's sleep routine has changed during this time and their sleep quality has worsened. This implies that monitoring sleep is crucial now more than ever to avoid the long-term issues that wrong sleep behaviour can lead to.

Understanding if a user is asleep or awake has significant implications: it may allow to guide sleep or wake related behavioral changes and recommendations to promote users' well-being; e.g. give suggestion on the right time to go to sleep and wake up (improving sleep quality by achieving sleep regularity[Soehner et al., 2011]), or by automatically turning off notifications during sleep.

The goal of this thesis is to develop an automatic approach to infer sleep and awake segments and sleep quality from physiological data. We focus in particular on the use of electrodermal activity, skin temperature and accelerometer data collected using wrist-worn devices.

While automatic detection of sleep quality has been studied extensively [Sano and Picard, 2012; Picard Rosalind W.], only a few studies explored automatic ways of segmenting physiological signals to detect sleep or awake segments [Zhai et al., 2020; Sano and Picard, 2012]. For instance, Sano and Picard [2014] used wearable sensors to infer sleep/awake, and by using electrodermal activity, accelerometer and skin temperature they achieve an accuracy of 85%.

The focus of this thesis is on how to detect robustly sleep quality and sleep/awake segments. To our knowledge, no prior studies have attempted to introduce artifact, peak epoch and storms in a machine learning pipeline in order to predict sleep/awake segments and sleep quality. Despite literature has shown correlation between physiological state (e.g. sleep/awake) and physiological responses. Moreover, several studies underline how the variability between users impacts overall performance. This convinced us to compare performance of a user-independent model, a general one that try to infer sleep/wake segments and sleep quality by using data from all users, and a user-dependent model that use only past data of the same user to infer his future instances.

Even if people all over the world use commercial devices to track their activities and also sleep, those methods are still prone to errors. As Stone et al. [2020] pointed out most of the existing commercial wearable devices that use physiological signals either over or underestimated sleep metrics (e.g., sleep efficiency and total sleep or wake time). Furthermore proprietary algorithm (e.g. Fitbit) [Menghini et al., 2021] make it difficult to understand which features they used and what these errors depend on.

The contributions of the thesis can be summarized as follows:

- An in-depth review of existing literature about sleep/wake and sleep quality detection, using mobile and wearable devices as well as traditional approaches.

- A data collection with 16 participants for one months to collect sensor data from mobile and wearable devices and ground-truth data about participants sleeping hours and quality.

- Dedicated tools to monitor the quality and quantity of the collected data.
- Cleaning data to discard data records with missing or incomplete self-reports.
- A dashboard to summarize the amount of collected sensor and ground-truth data.
- A rule-based approach to detect electrodermal activity storms and epochs.
- A machine learning pipeline to detect sleep/wake and sleep quality from wearable sensors using electrodermal activity, skin temperature e acceleration data collected with wristbands.

1.1 Overview

The structure of the thesis is as listed below:

Chapter 2: Related Work – This chapter presents an overview of the existent devices or techniques that are currently used to infer sleep is presented in this chapter.

Chapter 3: Background – This section clarifies the the main concepts used throughout the thesis that are necessary for a full understanding of how sleep works and the physiological signals connected to this state.

Chapter 4: Data Collection – In this chapter the steps took to run the data collection are described.

Chapter 5: Data Visualization – This chapter provides data visualizations to better understand the quantity and quality of the collected data as well as the patterns in collected data for sleep stages and sleep quality.

Chapter 6: Data Analysis – This part explain the choices and steps taken in the data analysis process.

Chapter 7: Results – In this chapter we present the results obtained and their discussion.

Chapter 8: Limitations and Future Works – This chapter describes the limitation of this work and suggestion to extend it.

Chapter 9: Conclusion – This chapter draws conclusions and deductions derived from this work.

Acronyms – This section includes all the acronyms used within this thesis.

Glossary – This section defines the main glossary and terminology used in the thesis.

Chapter 2

Related work

This chapter presents existing literature on sleep detection, we divided this chapter in: gold standard (medical devices), wearable sensors and non-wearable sensors.

Figure 2.2 shows in y-axis the accuracy of the models and in x-axis the user burden, polysomnography reaches the highest accuracy but also the user burden is the highest, instead bed sensors and wearable devices have still a good accuracy but require less user burden. In the same paper Perez-Pozuelo et al. [2020] show in Figure 2.2 a in depth comparison where for each device is evaluate its performance for the following metrics: sleep time, sleep quality, sleep stages, sleep disorders, scalability, usability.

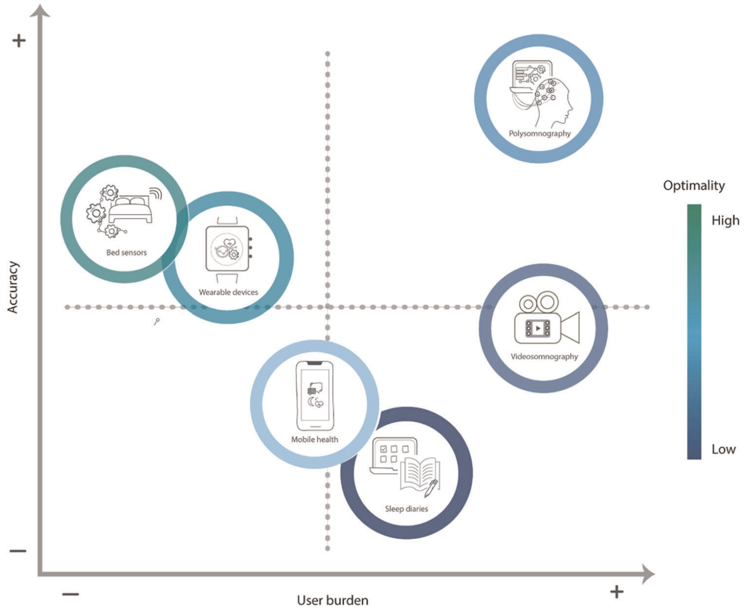
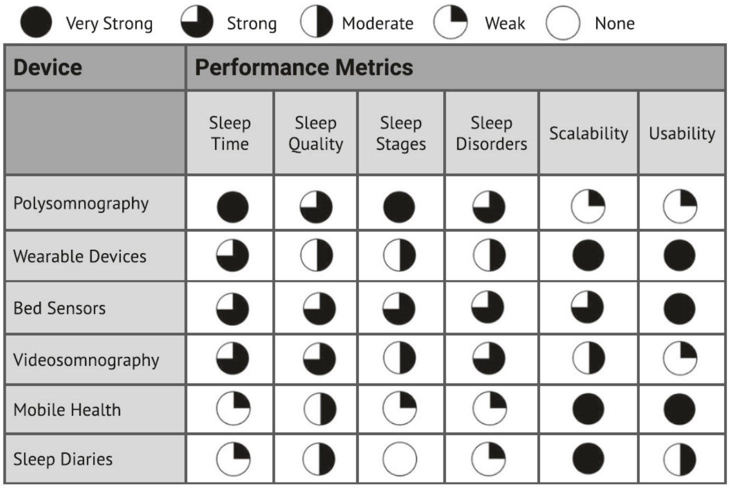


Figure 2.1. Different methods for infer sleep and their accuracy and usability trade-off [Perez-Pozuelo et al., 2020]



● Very Strong ◕ Strong ◑ Moderate ◔ Weak ○ None

Device	Performance Metrics					
	Sleep Time	Sleep Quality	Sleep Stages	Sleep Disorders	Scalability	Usability
Polysomnography	●	◕	●	◕	◔	◔
Wearable Devices	◕	◑	◑	◑	●	●
Bed Sensors	◕	◕	◕	◕	◕	●
Videosomnography	◕	◕	◑	◕	◑	◔
Mobile Health	◔	◑	◔	◔	●	●
Sleep Diaries	◔	◑	○	◔	●	◑

Figure 2.2. Evaluation of sleep-monitoring methods through the following performance metrics: sleep time, sleep quality, sleep stages, sleep disorders, scalability and usability [Perez-Pozuelo et al., 2020]

2.1 Medical devices as gold standard

Between all the different ways that researchers use to detect sleep the one that is globally recognized as gold standard is Polysomnography (PSG). Ibáñez et al. [2018] compares the different types of sleep assessment and indicated PSG as the best sleep detection method in terms of accuracy. Nevertheless, it is also the most intrusive approach and needed an expensive equipment and expert set-up. PSG requires: Electroencephalography (EEG), Electrocardiography (ECG), Electromyography (EMG) and Electrooculography (EOG).

PSG is also highly impractical and cumbersome to be used in real settings since it requires wearing multiple sensors and prevent the user to behave naturally [Sano et al., 2018].

To sum up, PSG is the gold-standard for sleep detection but is not scalable and is difficult to apply in contexts other than laboratory ones [Zhai et al., 2020].

2.2 Traditional manual methods

To overcome the issues with medical approaches and enable tracking sleep behavior at user's home, researchers usually used diaries and self-reports.

Asking people to take note on the time they fall asleep and they wake up is still reliable, especially if they take note in those exact moments.

However, self-reports and diaries are prone to recall biases, because users forget to take notes. Additionally, it is not easy for people to indicate exactly the sleep onset, since some people need minutes or even hours to fall asleep.

While self-reports and diaries might increases the burden on the user, they are reliable and widely adopted [Min et al., 2014; Sano et al., 2016] because they are very easy to apply and do not required special equipment, as for example require PSG.

2.3 Non-wearable sensors

Non-wearable devices are also extensively used in research [Wang et al., 2017; Chen et al., 2013; Min et al., 2014; Hao et al., 2013]. Wang et al. [2017], for instance, used smartphone, in particular light features, activity, phone lock state and microphone features to infer sleep duration.

Saeb et al. [2017] used only mobile phone sensors (geographic location, sound, motion, light and in-phone activities) to monitor sleep, and they used random forest classifier to build a global prediction model. They obtained an accuracy of 88.8% without improving the quality of data (by removing participants whose missing data were above 50% and correcting wrong reports), 91.8% with the improvement. They also underlined as the accuracy vary remarkably across the subject, ranging from 65.1% to 97.3%.

Min et al. [2014] collected one month of phone sensor data (e.g., accelerometer, microphone, ambient light, screen proximity, running process, battery state and display screen state) and sleep diary entries from 27 people. By using a Bayesian network with feature selection they were able to classify if a person was asleep or not in a 10-minute window with 93.06% accuracy and to predict a good or poor sleep quality with 83.97% accuracy.

An example of a commercial non wearable device for screening sleep disorder is Sleepiz [2021]. This device use radar signals to measure breathing, pulse rate and movement and from these measures it will diagnoses sleep disorders (e.g., sleep apnea, chronic obstructive pulmonary disease).

Another example of commercial non wearable device can be an app installed on a mobile smartphone. Some android apps that detect sleep are:

- Sleep Cycle app: uses accelerometer and microphone to detect sleep, it gives also a sleep quality score (required at least five nights to calibrate). It provides personalised alarm clocks based on ideal timings by ringing the alarm during light sleep to help users to wake up refreshed. Among all these app presented in this list, it is probably the most popular ones.

- Sleep Score app: uses "Sonar" function to detect sleep. In practice it uses the same principle of bats, it uses phone speakers to send silent signals and look at how those reflected waves are received back through the microphone. Anyway, only some smartphone supports this Sonar function (e.g., Samsung Galaxy S7, S8, S9, Note 8, Note 9 and Pixel 2 XL).

- Sleep Monitor: this app uses only microphone. Furthermore, during sleep recording, it stores some records (like short voice messages) when loud noises are detected to better investigate them.

- SnoreLab app: uses microphone and records all the night, this way it is possible to listen to all the sound caught by the microphone during sleep.

Though the promising results of non-wearable sensors, their application is just related to sleep/awake patterns since it will not perform that well to infer sleep quality or sleep disorders. Another issue that arises is the privacy one, since most of these devices are microphone based and not all the people will be willing to give companies the full access to their microphones.

2.4 Wearable sensors

Wearable technology market is plenty of commercial devices, such as Samsung Galaxy Watch, Apple Watch or Xiaomi Mi Band. Nevertheless, these devices are still struggling in accurate identify sleep stages as indicated by Stone et al. [2020], that is why usually they grouped sleep stages to increase the accuracy. Furthermore, most of manufacturing companies give little or no information regarding their reliability.

Several researchers have used wearable sensors to detect sleep [Sano and Picard, 2014; Sano et al., 2015; Sadeghi et al., 2019; Zhai et al., 2020]. For instance, Sano and Picard [2014] used wristbands and by using accelerometer, skin temperature and skin conductance, they obtained an accuracy of 86% in intra-subject classification and an accuracy of 74% in case of inter-subject classification.

Another way to measure sleep is using actigraphy. The measurement takes place through an actigraph, which is a wearable device that uses in most of the cases accelerometer, but sometimes also gyroscopes and magnetometers [Hibbing et al., 2017]. Actigraphs are unobtrusive: small and comfortable to wear, can record full days for a week and even longer [Smith et al., 2018]. Nevertheless, this technique is based on the observation that during sleep there is less movement than during wake time, anyway this is still not precisely alone, some people stay awake for some time when in bed [Devi, 2018].

Furthermore, as Lee-Chiong [2005] pointed out some sleep disorders change sleep patterns like usual movement, this means that almost ∼30% of actigraphy detection will be erroneous.

There are a lot of commercial devices that provide sleep quality and advice on how to improve it, a recent good review of some of those has been made by Stone et al. [2020] in which they compared them based on reliability of total sleep time (TST), total wake time (TWT), sleep efficiency (SE).

These devices are very promising as unobtrusive tracking systems. Nevertheless, there is still challenges, among all, the most urgent one is data quality. In real-world settings participants move freely and do not always wear devices properly. The presence of noise and missing data significantly hamper accuracy and robustness of system based on those signals. To this end, noise-robust solutions are needed to ensure high accuracy detection of sleep.

Chapter 3

Background

The following section describes the current knowledge of some key aspect of sleep and sleep quality, underlying also how physiological states can help to understand sleep and wake patterns and their impact on sleep quality.

3.1 Sleep Stages and Sleep Quality

The definition of sleep quality is not trivial as varies between individuals as Buysse et al. [1989] pointed out. Anyway, Buysse et al. [1989] gave a definition: "Sleep quality, on the other hand, is defined as one's satisfaction with the sleep experience, sleep quantity and feeling refreshed upon wakening [Buysse et al., 1989]".

There are two big sleep stages: sleep and awake. Awake can be defined as that stage in which a person is aware and conscious of his surroundings. At the contrary, a subject is sleeping when he is not aware of anything that happen around him.

To go deeper, sleep can be split in two main sleep phases: Rapid Eye Movement (REM) and non rapid eye movement (NREM). The NREM sleep is divided into three separate stages, usually called NREM Stage 1, NREM Stage 2 and NREM Stage 3 [Kales et al., 1968]. Stage 3 is the deepest stage of NREM defined "deepest" because in this phase it is much more difficult to wake an individual, it is also called Slow Wave Sleep (SWS).

Figure 3.1 shows an example of how these different sleep stages alternate between each other during a whole night. As we can see, these phases alternate cyclically over approximately 90 minutes with REM sleep periods getting progressively longer [Mary and William, 1979].

The sleep cycle usually starts with NREM stage 1, passes through the other stages of non-REM sleep and finishes with a short period of REM; then it restarts repeatedly until awakening.

According to Rechtshaffen and Kales (R & K) guideline, in total there are five sleep stages: REM, S1, S2, S3 and S4. There is also another guidelines, the one of the american academy of sleep medicine (AASM) which consider only four stages: REM, S1, S2, S3. The correspondence between these two can be founded by merging S2 and S3 of R & K with S3 of AASM.

Because it is difficult to distinguish N1 and N2, stages are sometimes grouped into light sleep and compared to deep sleep. This result in four stages: wake, REM, light and deep sleep. For instance, this is the case of the sleep detection made by Samsung wearables.

The following list illustrates the common characteristics of each sleep phase:

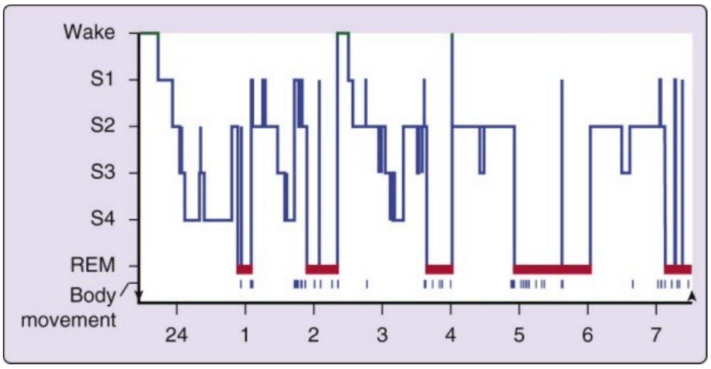


Figure 3.1. Alternation of sleep phases during a single night [Carskadon and Dement, 1989]

- REM: brain activity is close to the one seen in wakefulness and muscles are in a phase of atonia – completely absence of tone. Breathing becomes faster and irregular but heart rate and blood pressure are like waking. REM plays a key role in strengthening neural connections.
- NREM stage 1: is a transition phase between wakefulness and sleep, at this phase muscles are active but heartbeat and breathing slow.
- NREM stage 2: is the period of light sleep before deep sleep, muscles relax, heart rate and breathing slows and skin temperature decreases.
- NREM stage 3: deepest sleep phase, heart-beat and breathing reach their lowest levels and there is no muscle activity. This phase is responsible for tissue repair and regeneration; also dreaming and sleepwalking can occur.

Roebuck et al. [2014] analysed different physiological parameters and stated that for sleep analysis these four are the most important: hearth rate, respiration rate, temperature and body movement.

In our work we focused only on detecting sleep and awake segments since it is more feasible to obtain ground-truth and it does not require cumbersome equipment to validate our model, since the ground-truth in detecting sleep stages, as REM and NREM, is polysomnography.

3.2 Electrodermal activity (EDA)

Electrodermal Activity (EDA) reflects the changes in the electrical conductivity of the skin, which are due to sweat gland activity [Boucsein, 1992].

It is commonly recorded by two electrodes place on the device, usually the device is located on the palm, wrist or fingers of the non-dominant hand since it is less susceptible of artifacts.

The system that governed these changes is Sympathetic Nervous System (SNS) [Choi et al., 2011]. Since the activity of sweat glands through the sudomotor nerve is considered as a direct consequence of changes in the sympathetic activity, EDA can be considered an effective way to monitor it [Boucsein, 2013].

The most important signals obtainable from electrodermal activity (EDA) are peaks, they are defined as rapid evoked changes in the EDA signal due to stimuli, also known as Skin Conductance Response (SCR).

The signal can be composed in two main components:

- Phasic: is the rapidly changing peaks, these changes are due to short-term events. They can be observed during some environmental stimuli - cognitive processes, sight, sound, smell, etc.

- Tonic: represent slowly changing levels, it is the usual value of EDA in absence of environmental or external stimuli. These changes may depend on psychological state, skin dryness, autonomic regulation and hydration.

In several research these measures are mainly used to infer emotions, stress [Lanatà et al., 2015], engagement [Di Lascio et al., 2018], quality of social interactions [Riobo et al., 2014] as well as sleep. Sadeghi et al. [2019], for instance, observed the relation between EDA and all the sleep stages with the result that EDA is strongly associated to deep sleep, more than the other stages.

Figure 3.2 shows patterns of storms during night sleep, as reported by Sano and Picard [2011]. By investigating correlation between EDA and sleep stages, they found that higher percentage of storm epochs during slow wave sleep of the first quarter of the night was directly associated to a greater subjective sleep quality.

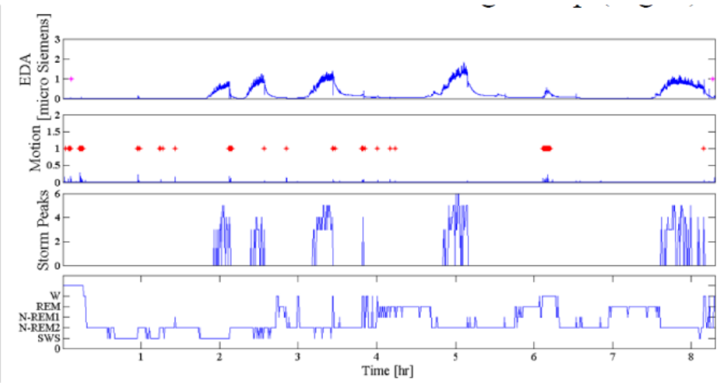


Figure 3.2. EDA during sleep, with storms and sleep stages [Sano and Picard, 2011]

3.2.1 Artifacts

Nonetheless, this signal is still prone to errors due to impact of artifacts that significantly hampers the quality of the data. A part of this work consists of extends the work by Gashi et al. [2020].

As pointed out by Gashi et al. [2020], artifacts can be divided in two groups:

- Shape artifacts: artifacts that cannot be linked to physiological responses since are not conformed to a normal physiological response, usually due to misplaced electrodes or their movement.

- Thermoregulation responses: similar to EDA responses but are not caused by electrodermal system but by the thermoregulation system. These responses are due to a change in the environmental temperature - the body performs a series of responses to keep the body temperature constant regardless of the external temperature - or a physical intense action - heat production in response to physical work.

We used the automatic approach develop by Gashi et al. [2020] to label EDA segments as artifacts.

3.2.2 Storms and Peak Epochs

Before talking of storms we needed first to define peak epochs, since Sano and Picard [2012] define storms in relation of peak epochs.

EDA peak epochs are when there is a minimum of 4 peaks in a time window of one minute.

Sano and Picard [2012] defined **storm** as peaks epochs that last more than 10 minutes.

Anyway, the first one to give a definition of storm was Burch, that stated: "Storm is a minimum of 5 galvanic responses (GSRs)/min for at least 10 consecutive minutes".

Then Picard Rosalind W. defined these regions manifest high frequency of electrodermal activity with between 4 and 10 peaks each minute.

In this study we stayed strict to the definition of Sano and Picard [2012] since they study extensively these regions during the past years.

The pattern of storms was tracked by Picard Rosalind W. and they found that storms tend to be spaced apart by 60-90 minutes, as sleep stages. In particular, this high frequency peak patterns usually shows up during deep sleep.

Also, EDA storms are related to activities or events before sleeps as well as they seem to be associated with storing information of daily events [Cacioppo et al., 2007]. Storms during night are not restricted to sleep domain but they are also positive correlated to anxiety [Boucsein, 2013].

We used these definitions to set the rules in order to label EDA segments as peak epoch or storm.

3.3 Accelerometer

Accelerometer is the most important sensor in actigraphy and it is with no doubt related to sleep. This is because there are more body movements during wake and less during sleep.

Body movements are also strongly related to sleep disorder, as the sleep-related rhythmic movement disorder (SRMD), thus accelerometer could help in diagnostic these kinds of problems.

3.4 Skin Temperature

Skin Temperature play an important role in sleep detection. Kräuchi et al. [2004] underline as distal skin temperature increase before going to bed, when body prepares for sleep, and decrease at wake up, more specifically during sleep onset both distal and proximal skin temperatures increase between 0.5 and 0.9 °C [Kräuchi et al., 2000].

Chapter 4

Data collection

The study procedure consisted of three main phases pre-study, study and post-study phase. We describe in details each step as well as the collect data at the end of the study. Figure 4.1 shows a summary of the procedure dividing it in pre-study, study and post-study.

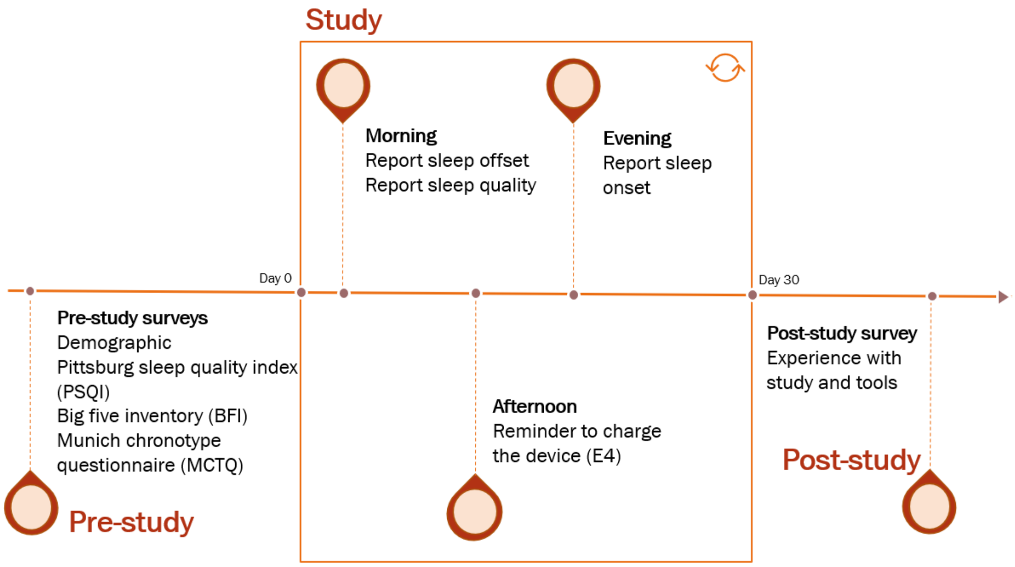


Figure 4.1. Summary of the procedure used during the study

We chose to carry a data collection in a real-world setting to ensure we obtain data about the natural sleeping behaviour of users.

4.1 Participants

Participants were recruited by advertising the study using snowball sampling, flyers and mouth-to-mouth propaganda. To use the android app to collect behavioural data and self-reports we had to restrict possible study participants to only those with an Android Phone (Android version

8.0+) as their primary phone.

In total, we collected data from 16 participants: ten students, three workers, two PhD students and one postdoc. Of which 11 are Male and 5 Female. Furthermore eight of them already track their sleep in some way (e.g., using a smartwatch or smartphone application).

4.2 Pre-study

In this phase we asked participants to complete four different surveys, gave at each participant a E4 wristband, a Paper&Pen physical diary to record sleep events, and we provided study description and tutorial on how to use and install the tools.

4.2.1 Surveys

Before the study, we asked participants to fill four questionnaires on Google Form about their demographics, sleep routine (pittsburgh sleep quality index (PSQI)), personality (big five inventory (BFI)) and chronotype (munich chronotype questionnaire (MCTQ)), in order to better understand the general characteristics and heterogeneity of the sample.

Demographic questionnaire : in this questionnaire we asked participants several questions about their demographic characteristics, as age, gender and occupation, for a total of 9 questions. In subsection A.2.1 we reported the exactly questions we asked to participants.

Pittsburgh sleep quality index (PSQI) : is a self-rated questionnaire created by Buysse et al. [1989] in 1989 that has 19 self-reported questions and five questions answered by bed partner or roommate if there is one. The aim is to assess disturbances and sleep quality over a 1-month time interval.

We slightly modified the original PSQI, in the question "During the past month, how would you rate your sleep quality overall?" the answers should be four: very good, fairly good, fairly bad, very bad. However, we prefer to keep consistency between this test and the options we gave to participants in their self-reports, so we used these five possible answers: excellent, good, normal, poor, very poor.

Therefore, to compute the score of that question, and so stretching and squeezing values in order to go from a scale of 0-4 to 0-3, we used this formula:

$$Y = n * \frac{X - X_{min}}{X_{range}}$$

Where X is the original variable, X_{min} is the minimum observed of X variable, X_{range} is the difference between the maximum and the minimum of X, n is the upper limit of the rescaled variable and Y is the rescaled value that we want to obtain.

subsection A.2.4 shows this survey as it was presented to participants.

Big five inventory (BFI) : is a self-report survey invented by John et al. [1991] and is composed by 44 questions. It is widely used to measure the Big Five dimensions:

- Openness: people who are high in this trait tend to be open to new challenges, creative and curious.

- Conscientiousness: is characterized by responsibility, a strong component of self-reflection and precision in every kind of work.
- Extraversion: high extroversion people tend to have a huge number of friends and acquaintances and love talking even with people they do not know.
- Agreeableness: is characterized by compliance and high empathy.
- Neuroticism: people who are high in this trait tend to be emotional instable, easily upsettable and worried about many things.

In subsection A.2.2 we reported this survey.

Munich chronotype questionnaire (MCTQ) : was developed by Roenneberg and Merrow [2003] and consists of 19 questions. It computes the chronotype as the middle point of sleep onset and offset in free days, this result can be obtainable only if the subject does not use alarm in those days. There are also other measures obtainable from PSQI, we computed also average weekly sleep duration and weekly sleep loss.

subsection A.2.3 shows the whole questionnaire.

4.2.2 Surveys results

Table 4.1 shows some of the demographics and psychological traits of the sample. The chronotype obtained from the munich chronotype questionnaire[Roenneberg and Merrow, 2003] is not shown in table since the chronotype can be computed only if a user does not use alarm in free days, which is the case of 10 participants on 16. So we prefer to report this data separately since it does not represent the whole dataset.

In munich chronotype questionnaire (MCTQ), chronotype has the format "hh:mm" and it is considered as the midpoint between sleep onset and offset. The aggregate chronotype of the 10 participants is respectively min, max, mean and standard deviation: 03:39, 04:55, 04:04, 00:22.

Measure	Min	Max	Mean	Std
Age	19	35	26.44	4.5
Extraversion (BFI)	22	30	26	2.76
Agreeableness (BFI)	22	33	28.25	3.32
Conscientiousness (BFI)	26	35	30.94	2.38
Neuroticism (BFI)	20	28	23.56	2.53
Openness (BFI)	28	43	35.38	4.9
PSQI score	3.25	8.5	5.55	1.53
Average weekly sleep duration (in hours) (MCTQ)	6	9.49	7.55	1.07
Weekly sleep loss (in hours) (MCTQ)	0	2.48	1.07	0.86

Table 4.1. Demographics and psychological traits statistics

To better understand the sample we had with our participants we shown below (Figure 4.2, Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6) some results derived from the pre-study questionnaires.

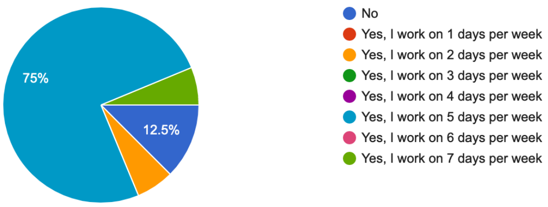


Figure 4.2. Regular work schedule

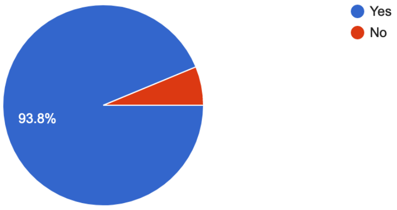


Figure 4.3. Alarm clock on workdays

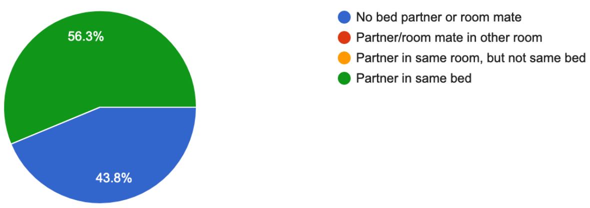


Figure 4.4. Bed partner or roommate

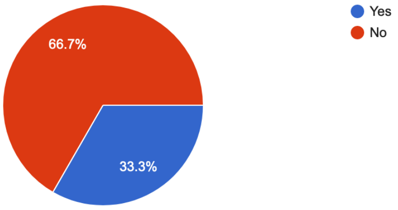


Figure 4.5. Wake up before alarm

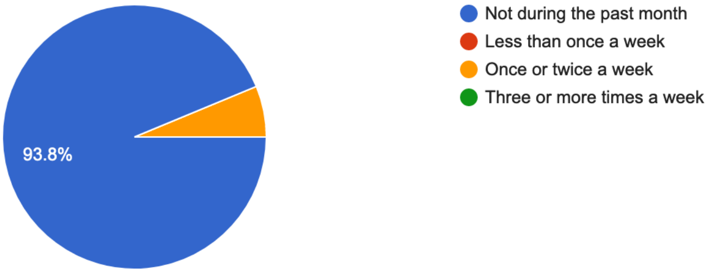


Figure 4.6. During past month taken medicine to help sleep (prescribed or "over the counter")

4.3 Study

After pre-study we ran the study, in this section we presented used tools and the procedure applied.

4.3.1 Tools

We used different tools for each type of data, as we described below.

4.3.1.1 Physiological signals

As we have seen in chapter 2, there are a plethora of different approaches to detect sleep, the same wide choice is the one we have when we are searching for wearable devices.

Different devices have of course different sensors and use different algorithms to obtain various physiological signals. Among all, we decided to use Empatica E4 wristband [1] [Garbarino

[1] Empatica E4: `https://www.empatica.com/en-eu/research/e4/`

Figure 4.7. Empatica E4 wristband

et al., 2014] since it is unobtrusive, small and lightweight. For these reasons we chose this device: it will allow a long-term data collection, without adding bias (e.g., discomfort due to the device), crucially to obtain participants real sleep behaviours. Moreover, it incorporates four sensors into one device such as: photoplethysmography (PPG), 3-axis acceleration (ACC), optical thermometer, electrodermal activity. These sensors collect the data as follows:

- **Blood volume pulse (BVP)**: derived from the green light of the photoplethysmography (PPG) sensor, the sampling rate is 64 Hz.
- **Electrodermal activity**: obtained from two electrodes in the strap, the unit of measure is μS, microSiemens, with a sampling rate of 4 Hz.
- **XYZ raw acceleration**: measurement of acceleration in the X, Y, and Z directions in the range of -2g and 2g, sampling rate of 32 Hz.
- **Skin temperature**: obtained by the optical thermometer, it is expressed in degrees on the Celsius (°C) scale, sampled at 4 Hz.

E4 can be used in two different modalities:

- Recording Mode: it allows recording up to 60 hours of data and stores it in its flash memory. Data remains on the device until sessions are synchronized, by uploading data to E4 connect servers, through E4 manager. Figure 4.8 shows how the recording mode works.
- Streaming Mode: physiological data can be monitored in real-time by using a Bluetooth connection and the E4 real-time App on a smartphone, or even by developing a personal application. In this way data are automatically uploaded to the E4 connect account after each session.

During the data collection we asked participants to upload their data through E4 connect by using an account we provide, in this way we were able to access their data.

We extracted from the sensors embedded in E4 wristband the following physiological signals:

- Blood volume pulse

[2] Empatica E4: https://www.empatica.com/en-eu/research/e4/

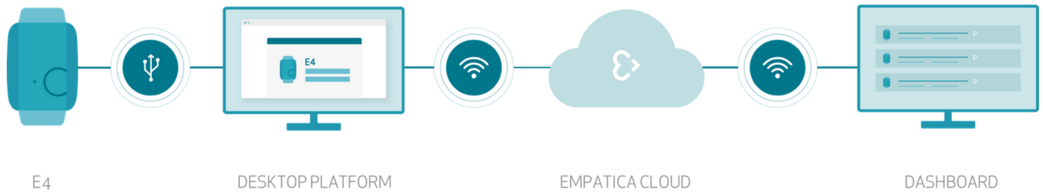


Figure 4.8. Empatica E4 - Record mode [2]

- Electrodermal activity
- Accelerometer data
- Skin temperature

4.3.1.2 Behavioral data

We used SleepApp, an Android application provided by the advisors to facilitate the collection of data for the study. Participants installed SleepApp from the Google Play store, this way they were able to: install it without too effort, receive update promptly, be reassured about the goodness of the app (e.g., no malicious app). Through this app we were able to collect behavioural data as well as self-reports.

Regarding behavioural data we collected:

- Time of phone lock/unlock events
- Time of screen on/off events
- Time and type of applications used on the phone
- Time and application from which a notification arrived on the phone
- Time and proximity of the phone screen to any object
- Time and amount of ambient light

Using this app we were also able to send app notifications every day in order to remind participants to:

- report waking up activity, by sending this message "Please don't forget to report your waking up activity!" every morning
- charge their wristbands, by sending this message "Please don't forget to charge E4 and upload the data" every afternoon
- report sleeping activity, by sending this message "Please don't forget to report your sleeping activity!" every evening

All data collected by this app was first stored locally, on device local database, and then uploaded every day remotely to the SwitchDrive folder created specifically for study purposes. This step did not require any effort from the users since the upload was automatic, anyway we provide an *Upload* button to force the upload in case of synchrony problems.

4.3.1.3 Self-reports

Self-reports were crucial to collect sleep data (sleep onset time, awake time and quality of the sleep) in order to obtain the sleep diaries used as ground-truth.

Since self-reports required effort to the user, remember to record a specific event and record it by providing some details, we prefer to give freedom on how record these events. For each type of event (sleep or awake) participants were able to choose between:

- Sleep App:
    - Widget
    - App home
- Google Form
- Paper and pen diary

The collected data from self reports was:

- Going to sleep time
- Waking up time
- Sleep quality score ('Very Poor', 'Poor', 'Normal', 'Good', 'Excellent')

Figure 4.9, Figure 4.10 and Figure 4.11 show screenshots of SleepApp. In particular Figure 4.9 shows its home from which users were able to record their sleep events by pushing the right button, in case of woke up events the app will show a pop-up that will ask users to score their sleep. Figure 4.10 shows the option that users will have when they decide to add a new event from their diaries, only in this case users were able, through the app, to indicate also the hour and the day of the new event.

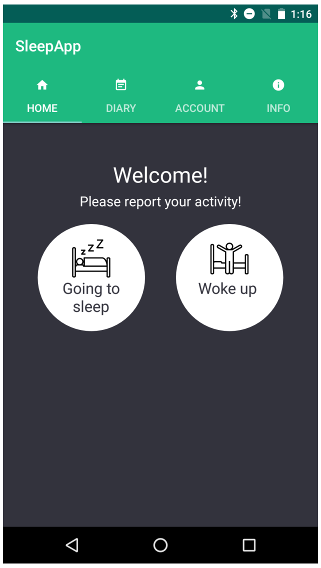
Figure 4.12 shows the structure of the self-report through Google Form. Instead, Figure 4.13 shows one standard page of the Paper&Pen we provided to users.

Google Form was used to allow participants to record sleep events also via laptop. In this way, for example, in case participants forgot to report an awake events and they were already start to work on their laptops, they were able to record their events without getting distracted by the phone.

4.3.1.4 Others tools

Google Form As we mention before, we used Google Form to collect self-reports as an additional optional methods. Anyway, we used it to provide questionnaires and collect response from users, participants fill in five surveys, four during the pre-study and one after the study.

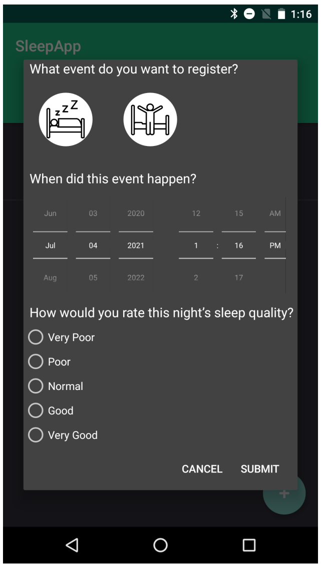
SWITCHDrive We used this cloud data storage service since every person that works or study in università della svizzera italiana (USI) has 50GB of free storage, plus it is widely used in Swiss higher educations, so it is fully implemented by following the Swiss data protection laws.



Welcome!

Please report your activity!

Figure 4.9. SleepApp - Home



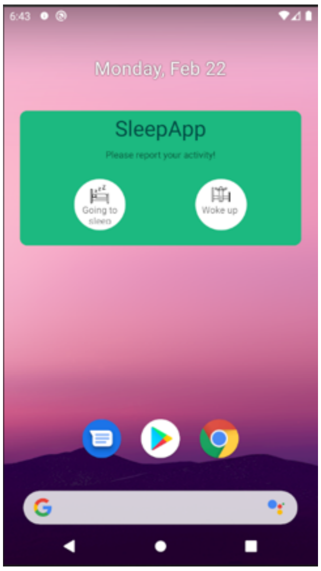
What event do you want to register?

When did this event happen?

How would you rate this night's sleep quality?

- ○ Very Poor
- ○ Poor
- ○ Normal
- ○ Good
- ○ Very Good

Figure 4.10. SleepApp - Diary



Please report your activity!

Figure 4.11. SleepApp - Widget

4.3.2 Procedure

During the study we asked participants to:

- Wear the E4 wristband on the non-dominant hand (which is proven is less prone to artifacts [Picard et al., 2015]), every night, starting from about four hours before going to sleep and four hours after they wake up. We chose four hours before and after to obtain a balanced dataset, since usually adults sleep duration should be in average between 7 and 9 hours [Chaput et al., 2018]. This way for each session we will obtain eight hours of sleep and eight hours of awake in total.

- Install an Android phone application that gathers behavioural data in the background.

- Provide self-reports about the time when they go to sleep and wake up as well as sleep quality when they wake up.

- upload the data from the E4 wristband using Empatica Manager in their laptop.

Furthermore, during the data collection, we monitored the quantity and quality of data through scripts created on purpose (chapter 5), in order to react promptly to any problem (e.g. synchronization problem, device issues).

4.4 Post-study

After the study, we asked participants to fill a questionnaire with questions regarding their experience with the study and the tools. In particular all the questions can be grouped in the following sections:

- Tools for self-reports
- Sleep App
- E4 wristband
- Willingness on sharing physiological data
- Opinions on wearable devices
- Comments about the overall study

In subsection A.2.5 we reported the conducted survey.

Participants were compensated through gift cards whose amounts were directly related to the amount of good data they provide during the study.

4.4.1 Survey results

Between all the questions we asked to participants the following observations are particularly interesting:

- Most of the participants likes the widget and the app in general since some of them use the phone right before close eyes (one uses the phone to set the alarm) and for most of them the first thing in the morning is looking at the phone (some as first thing in the morning switch off the alarm).
- One participant said “Using my phone (thus SleepApp) right when I was about to fall asleep tended to wake me up and influenced my sleeping schedule”.
- Two participants said that “Google form required more actions and effort”.
- "Sleep home and widget faster than others since they did not required setting time".
- None agree at "The device was distracting".

4.5 Privacy

When we talk about data inevitably privacy concerns arise. Not only we have to ensure that it is impossible to track a participant to his or her real name, but also we have to avoid that physiological[Fairclough, 2014] or behavioural data of someone become public since they can be or will be considered as unique identifier of a person as we already do with fingerprint or iris recognition [Piciucco et al., 2021].

Since SleepApp collected a lot of data in background, participants were reassured that no screenshots, page body content, notification content, notification sender or website visited were collected.

Furthermore, participants were aware that they were able to change their mind in any time by withdraw the permission to use their data.

All those personally identifiable information (PII) has to be stored safely and handled confidentially. We ensure this by:

- **Anonymizing data** through the assignment of an alphanumerical code, in this way participants' names were never mentioned as connected to data. The mapping of participants' random id to their actual names is kept separated from other project spaces to avoid any attempt of tracking.
- **Storing safely data** by using SWITCHDrive, a cloud data storage service often used in Swiss Universities, since it is fully implemented by following the Swiss data protection laws.

Only the authorized researchers had access to shared folders to ensure confidentiality, furthermore, data will be kept until 12 months after results have been first published.

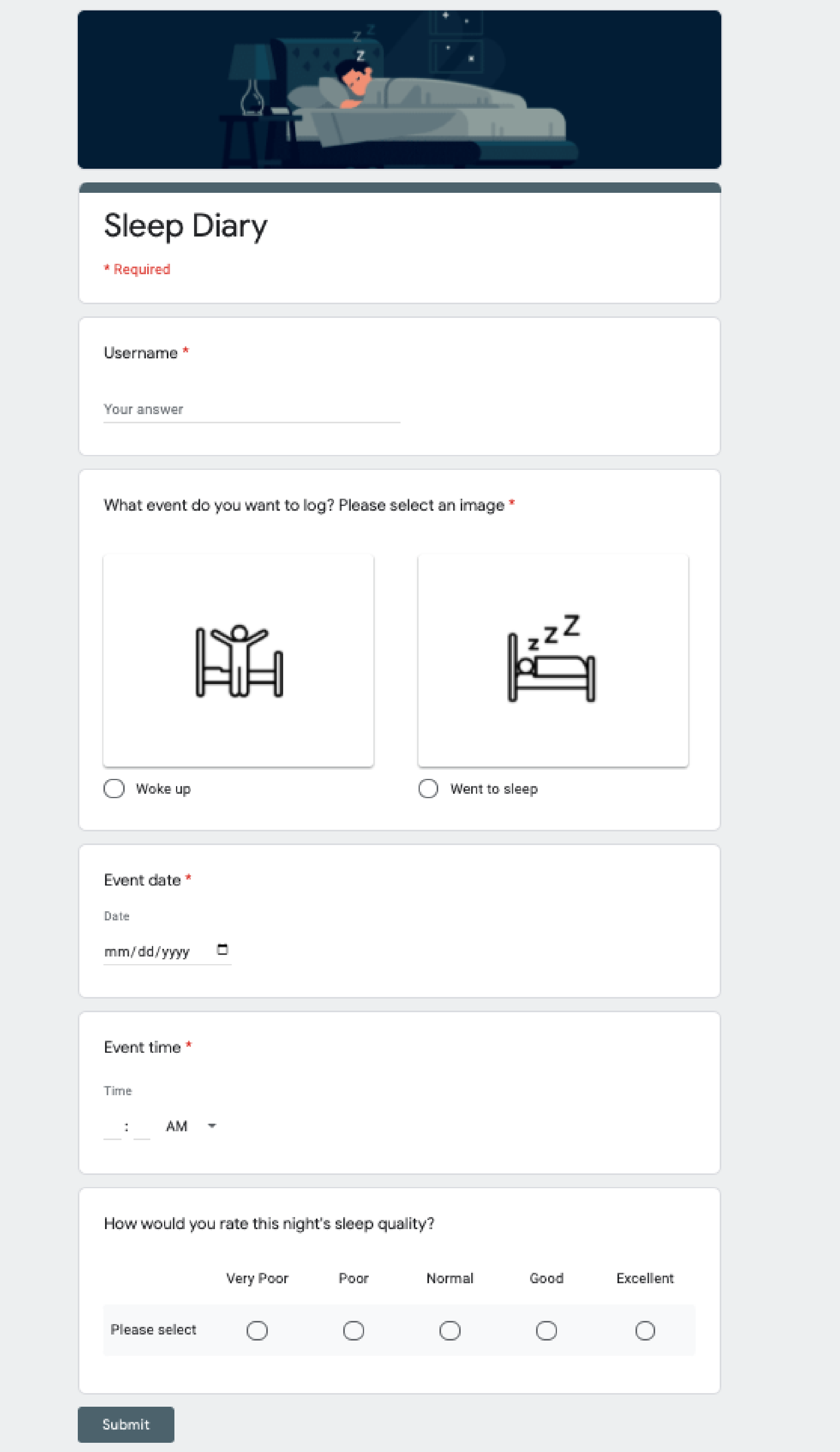
4.6 Collected Data

We collected sensor data from smartphone and E4 wristband in a real-world setting for 30 days, between the end of February 2021 and beginning of April 2021. The dataset was computed this way: for each sleep session we took the hour of sleep, four hours before and four hours after. For example, if we know from sleep diary that one user in a specific day slept from 23:00 to 5:00, the session will be from the first physiological data that we had from 19:00 and the end will be the last physiological data before 9:00. Gap between that data was set to 0 – we define gap as a period of time where we did not receive a sample from the participant's wristband. In total we did this for 130 hours among all users (2% of the total dataset).

In total we obtained 6557 hours, with the following distribution:

- Sleep/Awake problem:
  - Sleep: 49.16%.
  - Awake: 50.84%
- Sleep Quality (with five classes) problem:
  - Excellent: 5.10%.
  - Good: 36.31%.
  - Normal: 42.31%.
  - Poor: 14.35%.
  - Very Poor: 1.94%.
- Sleep Quality (with three classes) problem:
  - Low: 16.26%.
  - Normal: 42.43%.
  - High: 41.31%.
- Sleep Quality (with two classes) problem:
  - Low: 58.70%.
  - High: 41.30%.

The distribution of quality and quantity based on user is shown in the following figures (in Figure 4.18 quantity, in Figure 4.19 quality with five classes, Figure 4.20 quality with three classes, Figure 4.21 quality with two classes).
Similar figures but with an aggregate view are shown in Figure 4.22, Figure 4.23, Figure 4.24, Figure 4.25.



Sleep Diary

* Required

Username *

Your answer

What event do you want to log? Please select an image *

○ Woke up

○ Went to sleep

Event date *

Date

mm/dd/yyyy

Event time *

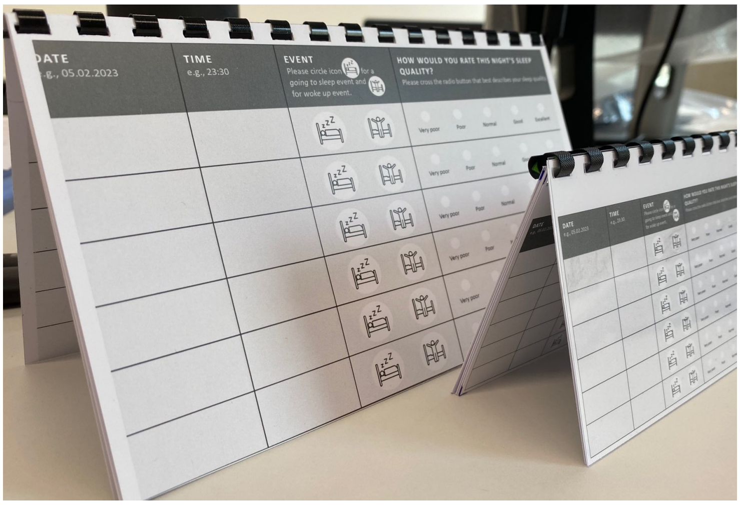
Time

_ : _ AM

How would you rate this night's sleep quality?

	Very Poor	Poor	Normal	Good	Excellent
Please select	○	○	○	○	○

Figure 4.12. Self-reports - Google Form



DATE
e.g., 05.02.2023

TIME
e.g., 23:30

EVENT
Please circle icon for a going to sleep event and for woke up event.

HOW WOULD YOU RATE THIS NIGHT'S SLEEP QUALITY?
Please cross the radio button that best describes your sleep quality

Figure 4.13. Self-reports - Paper&Pen

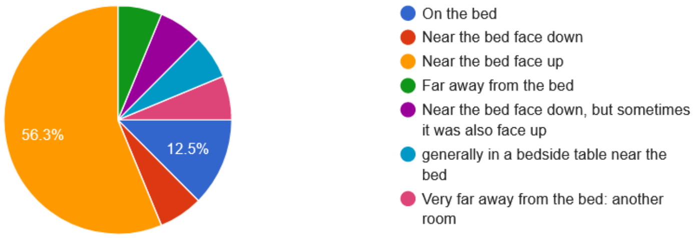


Figure 4.14. Position of phones' participants during nights

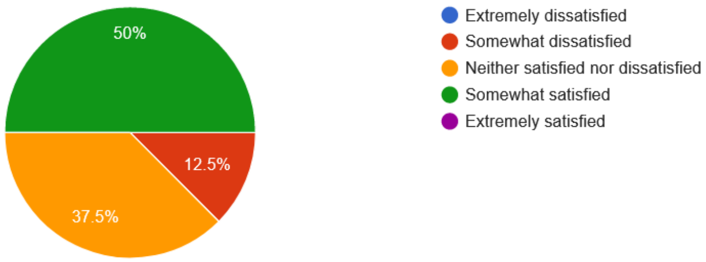


Figure 4.15. Overall experience with the Empatica E4 wristband

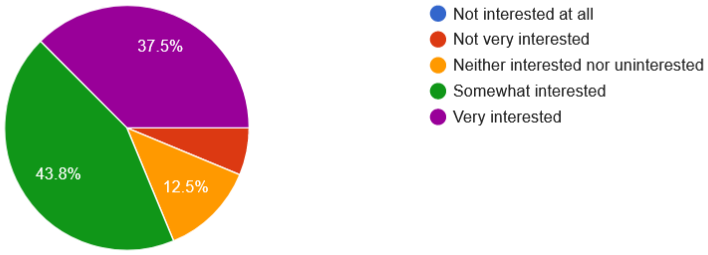


Figure 4.16. Interested on knowing physiological data (e.g., heart rate, body temperature, etc.) throughout the day and night

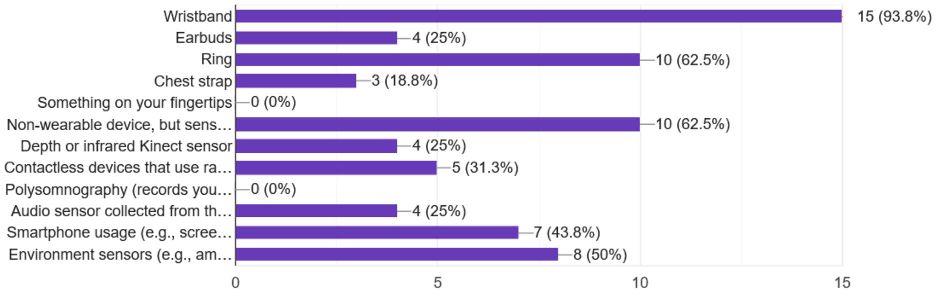


Figure 4.17. Best device to measure sleep

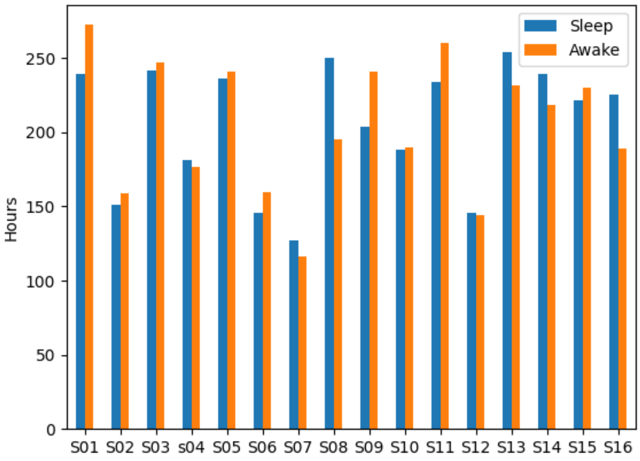


Figure 4.18. Amount of hours of sleep and awake moments for each user

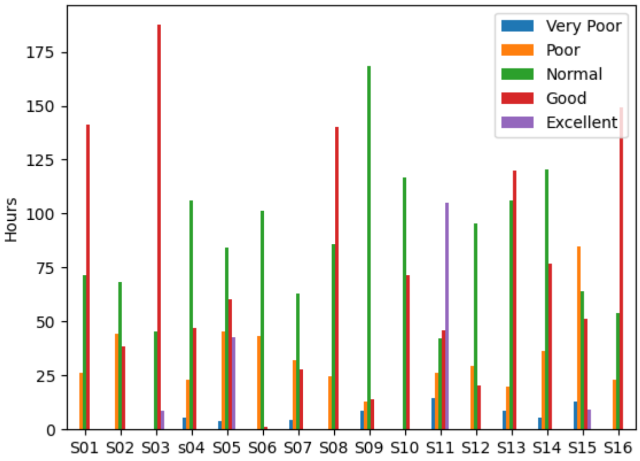


Figure 4.19. Amount of hours of sleep quality (very poor, poor, normal, good, excellent) for each user

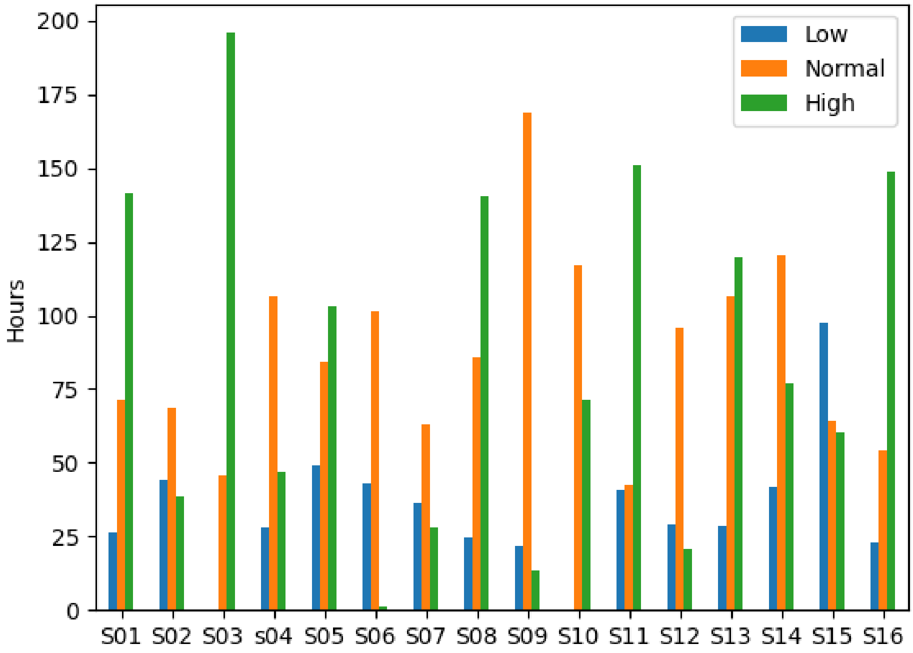


Figure 4.20. Amount of hours of sleep quality (low, normal, high) for each user

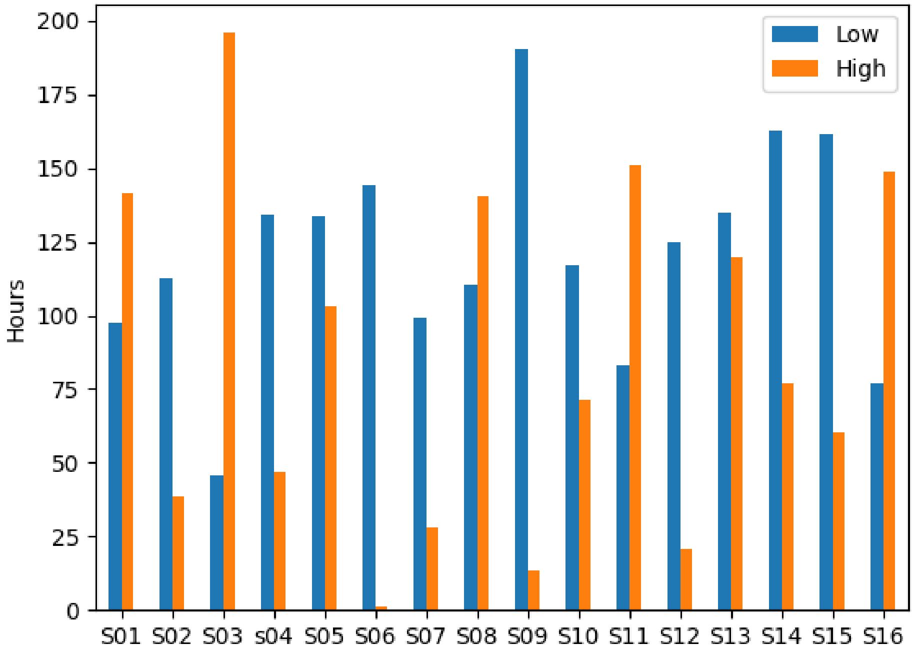


Figure 4.21. Amount of hours of sleep quality (low, high) for each user

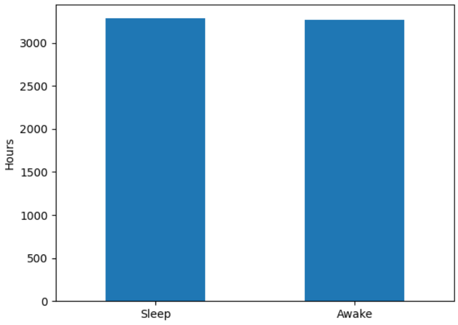


Figure 4.22. Amount of hours of sleep and awake moments

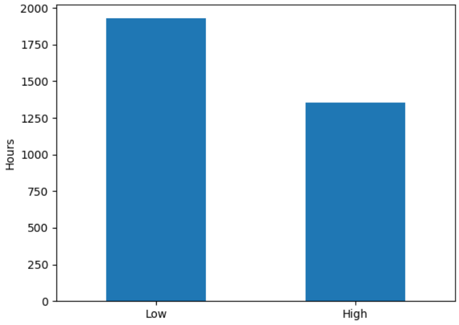


Figure 4.23. Amount of hours of sleep quality (low, high)

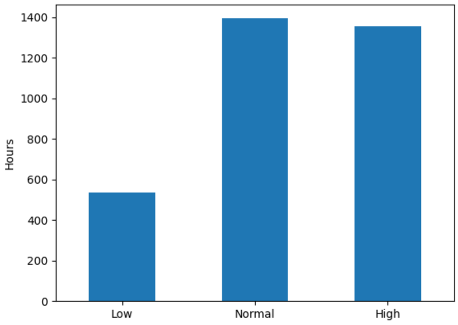


Figure 4.24. Amount of hours of sleep quality (low, normal, high)

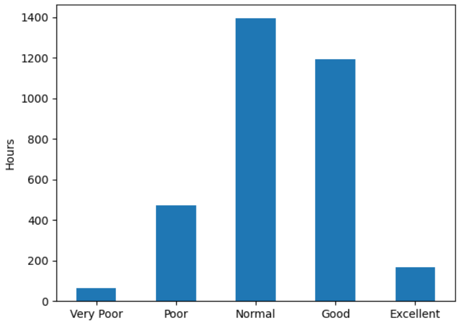


Figure 4.25. Amount of hours of sleep quality (very poor, poor, normal, good, excellent)

Chapter 5

Data visualization

Part of this work consisted of developing a dashboard in order to visualize data collected during the previous phase.

5.1 Dashboard

We used Streamlit [1] , which is an open-source framework that allows to turn data scripts into a web app in few lines of code, by just declaring widgets and connect to each one a function without taking care of front-end or back-end.

We created two different dashboards, one for phone data and the other one for wristband data. Both dashboards have the same structure: a side menu with a burger icon from which we can select the user and a plot, regarding the sensor and the user we select, will show up.

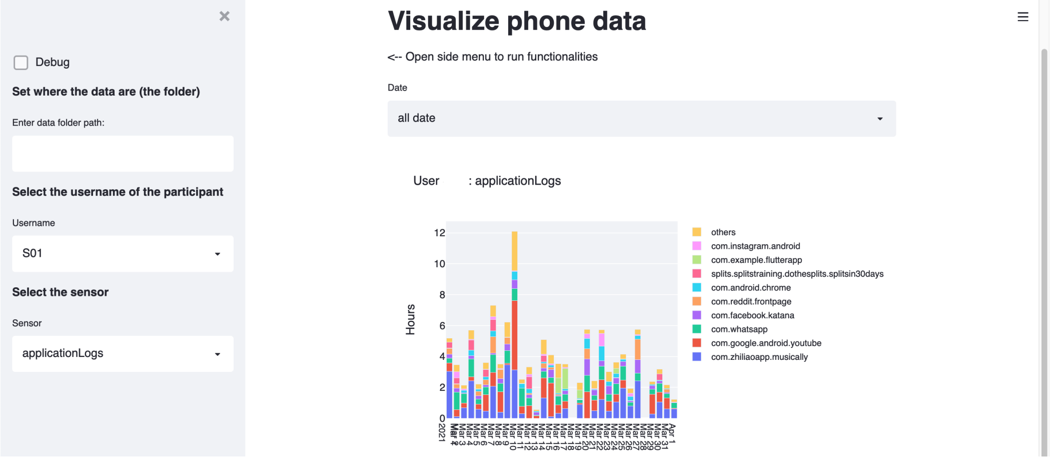
An example of the web app home is in Figure 5.1.

5.1.1 Phone data

For each participants we can select:

- phone lock
- screen events
- applications usage
- notifications
- sleep events
- proximity
- light
- sleep plot

[1] Streamlit: `https://docs.streamlit.io/en/stable/api.html`



×

□ Debug

Set where the data are (the folder)

Enter data folder path:

Select the username of the participant

Username

S01

Select the sensor

Sensor

applicationLogs

Visualize phone data

<-- Open side menu to run functionalities

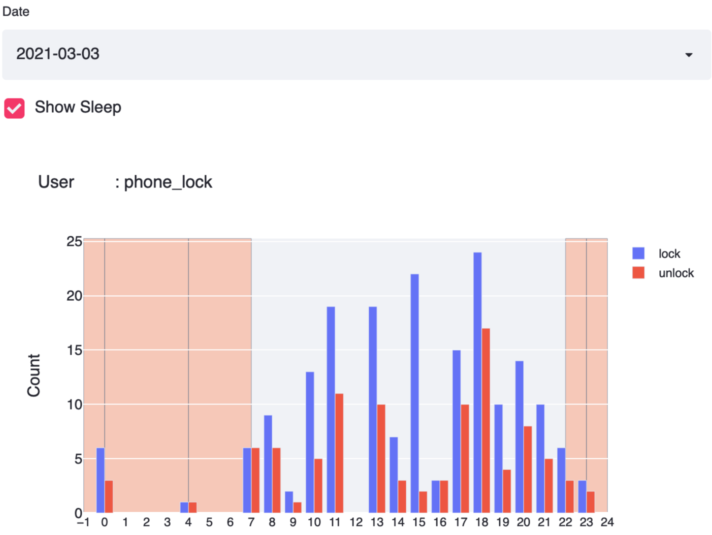
Date

all date

User : applicationLogs

Figure 5.1. Dashboard phone data - Home

For each of those sensors we can look at a general view with the data for each day or we can zoom into a single day and see how that sensor data is distributed within 24 hours. Furthermore, in all day plots we can choose, by clicking on a checkbox, to see also sleep time. An example can be found in Figure 5.2, in that figure there is a sleep time window where the background has a salmon pink colour and the grey line in the background indicates that there was a slightly (under one hour) sleep interruption. As we can see, the participant correctly reports his sleep since based on the phone lock we know that he woke up during that night.



Date

2021-03-03

☑ Show Sleep

User : phone_lock

Figure 5.2. Dashboard phone data - Lock/unlock and sleep

Once we collected all the sleep data, we did a further investigation on sleep onset, offset and duration; for each participant and between all of them.

We computed that plot shown in Figure 5.3 by computing a sort of baseline (defined as the average time, in this case the average time of sleep onset time for the blue line and the average time of sleep offset for the red line) and each point represent how many hours are different from his usual bed/awake time. Furthermore, in all of these sleep plots there was a vertical line on March 28th, that line indicates that night there were full moon and time zone switch to summertime (+1 hours). This information was interesting when we compared our results with the related work.

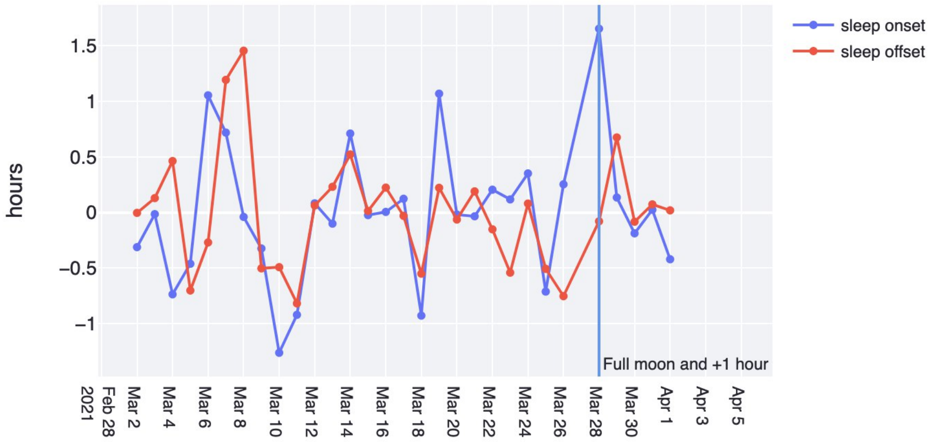


Figure 5.3. Dashboard phone data - Plot sleep onset and offset for one user

To understand if some sleep patterns (e.g., on day X most of users woke up earlier) were equal among all participants we created a box plot computed by using the same values in each plot like in Figure 5.3 but for each user. So, in Figure 5.4, Figure 5.5 and Figure 5.6 we can see that information in an aggregate view.

5.1.2 Wristband data

The dashboard created to visualize the wristband data is very similar to the one created for phone data. This time for each user we will have three sub-plots (by declaring them as subplots in matplotlib a zoom in one of them will result in the same zoom for all of them) that share x-axis which represents time. As we can see in Figure 5.7 and in Figure 5.8, the first sub-plot represents EDA values, the second one is the accelerometer data and the last one is the distal skin temperature. The salmon pink background indicates that a sleep was taken during that time, in this way we are able to further investigate some patterns along physiological data, as that distal skin temperature increase during the transition from wakefulness to sleep.

In the dashboard by clicking on a checkbox we can choose to see also artifacts, peak epochs and storms. As we can see in both the previous figures.

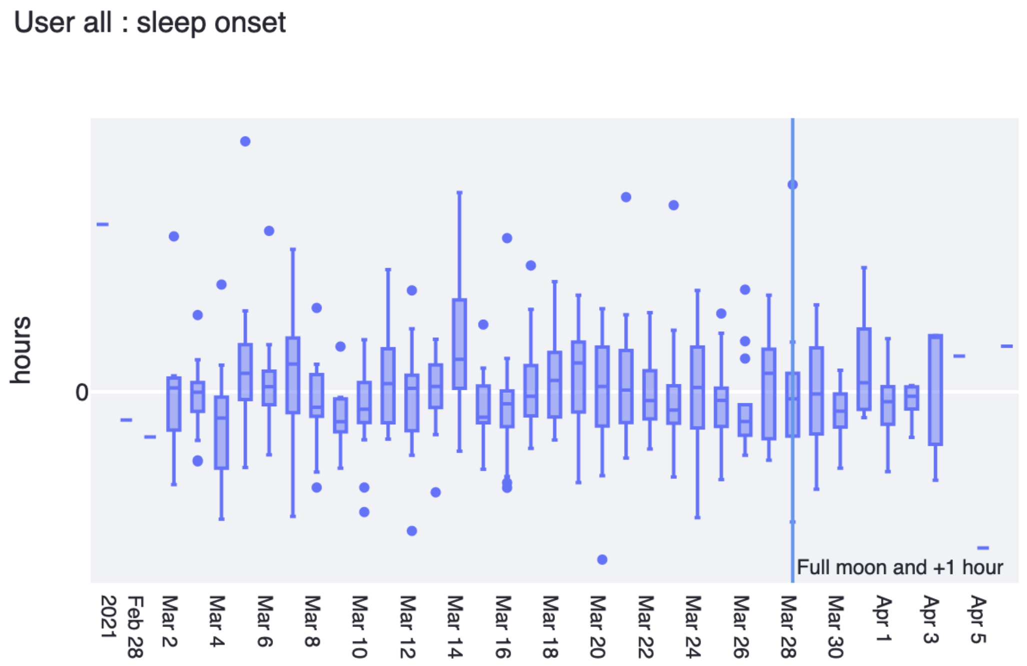


Figure 5.4. Sleep onset among all users, the deviation is based on their personal average

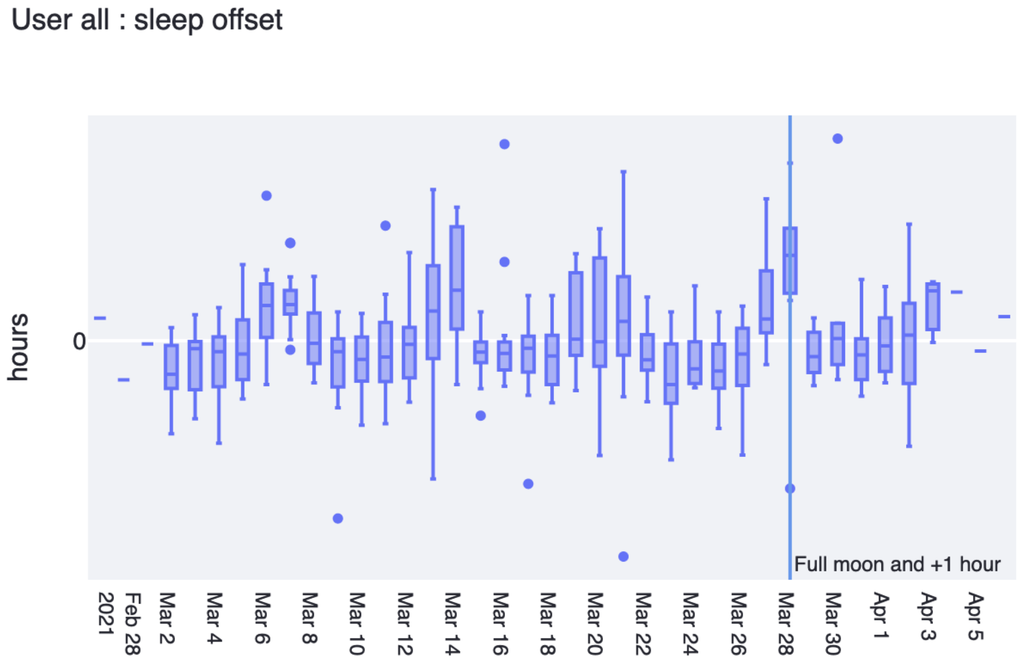


Figure 5.5. Sleep offset among all users, the deviation is based on their personal average

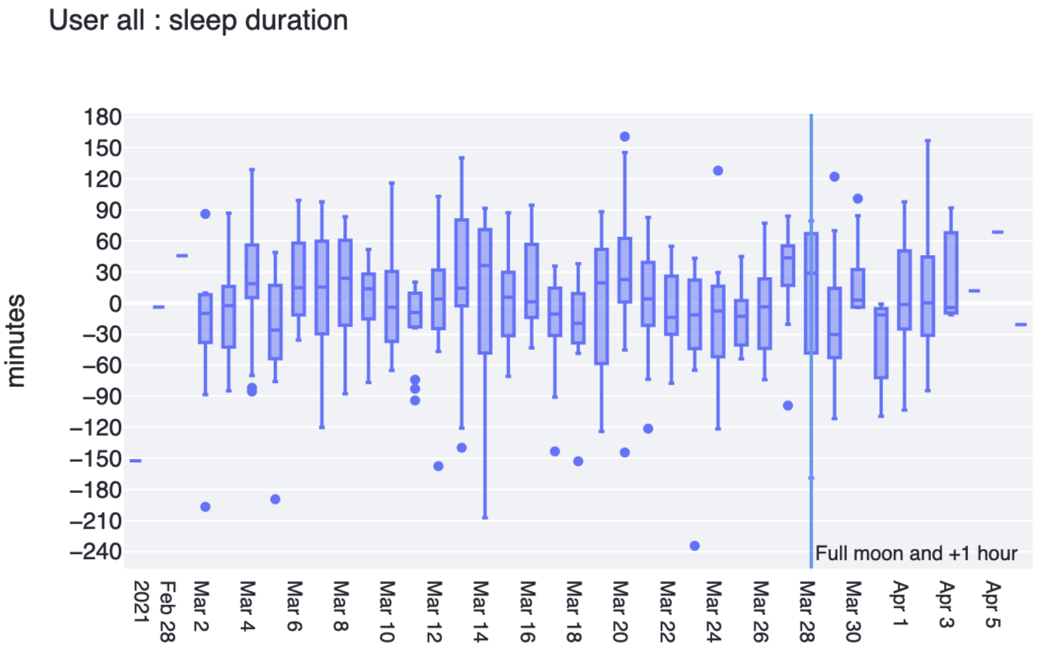


Figure 5.6. Sleep duration among all users, the deviation is based on their personal average

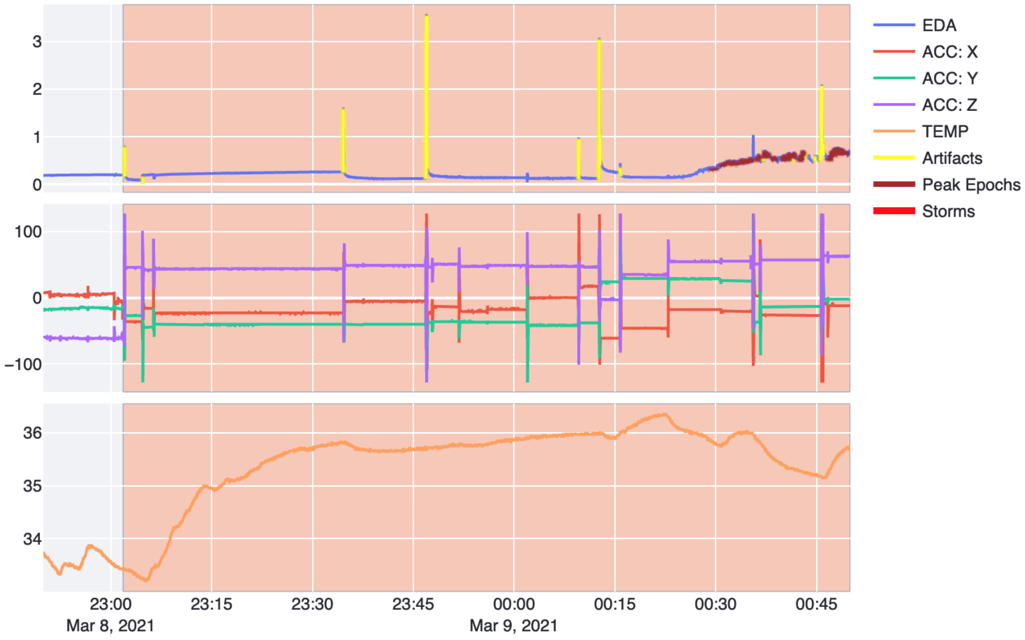


Figure 5.7. Dashboard wristband data - Example 1

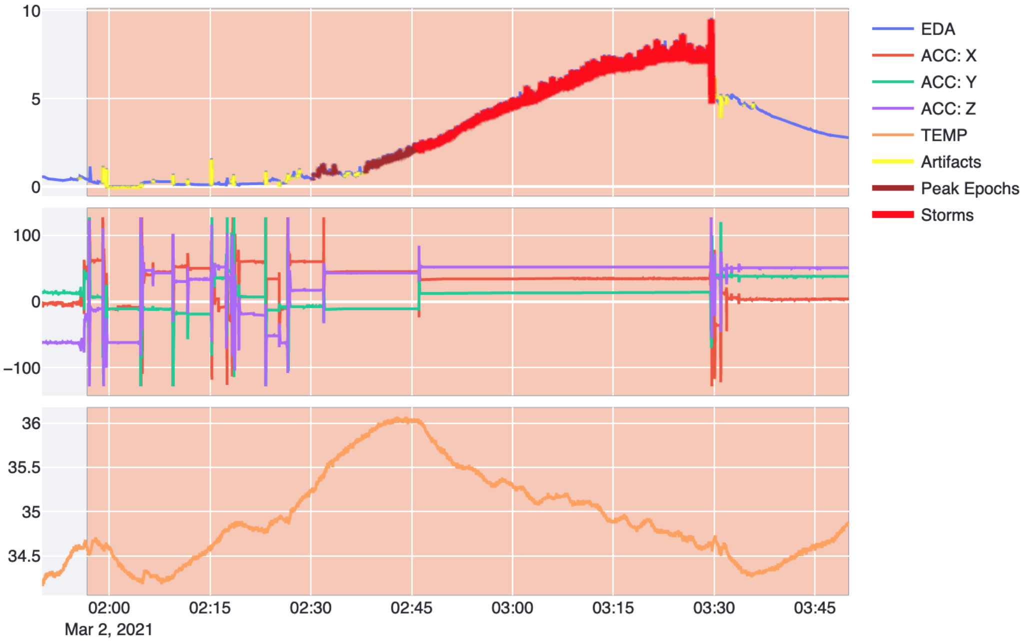


Figure 5.8. Dashboard wristband data - Example 2

Chapter 6

Data analysis

In this chapter we described the following steps: data imputation, preprocessing, segmentation, feature extraction, feature interpretation and classification. Figure 6.1 shows the data analysis pipeline.

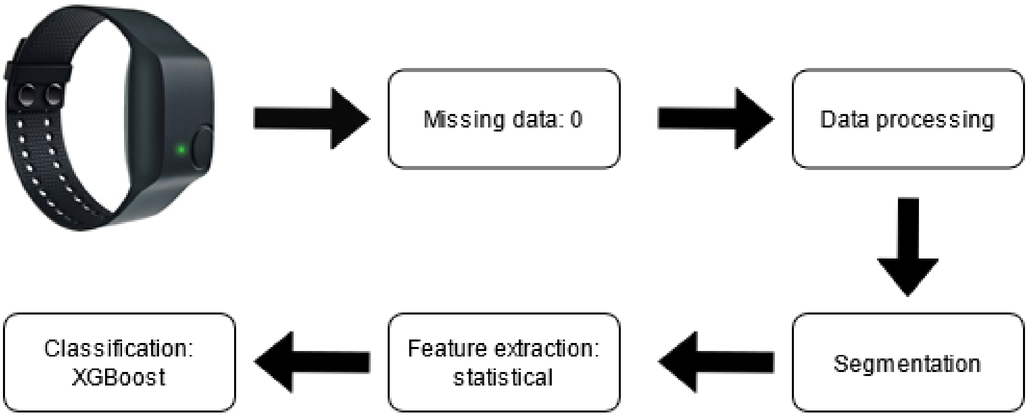


Figure 6.1. Data analysis pipeline

6.1 Data Cleaning

6.1.1 Self-reports

We wanted to track two different things: sleep and sleep quality. To obtain ground-truth we asked participants to reports these events, anyway self-reports suffer from missing data or errors due to manual registration processes. E.g., we noticed that some participants report sleep time but, probably in a hurry to start the day, forget to report also wake time.

To avoid having missed reports or even wrong one (e.g., sleep onset and then sleep offset 14 hours after) we visual inspected self-reports and compared them with behavioural data as screen and phone lock in order to clean that misleading information.

Since we gave to participants multiple ways of report sleep events, we also had to take care of merge reports from different sources (SleepApp, Google Form and Paper&Pen). Merging SleepApp and Google Form was easy since both can be obtainable as csv files. To add also sleep events from Paper&Pen we asked participants to return this physical diary at the end of the study.

One problem of having different sources was that in some cases participants record same events multiple times in different tools, in that case we contacted them by email and asked them which one we had to consider as the correct one.

Since we caught all type of sleep, night one and naps, we were not able to just look at the time between sleeps, since if there is a sleep four hours after another sleep it is very likely that one of the two sleep is a nap.

After some attempts the following rules obtained good results (by looking at sample checks):

- in case of consecutive sleep events, always take the last one and delete the first one
- in case of consecutive awake: if time between both is below 2 hours delete the first and take as good the second, otherwise delete the second and save the first one
- if between sleep and awake there are more than 14 hours, delete both

The previous rules only deal with sleep time, regarding sleep quality we decided to remove sleep events without sleep quality as this choice only makes us discard 2 sessions. Given the amount of data it seemed reasonable enough to make this choice.

6.1.2 Physiological data

As mentioned in chapter 4, to obtain a dataset balanced as possible we decided to consider a session as the time between four hours before the sleep events and four hours after awake events. The problem of taking just time sections when we had physiological data was that, during pre-processing phase, we could not apply some filter if we had holes in the time series, e.g., the low filter necessary to decompose EDA signals in phasic and tonic.

Once we took the first data point in those four hours before the sleep onset and the last data point in the four hours after sleep offset, all data (EDA, ST and ACC) that is missing in between was set to 0.

6.2 Data Processing

For the data processing phase we focused especially on EDA, since between all the sensors we were taking in account it is the one that is more affected by artifacts. For example, Wang et al. [2018] had to discard weeks of collected EDA signals due to the low quality of that data.

To clean and pre-process EDA signals we used an approach similar to the ones done in literature [Gashi et al., 2020]:

1. **Cleaning**: We applied a first order Butterworth low-pass filter with a cut-off frequency of 0.6 Hz to remove high frequency noise fluctuations [Gashi et al., 2020].
2. **Decomposition**: As it was extensively studied, EDA signals can be decomposed in *phasic* and *tonic*, we did that by applying cvxEDA, an algorithm that use convex optimization to decompose the signal [Greco et al., 2016].

3. **Recognize artifact, epoch and storms**: We used a rule-based approach to recognize peak epoch and storm, the rules are the ones described in subsection 3.2.2 and a visual description can be found in Figure 6.2. Regarding artifact we used the EDArtifact module created by Gashi et al. [2020]. In this way we were able to label each EDA value as artifact, peak epoch or storm. A detailed description of this step can be found at subsection 6.2.1.

During the data processing we had to deal with different sampling between EDA, body temperature and accelerometer since EDA and ST have a sampling rate of 4 Hz instead ACC has a sampling rate of 32 Hz. To deal with this difference we decided to down sample the accelerometer data to 4 Hz by taking for each value of X,Y,Z we took the mean after the down sampling. This choice seemed the correct one since a high sampling of accelerometer is more necessary for detecting sports activities and since the segmentation will be higher (1m and above) this choice will not affect the final result.

6.2.1 Identify peak epochs, storms, and shape artifacts

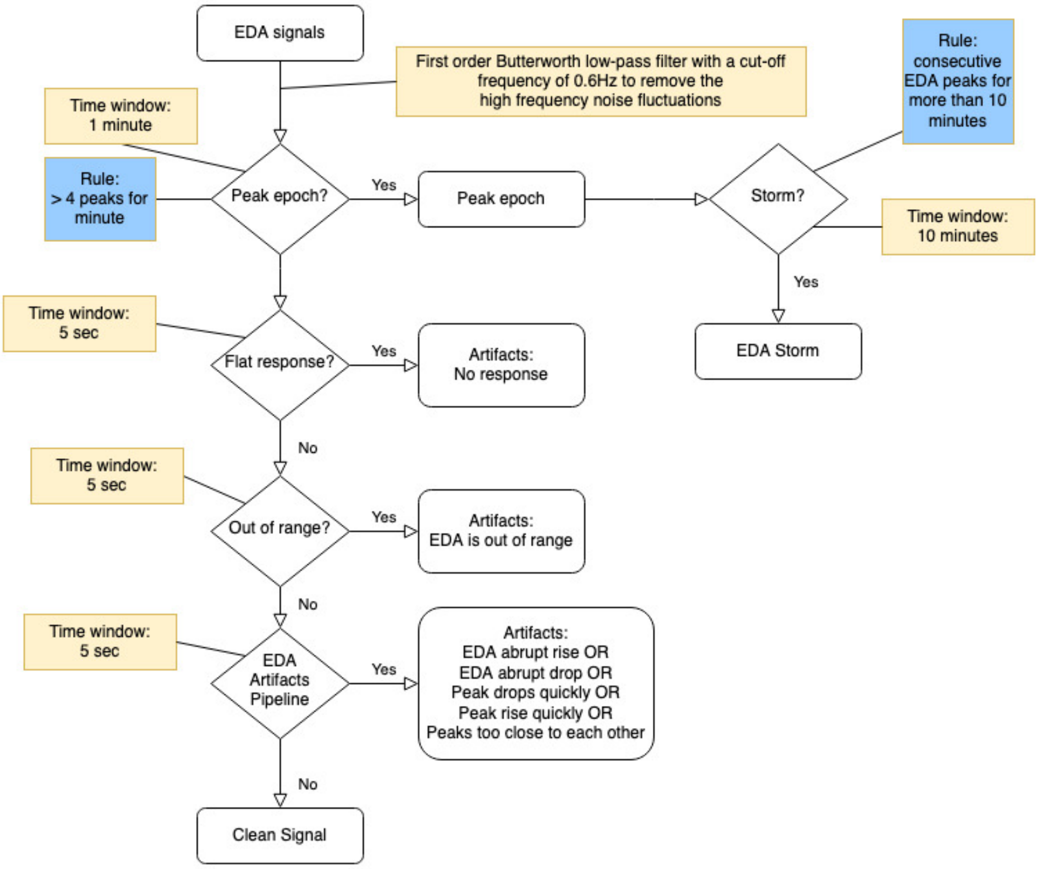

EDA signals
First order Butterworth low-pass filter with a cut-off frequency of 0.6Hz to remove the high frequency noise fluctuations
Time window: 1 minute
Rule: > 4 peaks for minute
Peak epoch?
Yes
Peak epoch
Storm?
Rule: consecutive EDA peaks for more than 10 minutes
Time window: 10 minutes
Yes
EDA Storm
Time window: 5 sec
Flat response?
Yes
Artifacts: No response
No
Time window: 5 sec
Out of range?
Yes
Artifacts: EDA is out of range
No
Time window: 5 sec
EDA Artifacts Pipeline
Yes
Artifacts: EDA abrupt rise OR EDA abrupt drop OR Peak drops quickly OR Peak rise quickly OR Peaks too close to each other
No
Clean Signal

Figure 6.2. Classifier for peak epoch, artifact and storm

As described in Figure 6.2 once we had the filtered EDA signals, so after that the first order Butterworth low-pass filter was applied with a cut-off frequency if 0.6 Hz, we follow this path:

1. **Peak Epoch –** We kept the implementation of Gashi et al. [2020], they used EDAExplorer [Taylor et al., 2015] to obtain the number of peaks in a time window of 5 seconds. Then we used a sliding window with a time of 1 minute and if the number of peaks in that 1-minute window was above 4 then all values in that time window are labelled as "Peak_Epoch". We chose to use a sliding window since 1 minute is a large time window and by using a no-overlapping one we could risk to lose too much information.

2. **Storm –** With the same approach used before we created another sliding window that seeks for consecutive EDA values labelled as *Peak_Epoch* that last longer than 10 minutes. The EDA values that obey to the rule just described were labelled as "Storm".

3. **Artifact –** At this point we continued with the same procedure used by Gashi et al. [2020], and we obtained the "Artifact" label. In this part we prefer not to label as "not Artifact" the EDA values already labelled as "Peak_Epoch" or "Storm", so, we leave to the user the decision to choose if they want to have "Peak_Epoch" and "Storm" not labelled also as "Artifact".

6.3 Labelling

By using the self-reported sleep and wake up time we labelled segmented windows before the time of sleeping and after the time of waking up as *Awake* the others to *Sleep*.

Regarding sleep quality we used only the data when participants were sleeping and we predicted sleep quality only in that time window even if, of course, we had for each whole block of sleep a singular quality label. This choice was made since we did not have enough data to use each whole session as train or test, in chapter 8 we discussed possible variant of this approach.

In self-reports we had a scale with five labels: *Very Poor*, *Poor*, *Normal*, *Good*, *Excellent*. In this case the dataset was really unbalanced so we decided to try also a classification of three classes *Low*, *Normal* and *High* by mapping the two highest (*Good* and *Excellent*) to *High* and the two lowest (*Very Poor* and *Poor*) to *Low*. For the same reason we tried also with two classes *High* and *Low* in this case *Very Poor*, *Poor* and *Normal* were mapped into *Low*.

6.4 Segmentation and feature extraction

The feature extraction is strictly related to the segmentation phase [Ploetz, 2021]. Since based on the segmentation we computed statistical features. For the segmentation we tried different values: 1 minute, 5 minutes and 10 minutes. Values that were chosen by looking at the current state-of-art of other research [Sano et al., 2018; Min et al., 2014].

Based on the segmentation window we then computed for EDA, ACC, ST: mean, standard deviation, sem (standard error of the mean of values within each group), maximum, minimum, median, variance, 7-quantiles.

Regarding ground-truth labels we assigned the window label at the majority in that window.

During data processing we labelled *artifact*, *peak epoch* and *storm* as a binary value: 1 as True and 0 as False.

Despite that, we decided to consider, during segmentation, Artifact, Peak_Epoch and Storm as sum. This way we were not losing information if in a time window there were just few values labelled. Accordingly, in the segmented dataset, the values of "Artifact", "Peak_Epoch" and "Storm" should not be considered as number of those events but more as number of segments of 250*ms* that are that kind of event. At the end we obtained a total of 59 feature.

6.5 Classification

We used as classifier XGBoost (Extreme Gradient Boosting) [1] [Chen and Guestrin, 2016] since is fast, and its performance are usually better than other classification algorithms [Qiang et al., 2018; Chollet, 2017].

XGBoost is a boosting method that by adding tree on top of other trees it is able to correct the errors of the previous one.

Classification tasks:

- Recognize sleep or awake segments
- Sleep Quality with 5 classes (*Very Poor*, *Poor*, *Normal*, *Good*, *Excellent*)
- Sleep Quality with 3 classes (*Low*, *Normal*, *High*)
- Sleep Quality with 2 classes (*Low*, *High*)

6.6 Evaluation

To evaluate models we split the data into train and test sets by using two different validation approaches: user-dependent and user-independent.

User-independent is a special case of k-fold cross validation where k is equal to the number of subjects, in our case k is equal to 16. One subject is selected as test data while the other subjects are used for training the model. This procedure was repeated until all the subjects have been used as test. In this way we ensured that our model does not contain subject bias, this is why it was used to test the generalizability of the model to a new unseen user. The evaluation metrics were computed as the mean between all the results obtained for each model that was tested with the left-out user.

User-dependent in this approach, for each user we selected a session as test (which is the 5th or more, when we ordered the sessions in ascending chronological order) and by using all the past sessions in the chronological order we predicted the selected one. This was repeated for each session with at least four previous sessions and for each user. This why we ensure that no future data will be used to predict past data, avoiding temporal leak Chollet [2017]. As like in the user-independent model, the evaluation metrics were computed as the mean between all the results obtained for all the evaluations, as describe before. This way we were testing the capability of the classifier to generalize to unseen data of a known user.

[1] XGBoost: `https://xgboost.readthedocs.io/en/latest/#`

6.6.1 Baselines

To better understand the goodness of our models we compared them with the following baselines:

- **Pittsburgh sleep quality index (PSQI)**: a rule-based prediction in which we used the time that subjects indicated, in the PSQI as their usual sleep onset and offset and their usual quality. For the user-independent model we used the mean sleep onset and offset time among all the users and the most frequent usual quality, based on the answer we received from PSQI.
- **Dummy stratified** [2] : A dummy classifier created by using sklearn and by indicating as strategy: stratified. In this way, the prediction is based on the dataset's class distribution.
- **Dummy most frequent** [3] : A dummy classifier created by using sklearn and by indicating as strategy: most_frequent. In this way the prediction will always be the most frequent label in the dataset.

After having the prediction labels and the self-reports labels we just computed the metrics as we describe in the following section.

6.6.2 Metrics

The problems we were facing are all classification problems therefore we decide to use as evaluation metrics: balance accuracy, accuracy, recall and precision. We believe that balance accuracy in case of unbalanced dataset is most representative than accuracy, since it is computed as the average of the correct ratio of each classes individually, so even if a class has more entries it will weight the same. Despite this we decided to include also the accuracy since most of the related work use just accuracy.

Having in mind the confusion matrix Figure 6.3 (in the figure it is shown a case of a binary problem but even with 5 classes it easy to scale to a binary confusion matrix) we can describe the evaluation measures as follow:

Accuracy [Gron, 2017] is the percentage of correctly classified instances. It is calculated as

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Recall [Gron, 2017] is also called positive predictive value (PPV) and it can be considered as how many of the actual positives are true positive. Recall is a crucial metric to look at when there is a high cost associated to a false negative (e.g., disease detection). It is computed as

$$Recall = \frac{TP}{TP + FN}$$

Precision [Gron, 2017] is also called true positive rate or sensitivity and can be defined as how precise and accurate the model is, by looking at how many of those predicted positive are actual positive. The formula to obtain this metric is

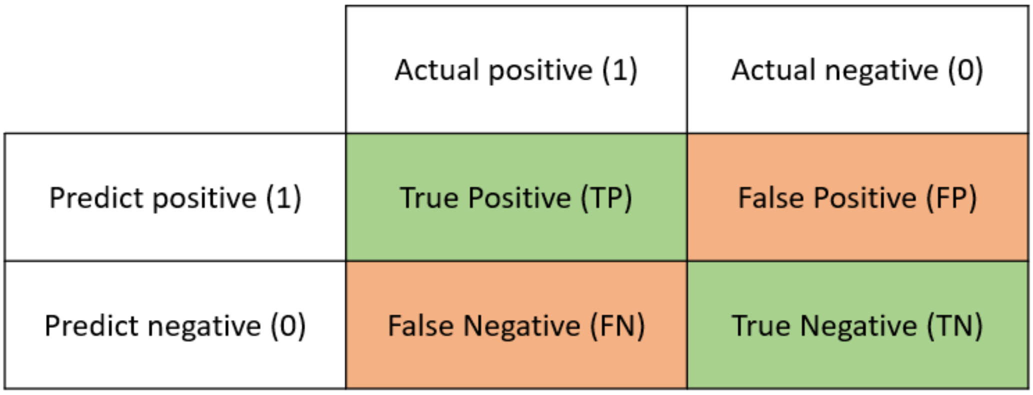
$$Precision = \frac{TP}{TP + FP}$$

[2] Sklearn dummy stratified: `https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html`

[3] See footnote 2

Balanced accuracy [Gron, 2017] explains as a percentage how good a classifier is by also taking into account the classes balance. It is computed as

$$Balanced\ accuracy = \frac{1}{2} * \frac{TP}{TP + FN} + \frac{1}{2} * \frac{TN}{FP + TN}$$



	Actual positive (1)	Actual negative (0)
Predict positive (1)	True Positive (TP)	False Positive (FP)
Predict negative (0)	False Negative (FN)	True Negative (TN)

Figure 6.3. Example of a confusion matrix

Chapter 7

Results

In this chapter the results obtained, by using the procedure explained before in chapter 6, are presented.

7.1 Sleep vs Awake

Table 7.1 shows the results we obtained using our models to detect sleep/awake segments. The overall results are very promising. In particular for windows of 5 and 10 minutes in all models the balanced accuracy is above 90%, the reason why these time windows are better can be that larger time windows are more capable to catch sleep patterns. We firstly made comparison with 1-minute window, when we aspect to understand better how different sensors and features contributes to the final results and then we continue the evaluation with 5 and 10 minutes only with the model that use all the sensors.

Performance sleep/awake between the user-independent and user-dependent model are very close, this can suggest that sleep patterns are usually very similar among all participants.

For 10-minute window model we also present the results to a model without artifact, in this model window labelled as artifact had EDA value (filtered, phasic and tonic) turn into 0. Anyway the results still remain pretty much unchanged.

7.1.1 Comparison between features

As we can see in Table 7.1 on a 1-minute window the difference between the model with only EDA and only ACC is of 10% in the user-independent model, this observation are in line with previous finding in literature where there is a clear primacy of accelerometer features above others (balanced accuracy with only ACC is 88.11% against the 78.13% of only EDA). The same difference is also in the user-dependent model when ACC obtains 87.99% and EDA acquires 77.32%.

Also, distal skin temperature gave a good balance accuracy: 79.64% in the user-independent and 80.73% in the user-dependent.

Along all the different time windows there is no big difference between adding also storm, peak epoch and artifact in some cases balance accuracy is slightly better when there are also peak epoch labels, however the difference is never above 0.10% so it is statistical insignificance.

7.1.2 Comparison with baselines

We observed that the baseline computed by using PSQI has a high recall value in the user-dependent model this probably indicates that users had strong sleep routine and their usual sleep onset is in average before their real onset and reported offset is usually later than the actual one, therefore there are very few false negative.

Except for the recall we were always able to beat all the baselines in almost all the time windows, the baseline harder to defeat was the one based on the PSQI especially by looking at the user-dependent model the differences are of 2% or 3%.

	Model	User-independent Model				User-dependent Model			
		Accuracy	Balance Accuracy	Recall	Precision	Accuracy	Balance Accuracy	Recall	Precision
1 min	EDA	77.97%	78.13%	78.86%	78.01%	77.57%	77.32%	74.93%	77.29%
	EDA+Artifact	77.94%	78.10%	78.94%	77.88%	77.52%	77.34%	75.09%	77.68%
	EDA+Peak Epoch	78.09%	78.24%	78.69%	78.25%	77.68%	77.49%	75.29%	77.73%
	EDA+Storm	78.01%	78.17%	78.79%	78.09%	77.61%	77.36%	75.05%	77.20%
	TEMP	79.45%	79.64%	84.56%	78.13%	81.34%	80.73%	83.17%	78.02%
	ACC	88.02%	88.11%	88.36%	88.02%	88.29%	87.99%	85.77%	87.96%
	ACC+TEMP	88.68%	88.78%	89.45%	88.39%	89.91%	89.37%	88.24%	88.66%
	EDA+ACC+TEMP	89.64%	89.75%	90.36%	89.28%	90.57%	89.93%	88.77%	89.26%
	EDA+ACC+TEMP+Artifact	89.64%	89.75%	90.23%	89.36%	90.50%	89.88%	88.80%	89.19%
	EDA+ACC+TEMP+Peak Epoch	89.69%	89.80%	90.22%	89.44%	90.46%	89.78%	88.58%	89.16%
	EDA+ACC+TEMP+Storm	89.66%	89.77%	90.36%	89.31%	90.51%	89.91%	88.85%	89.18%
5 min	EDA+ACC+TEMP	90.04%	90.14%	90.73%	89.58%	91.09%	90.58%	**90.14%**	89.47%
	EDA+ACC+TEMP+Artifact	90.07%	90.18%	90.56%	89.79%	91.10%	90.59%	90.05%	89.48%
	EDA+ACC+TEMP+Peak Epoch	90.14%	90.25%	90.70%	89.81%	91.09%	90.58%	90.07%	89.47%
	EDA+ACC+TEMP+Storm	90.00%	90.11%	90.62%	89.63%	91.14%	**90.63%**	90.10%	89.56%
10 min	EDA+ACC+TEMP	**90.48%**	**90.58%**	91.16%	**89.99%**	91.16%	90.61%	89.86%	89.65%
	EDA+ACC+TEMP (without Artifacts)	90.39%	90.51%	**91.51%**	89.62%	91.03%	90.62%	90.03%	89.56%
	EDA+ACC+TEMP+Artifact	90.22%	90.32%	90.85%	89.77%	91.11%	90.55%	89.76%	89.63%
	EDA+ACC+TEMP+Peak Epoch	90.44%	90.55%	91.11%	89.97%	91.10%	90.55%	89.76%	89.65%
	EDA+ACC+TEMP+Storm	90.37%	90.49%	91.27%	89.74%	**91.18%**	**90.63%**	89.79%	**89.74%**
	SHAP_top_20	90.31%	90.42%	90.73%	90.00%	90.85%	90.35%	89.69%	89.29%
	Storm + Peak Epoch	54.08%	54.49%	84.02%	52.74%	57.78%	57.78%	54.77%	59.79%
Baseline (prediction with time of PSQI)		85.87%	85.86%	88.73%	83.95%	88.83%	88.81%	**92.41%**	86.57%
Baseline (Dummy stratified)		49.99%	49.99%	49.94%	50.04%	50.15%	50.00%	50.10%	50.11%
Baseline (Dummy most frequent)		50.05%	50.00%	**100.00%**	50.05%	52.14%	50.00%	43.75%	23.00%

Table 7.1. Sleep vs Awake. SHAP_top_20 is the model obtained by using only the first 20 features that were more important according to SHAP Figure 7.3.

7.2 Sleep Quality

Contrary to what we observed in sleep/awake here the difference between user-dependent and user-independent model is more tangible, this can be due to two different hypotheses: physiological characteristics of sleep quality are more individual and personal so can vary a lot along people, or sleep quality is more subjective so even its definition is not trivial, for example one user can be defined sleep quality as feeling rested and another as no wake up during the night.

We observed that in sleep quality detection user-dependent balance accuracy is always higher than user-independent one:

- in sleep quality with five classes user-dependent balance accuracy is higher by 17 percentage points
- in sleep quality with three classes user-dependent balance accuracy is higher by 13 percentage points
- in sleep quality with two classes user-dependent balance accuracy is higher by 12 percentage points

Patterns between the different sleep quality problems, so with two, three and five classes, are very similar to each other so, when it is not indicated which problem we are refereeing to, statements are referred to all. As we expected, even if patterns are very similar, there is a clear increase of the results goodness when we reduce the number of classes.

For 10-minute window model, as in the sleep/awake problem, we also present the results to the model without artifact, in this model window labelled as artifact had EDA value (filtered, phasic and tonic) turn into 0. Again, the results obtained are still no so different from the one without any knowledge of *Artifact* (balance accuracy user-dependent model without knowledge of *Artifact*: 62.63%, balance accuracy user-dependent model with *Artifact* segments to 0: 61.25%), this can be due to short time windows that by containing few artifacts, they do not impact enough the final result.

The results presented could be seeing as far worse than the ones obtained in the sleep/awake problems, nevertheless the dataset is really unbalanced with five classes and three classes, results are above 60% in the user-dependent model and by considering only two classes.

7.2.1 Comparison between features

As we can see by looking at Table 7.4, Table 7.3 and Table 7.2 in the 1-minute time window, electrodermal activity achieve the highest balance accuracy value in comparison with ACC and TEMP, even if in some cases, as in sleep quality with three classes, TEMP performed very closely to EDA.

Between the models with knowledge of artifacts, peak epochs or storms there is no remarkable difference, not in the user-dependent model nor in the user-independent one. This can be due to high noise from all the features that hinders the final impact on all the model.

To better understand how peak epochs and storms impact the final model we decided to run a model with just these two features, the results for sleep/awake were not better as expected, since they alone are not representative. At the contrary, the model with only storms and peak epochs in high/low sleep quality performed better than other models (balance accuracy user-dependent: 65.47%, user-dependent EDA + ACC + TEMP: 62.63%). These results are very promising, a future furthermore investigation could better understand the correlation between these features and perceived sleep quality.

7.2.2 Comparison with baselines

User-independent baselines created with PSQI and the most frequent baseline are equal, this means that the most frequent PSQI is also the most frequent answer in self-reports.

Along all variants the baselines are just faintly worse than our models.

Model		User-independent Model				User-dependent Model			
		Accuracy	Balance Accuracy	Recall	Precision	Accuracy	Balance Accuracy	Recall	Precision
1 min	EDA	38.58%	29.26%	38.58%	44.78%	47.35%	47.15%	47.35%	87.69%
	EDA+Artifact	38.56%	29.07%	38.56%	44.64%	47.45%	47.27%	47.45%	87.45%
	EDA+Peak Epoch	38.97%	**29.64%**	38.97%	45.15%	47.50%	47.31%	47.50%	86.80%
	EDA+Storm	38.59%	29.26%	38.59%	44.59%	47.30%	47.09%	47.30%	87.66%
	TEMP	**39.26%**	28.27%	**39.26%**	41.53%	**48.17%**	**47.96%**	**48.17%**	**98.11%**
	ACC	38.21%	28.38%	38.21%	43.34%	45.57%	45.34%	45.57%	96.21%
	ACC+TEMP	38.04%	28.07%	38.04%	43.47%	45.65%	45.36%	45.65%	96.47%
	EDA+ACC+TEMP	37.93%	28.58%	37.93%	43.11%	46.73%	46.49%	46.73%	90.28%
	EDA+ACC+TEMP+Artifact	38.27%	28.77%	38.27%	43.33%	47.38%	47.11%	47.38%	91.27%
	EDA+ACC+TEMP+Peak Epoch	38.08%	28.57%	38.08%	44.72%	47.05%	46.80%	47.05%	91.56%
	EDA+ACC+TEMP+Storm	38.12%	28.80%	38.12%	42.90%	47.33%	47.07%	47.33%	91.12%
5 min	EDA+ACC+TEMP	38.41%	28.89%	38.41%	43.44%	47.20%	46.86%	47.20%	87.42%
	EDA+ACC+TEMP+Artifact	38.21%	28.37%	38.21%	43.42%	47.14%	46.82%	47.14%	87.23%
	EDA+ACC+TEMP+Peak Epoch	38.09%	28.40%	38.09%	**45.93%**	47.03%	46.76%	47.03%	87.36%
	EDA+ACC+TEMP+Storm	37.85%	28.41%	37.85%	43.09%	47.22%	46.91%	47.22%	87.83%
10 min	EDA+ACC+TEMP	38.39%	28.89%	38.39%	42.66%	46.90%	46.54%	46.90%	85.87%
	EDA+ACC+TEMP+Artifact	38.61%	28.97%	38.61%	43.61%	47.00%	46.67%	47.00%	85.94%
	EDA+ACC+TEMP+Peak Epoch	38.27%	28.23%	38.27%	44.16%	47.35%	47.06%	47.35%	85.89%
	EDA+ACC+TEMP+Storm	38.38%	28.24%	38.38%	43.21%	47.08%	46.72%	47.08%	85.37%
Baseline (Predict with usual quality in PSQI)		**42.43%**	20.00%	**42.43%**	18.00%	46.95%	29.27%	46.95%	26.19%
Baseline (Dummy stratified)		33.61%	20.03%	33.61%	33.62%	44.32%	29.36%	44.32%	44.30%
Baseline (Dummy most frequent)		**42.43%**	20.00%	**42.43%**	18.00%	**56.77%**	29.27%	**56.77%**	33.93%

Table 7.2. Sleep Quality with 5 classes Excellent, Good, Normal, Poor and Very Poor

Model		User-independent Model				User-dependent Model			
		Accuracy	Balance Accuracy	Recall	Precision	Accuracy	Balance Accuracy	Recall	Precision
1 min	EDA	41.52%	35.57%	41.52%	48.49%	50.49%	50.22%	50.49%	89.20%
	EDA+Artifact	**41.80%**	35.97%	**41.80%**	**48.73%**	50.50%	50.24%	50.50%	88.28%
	EDA+Peak Epoch	41.74%	36.00%	41.74%	48.47%	**50.65%**	**50.37%**	**50.65%**	89.17%
	EDA+Storm	41.75%	**36.12%**	41.75%	48.61%	50.37%	50.07%	50.37%	89.17%
	TEMP	38.69%	35.12%	38.69%	46.12%	50.22%	49.96%	50.22%	**97.63%**
	ACC	39.74%	34.33%	39.74%	46.22%	47.52%	47.23%	47.52%	96.19%
	ACC+TEMP	39.39%	33.92%	39.39%	44.50%	47.85%	47.53%	47.85%	96.20%
	EDA+ACC+TEMP	40.72%	34.72%	40.72%	46.98%	49.68%	49.34%	49.68%	91.55%
	EDA+ACC+TEMP+Artifact	40.76%	34.97%	40.76%	47.22%	50.28%	49.94%	50.28%	92.26%
	EDA+ACC+TEMP+Peak Epoch	41.18%	35.41%	41.18%	47.29%	49.92%	49.58%	49.92%	92.53%
	EDA+ACC+TEMP+Storm	40.87%	35.13%	40.87%	47.14%	50.18%	49.85%	50.18%	92.21%
5 min	EDA+ACC+TEMP	41.04%	35.94%	41.04%	47.07%	49.96%	49.54%	49.96%	88.90%
	EDA+ACC+TEMP+Artifact	40.78%	34.92%	40.78%	46.75%	49.91%	49.55%	49.91%	88.49%
	EDA+ACC+TEMP+Peak Epoch	41.40%	35.34%	41.40%	47.66%	49.78%	49.43%	49.78%	88.96%
	EDA+ACC+TEMP+Storm	41.17%	35.57%	41.17%	47.50%	50.13%	49.74%	50.13%	90.20%
10 min	EDA+ACC+TEMP	41.04%	34.99%	41.04%	46.63%	49.60%	49.22%	49.60%	86.08%
	EDA+ACC+TEMP+Artifact	40.92%	35.30%	40.92%	46.74%	49.65%	49.28%	49.65%	86.02%
	EDA+ACC+TEMP+Peak Epoch	40.97%	35.05%	40.97%	46.57%	50.01%	49.58%	50.01%	86.86%
	EDA+ACC+TEMP+Storm	41.30%	35.63%	41.30%	47.27%	49.76%	49.31%	49.76%	85.71%
Baseline (Predict with usual quality in PSQI)		**42.43%**	33.33%	**42.43%**	18.00%	48.58%	35.42%	48.58%	27.93%
Baseline (Dummy stratified)		37.61%	33.27%	37.61%	37.62%	46.89%	35.21%	46.89%	46.86%
Baseline (Dummy most frequent)		**42.43%**	33.33%	**42.43%**	18.00%	**59.06%**	35.42%	**59.06%**	36.30%

Table 7.3. Sleep Quality with 3 classes High, Normal and Low

Model		User-independent Model				User-dependent Model			
		Accuracy	Balance Accuracy	Recall	Precision	Accuracy	Balance Accuracy	Recall	Precision
1 min	EDA	56.99%	50.27%	22.73%	40.87%	63.67%	63.60%	23.17%	37.53%
	EDA+Artifact	57.12%	50.53%	22.91%	**40.89%**	63.74%	63.68%	23.45%	38.05%
	EDA+Peak Epoch	57.09%	50.48%	22.86%	**40.89%**	63.44%	63.38%	23.09%	37.79%
	EDA+Storm	56.82%	50.62%	23.82%	40.76%	63.26%	63.19%	23.16%	37.80%
	TEMP	57.80%	48.89%	7.08%	32.70%	63.66%	63.70%	22.61%	**39.63%**
	ACC	53.13%	48.91%	28.07%	36.93%	60.71%	60.71%	21.89%	39.29%
	ACC+TEMP	53.31%	48.90%	30.19%	37.02%	60.92%	60.88%	21.78%	39.02%
	EDA+ACC+TEMP	54.22%	49.89%	28.64%	38.60%	62.86%	62.78%	22.95%	38.69%
	EDA+ACC+TEMP+Artifact	54.58%	50.27%	29.32%	38.75%	62.95%	62.85%	22.89%	38.76%
	EDA+ACC+TEMP+Peak Epoch	54.80%	50.29%	28.87%	39.03%	62.69%	62.57%	22.94%	38.54%
	EDA+ACC+TEMP+Storm	54.54%	50.15%	28.80%	38.86%	62.48%	62.35%	22.80%	38.67%
5 min	EDA+ACC+TEMP	54.48%	49.97%	29.14%	38.82%	62.70%	62.62%	22.92%	37.77%
	EDA+ACC+TEMP+Artifact	53.88%	50.32%	31.00%	38.62%	62.75%	62.74%	22.97%	37.80%
	EDA+ACC+TEMP+Peak Epoch	54.71%	50.37%	29.58%	38.87%	62.66%	62.61%	22.98%	37.77%
	EDA+ACC+TEMP+Storm	54.44%	50.19%	29.94%	38.99%	62.61%	62.55%	22.96%	38.03%
10 min	EDA+ACC+TEMP	54.36%	**51.27%**	**33.02%**	39.38%	62.61%	62.63%	23.26%	36.99%
	EDA+ACC+TEMP (without Artifacts)	54.01%	49.71%	29.38%	38.41%	61.40%	61.25%	22.20%	37.99%
	EDA+ACC+TEMP+Artifact	53.83%	49.07%	28.59%	38.41%	62.44%	62.45%	23.09%	37.00%
	EDA+ACC+TEMP+Peak Epoch	53.99%	50.15%	30.27%	38.78%	62.26%	62.24%	22.64%	36.74%
	EDA+ACC+TEMP+Storm	54.26%	49.55%	29.01%	39.13%	62.40%	62.43%	23.09%	36.73%
	SHAP_top_20	54.25%	49.90%	29.49%	38.83%	61.46%	61.46%	22.76%	36.99%
	Storm + Peak Epoch	**60.61%**	49.89%	3.38%	37.90%	**65.55%**	**65.47%**	**24.94%**	37.43%
Baseline (Predict with usual quality in PSQI)		58.70%	50.00%	0.00%	0.00%	63.89%	50.00%	**31.25%**	16.61%
Baseline (Dummy stratified)		51.61%	50.10%	**41.42%**	**41.42%**	62.91%	50.16%	38.22%	38.31%
Baseline (Dummy most frequent)		**58.70%**	50.00%	0.00%	0.00%	**71.52%**	50.00%	**31.25%**	20.43%

Table 7.4. Sleep Quality with 2 classes High and Low. SHAP_top_20 is the model obtained by using only the first 20 features that were more important according to SHAP Figure 7.5

7.3 Features

To really understand the best features that concretely help the model in its predictions, and in what extent, we decided to use SHAP (SHapley Additive exPlanations). SHAP was first presented in 2017 by Lundberg and Lee [2017], is an interpretability method that uses Shapley values. A Shapley value is a game theory concept that can be defined as a measure for the marginal contribution of a feature after all possible combinations have been considered.

By looking at those Shapley values, computed for each feature, we are able to examine the decision making of our model.

There are different things and different visualizations obtainable from SHAP, some are more general and others goes deeper, we will first look at a bar plot showing for each feature the mean absolute value of the Shapley values and then we will go deeper by looking also at how each feature contributes to every sample in an aggregate way.

To understand further our explanations and hypothesis we use SHAP only with binary problem, so sleep/awake and high/low sleep quality, and only with a segmentation of 10 minutes.

In the bar plot (e.g., Figure 7.5) for each feature we have the average impact on model outputs, a higher value means a higher impact.

In the other plot (e.g., Figure 7.6) y-axis shows the feature and x-axis shows the Shapley value for each instance. In case of point overlapping, dots are pile up along y-axis so the sense of the distribution is not lost. Each point represent a different instance, whose colour represents the feature value (from low, blue, to high, red). Positive Shapley value means higher contribution of that feature to a positive (1, "Sleep"/"High") output value. At the contrary, more in that instance that feature has a Shapley value under 0 more it will contribute to a negative (0, "Awake"/"Low") answer.

In Figure 7.3 and Figure 7.5 we show the top 20 features, in Figure 7.6 and Figure 7.4 there are the same 20 top features but we will look at how they contribute, which feature explains better sleep or awake segments, and if they are positive correlated or not, e.g., high negative value of a feature indicates a sleep event.

To understand how much the information of the user will impact the final prediction, we also show in Figure 7.2 and Figure 7.1 how sleep and quality detection changed based on SHAP evaluation. As we can see adding the knowledge of the user, place this feature in first position in high/low sleep quality detection and in third position for sleep/awake recognition. This confirmed what we also observed in Table 7.1 and Table 7.4 where the user-dependent model in sleep quality recognition is always better than the independent one.

7.3.1 Sleep vs Awake

By looking at Figure 7.3 it is clear that just ACC features are enough to classify sleep and awake with a good degree of certainty, which is perfectly in line with our results. In the top 20 there is also three features obtained from TEMP and five from EDA. The latter are from EDA values filtered, phasic and tonic components even if among this five only one is an EDA filtered feature.

Regarding patterns in features values, we have to look at Figure 7.6 and we understand that, for example a high value of *Y_sem* (we remind that sem is a standard deviation not for the values in that window but the standard deviation of that window compare to the entire session mean value) is usually an indicator of an awake segments, at the contrary a low value is related to a sleep segment; this pattern is the same for *Z_sem* and *X_sem*.

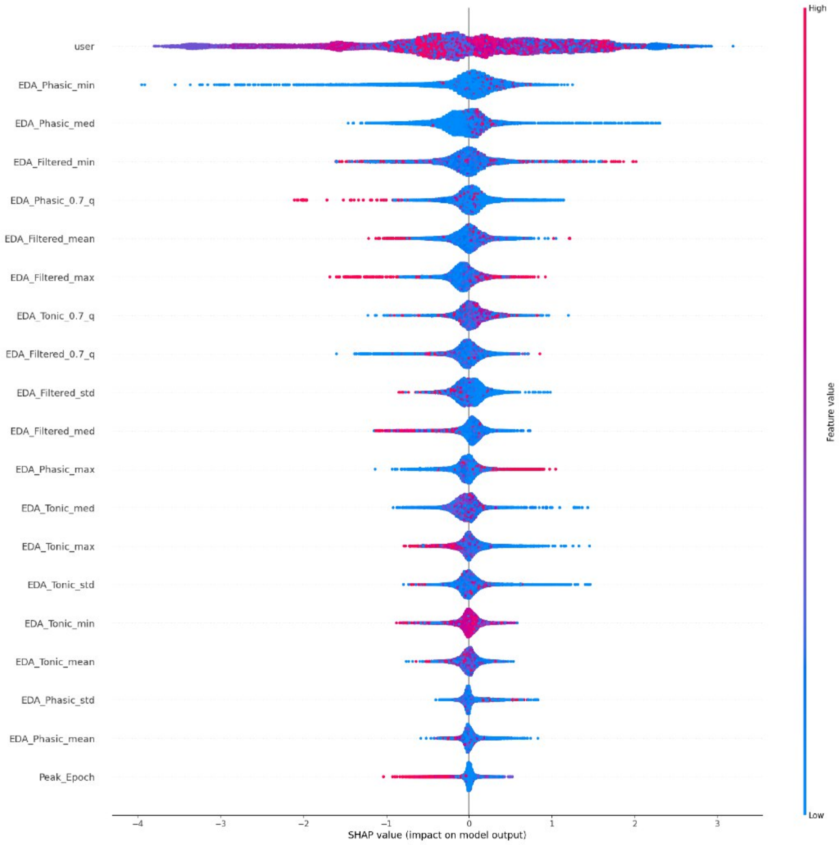


Figure 7.1. Shap plot of binary sleep quality classification with the knowledge of which user we are using as test

We also notice that for *EDA_Phasic_med* and *EDA_Phasic_0.7_q* a high value is related to an awake window and low values indicate a sleep event; the opposite of EDA min, max and standard deviation where a low value indicates sleep.

7.3.2 Sleep Quality

From Figure 7.5 it is clear that for the binary sleep quality problem EDA has a key role: 13/20 are EDA features. In the top 10 there are also two features obtained from TEMP, instead ACC has five features in top 20 but all in low positions.

In detail, as we can see in Figure 7.4, the high or low features values are very difficult to interpret, this could suggest that there is not a clear predominance but more a combination of

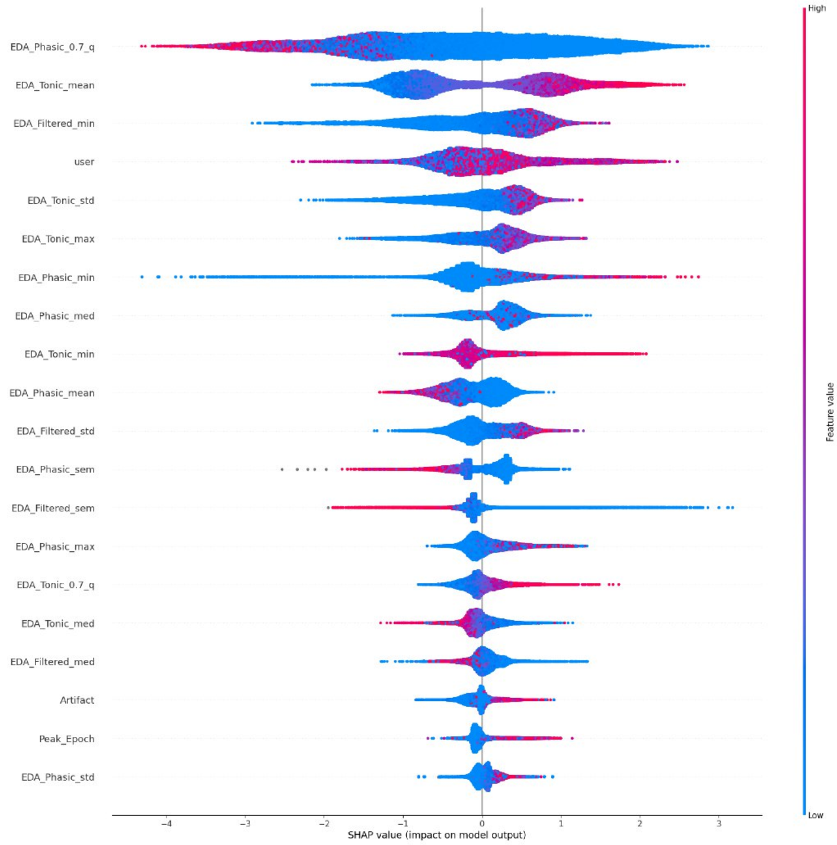


Figure 7.2. Shap plot of Sleep/Awake classification with the knowledge of which user we are using as test

different features.

We are able just to reason on *TEMP_max*, *TEMP_std*, *EDA_Filtered_max* and *EDA_Phasic_max* whose high values are associated to a higher sleep quality and lower values indicate a low sleep quality.

7.4 Comparison with related work

Results taken from literature are presented in Table 7.5. Higher accuracy in sleep/awake problem is obtained from Min et al. [2014], by using just phone sensors they were able to achieve in the individual model 94.52%. Among all the accuracy values in literature, they are very similar

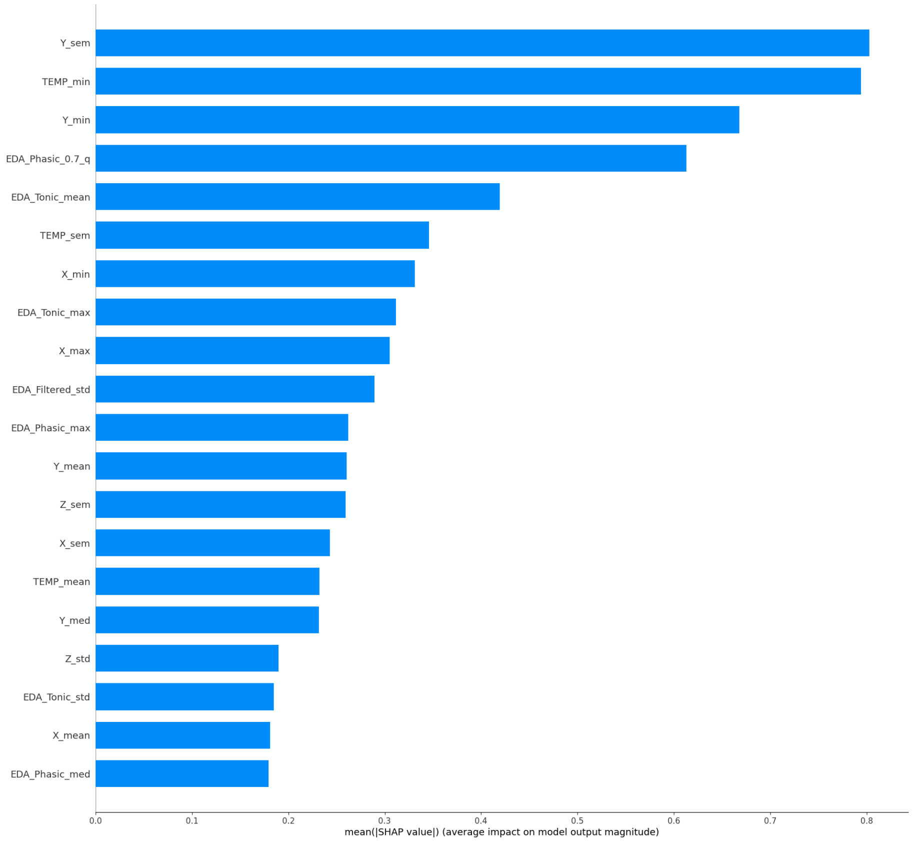


Figure 7.3. Shap bar plot of Sleep/Awake classification

to the one we obtained (90.58%).

Only few studies focus on sleep quality as daily measure. For example Sano et al. [2015] try to infer the PSQI score and not a daily sleep quality.

Sano et al. [2015] and Min et al. [2014] reach an accuracy in sleep quality detection above 80%, as we mentioned also in the previous section, sleep quality is still challenging since its definition vary between different studies.

Sano and Picard [2014] conclude that the combination of ACC and TEMP played a key role in Sleep/Awake classification, we also confirm that since in Figure 7.3 the first three features are from ACC and TEMP.

We have also found that skin temperature tends to increase in the first phase of sleep (as can be infer from Figure 7.6), which is consistent with the previous finding [Kräuchi et al., 2000].

Data collection was particularly interesting since, during that 30 days, 28 of March was full

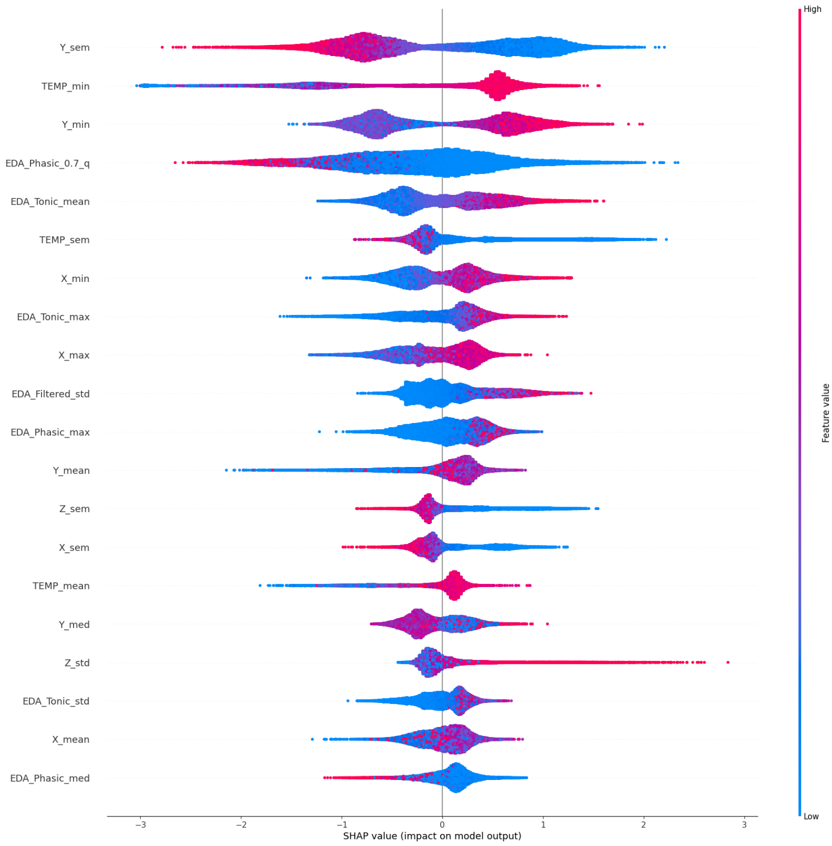


Figure 7.4. Shap plot of Sleep/Awake classification

moon and also time zone changed to summer time (+1 hour), Casiraghi et al. [2021] observed that "on nights before a full moon, people go to bed later and sleep less", by looking at Figure 5.4 we can see clearly that except for the very previous day of the full moon we did not observed a strong trend on a late sleep onset. Differently, a reduction on the sleep time on days before was observed also in our study, as we can see in Figure 5.6.

Paper	**Problem**	**Features**	**Approach**	**Results**
Wang et al. [2017]	Sleep/Awake	4 smartphone features: light, phone usage, activity, microphone.	Linear combination. Validation approach: NA	95% of the inferences had an accuracy of ± 25 minutes

Min et al. [2014]	Sleep/Awake	Phone sensors in a 10-minute windows: an accelerometer, microphone, ambient light sensor, screen proximity sensor, running process, battery state, and display screen state	Bayesian Network. Validation approach: leave-one-user-out cross validation for general and leave-one-day-out cross validation for individual	Accuracy general (same as our user-independent model): 93.06%, accuracy individual (same as our user-dependent model): 94.52%
Min et al. [2014]	Sleep Quality (not classifying daily quality but were detecting good and poor sleepers)	Phone sensors in a 10-minute windows: an accelerometer, microphone, ambient light sensor, screen proximity sensor, running process, battery state, and display screen state	Bayesian Network. Validation approach: leave-one-subject-out (LOSO) cross validation for general and leave-one-day-out (LODO) cross validation for individual	Accuracy: 83.97%
Sano et al. [2018]	Sleep/Awake	Combination of smartphone and wristband sensors	Neural networks with long-short term memory (LSTM) cells. Validation approach: 5-fold cross-validation	Accuracy: 96.5%
Sadeghi et al. [2019]	Sleep Quality	Wristband sensors: heart rate variability, electrodermal activity, body movement and skin temperature	Random forest. Validation approach: NA	Accuracy: 75%

Sano et al. [2015]	Sleep Quality (PSQI score)	Combination of smartphone and wristband sensors	Support vector machine (SVM). Validation approach: leave-one-subject-out (LOSO)	Accuracy: 88%
Zhai et al. [2020]	Sleep/Awake	Heart rate and actigraphy	Convolutional neural networks (CNNs). Validation approach: 5-fold cross-validation	Accuracy: 84.4% ± 1 standard error at 95% confidence interval
Sano and Picard [2014]	Sleep/Awake	ACC+SC+TEMP	Support vector machine (SVM). Validation approach: NA	Intra-subject accuracy: 86%, inter-subject accuracy: 74%.
Guo [2016]	Sleep/Awake	Phone sensors in a 10-minute windows	Random forest. Validation approach: 10-fold cross-validation	Accuracy: 95.48%

Table 7.5. Results from related work. If validation technique is not indicate we use "NA" as "Not Available"

7.5 Commercial devices

In pre-study we asked participants if they were used to wear some commercial devices that track sleep and seven participants answered yes. Before the end of the study we contacted those participants to ask them if they could share with us the sleep tracked by those devices during the study. Four participants have agreed to share their data with us, of which two used MiBand, one used FitBit and two wore Garmin.

Table 7.6 shows the result that we obtained by using sleep and awake time from their commercial wearable devices. Since MiBand provides also a sleep quality score (range from 0 to 100) we mapped their score to our scale (1-5) and we compute the same metrics we used in our study. Table 7.7 shows the results for sleep quality problem

MiBand tracks also nap, but it does not provide sleep quality in case of naps. Therefore, we decided to use also naps onset and offset to evaluate sleep/awake problems but we used only night sleeps to evaluate sleep quality. Another issue was that in our study for each sleep segments (from sleep onset to sleep offset) we had a score, MiBand instead give a score for all night even if user woke up during night. So in case of multiple sleep qualities during a night we compared the mean of those values as ground-truth and we compered it with the sleep score provided by MiBand.

For a better understand of these results we provide also the distribution of the classes.

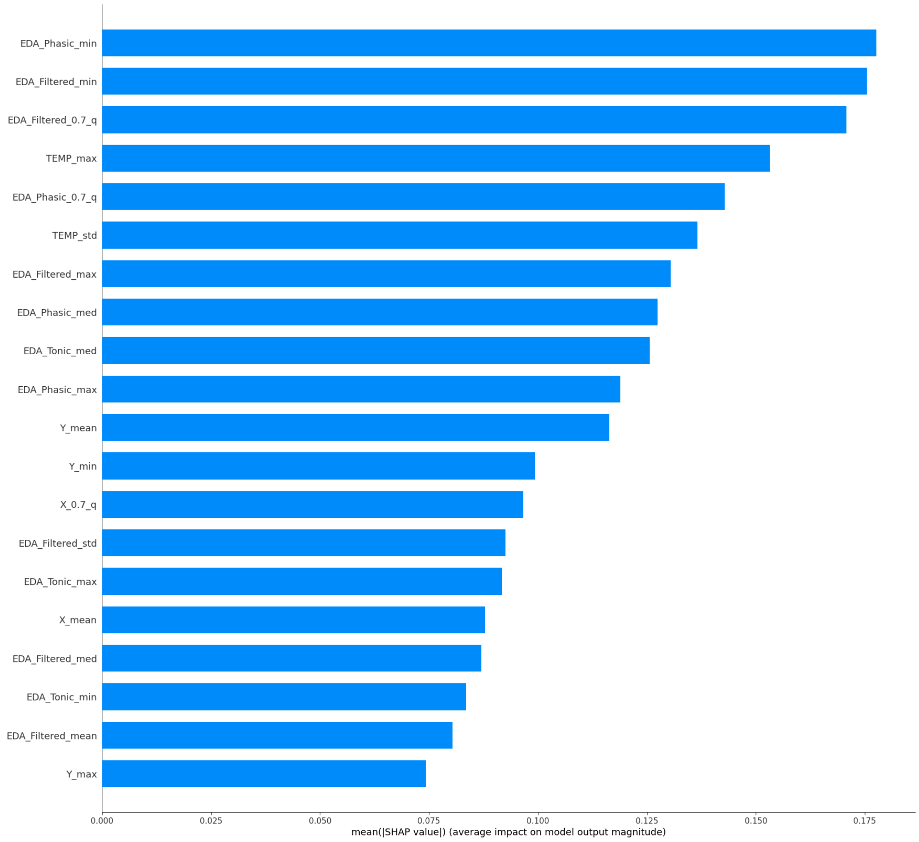


Figure 7.5. Shap bar plot of binary Sleep Quality

Device	Accuracy	Balance Accuracy	Recall	Precision	Distribution of classes
MiBand1	97.75%	97.82%	97.75%	97.77%	Sleep: 47%, Awake: 53%
MiBand2	96.42%	96.41%	97.08%	95.95%	Sleep: 50%, Awake: 50%
FitBit	97.90%	97.89%	99.73%	96.22%	Sleep: 50%, Awake: 50%
Garmin1	93.80%	93.59%	97.18%	91.59%	Sleep: 53%, Awake: 47%

Garmin2	92.69%	92.82%	99.62%	87.29%	Sleep: 49%, Awake: 51%
Our approach	**91.16%**	**90.61%**	**89.86%**	**89.65%**	Sleep: 49%, Awake: 51%

Table 7.6. Performances of commercial devices for sleep

Device	Accuracy	Balance Accuracy	Recall	Precision	Distribution of classes
MiBand1	6.45%	5.71%	6.45%	47.31%	Very Poor: 0% , Poor: 23%, Normal: 32%, Good: 45%, Excellent: 0%
MiBand2	41.38%	17.14%	41.38%	23.17%	Very Poor: 3% , Poor: 14%, Normal: 14%, Good: 21%, Excellent: 48%
Our approach	**46.90%**	**46.54%**	**46.90%**	**85.87%**	Very Poor: 2% , Poor: 15%, Normal: 42%, Good: 36%, Excellent: 5%

Table 7.7. Performances of commercial devices for sleep quality (with five classes)

MiBand and Fitbit reach higher balance accuracy (MiBand: 97.75% and 96.42%, Fitbit: 97.90%) than Garmin (93.80% and 92.69%). However, sleep quality was very difficult to infer even for MiBand commercial devices (balance accuracy of 5.71% and 17.14%).

Our approach had similar performance, or in some cases slightly worse than commercial devices, in infer sleep/awake times. This can be also due to less training data that we had in our approach in comparison to commercial devices that were worn for more time.

In the sleep quality problem with five classes our approach performed better than the commercial ones.

7.6 Main findings

To sum up:

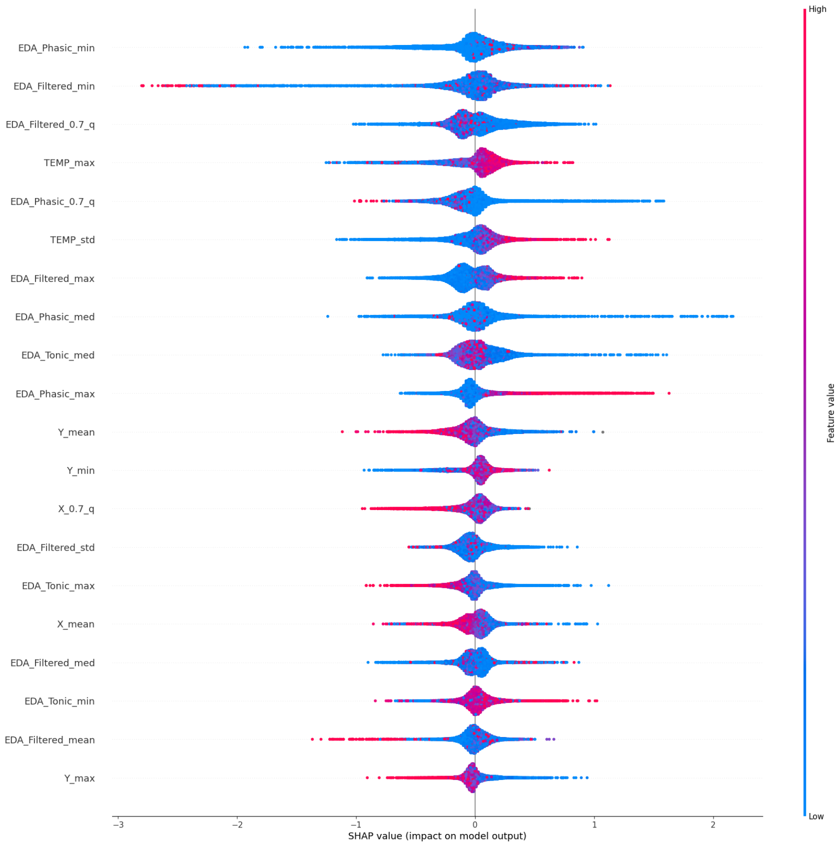


Figure 7.6. Shap plot of binary Sleep Quality

Sleep vs Awake Overall, ACC plays an important role in detecting sleep/awake segments (balance accuracy of 88.11% in 1 minute time window user-independent model), as also remarked by previous works. Anyway, we were able to increase this accuracy by using also TEMP and EDA, that, as also reported by SHAP, are still important for the final result. Best balance accuracy value was achieved by user-dependent 10 minutes window model (EDA + TEMP + ACC) with also the labels of storms: 90.63%.

Sleep Quality User-dependent model much better than user-independent, this can be because sleep quality depends on individual's interpretation. As expected, reducing the number of classes to predict results in an increase of the overall performances. By looking at SHAP plots we understood that EDA features are important in detecting sleep quality (in the top 20 features 13 are EDA features). Also TEMP seems important, along the three sensors only ACC seems the less important. The model that use just peak epochs and storms to infer high/low sleep quality performed almost three percentage points better than the model with all the features.

Chapter 8

Limitation and ideas for future work

While there is a lot of potential, there are some limitations, at each one we provided a possible solution or at least a hint on where to start looking for a solution:

- Data collection:
  - We keep self-reports as simple as possible to avoid overload users while they are tired (before sleeping) or when they are newly awake (right after sleeping). Anyway, might be worthwhile to allow users to report also other kind of information, such as "still awake, struggle to fall asleep", "I was woken up", "Wake up suddenly due to nightmares" and so on.
  - Seeing the not so good results for sleep quality prediction, we could try to not predict a sleep quality value but instead predict "Yes" or "No" answers at question as: "do you feel rested?", "do you feel as you have enough energy to start your day?", "do you struggle to fall asleep?".
  - Sleep is not an easy event to track since we are not aware when we really fall asleep so a way to increase the reliability of self-reports could be obtained by combining self-reports with other methods (e.g., accelerometer or phone usage).
- Data analysis:
  - As explained before, during this work missing data were turning into 0 values for each sensor. Maybe valuating if model performance improves by choosing a different data imputation strategy, for instance by replacing missing data with a mean value under certain circumstances (e.g., EDA mean computed only from segments without artifacts) can lead to better results [Jadhav et al., 2019].
  - We used three different sensors: EDA, ACC, ST. But we only applied a filter function to EDA, an improvement should come from using also filtering methods for ACC, for example by removing all little movements (e.g., applying a high pass filtering)[Bujari and Licar, 2012].
  - In our work we used XGBoost since it is also very used with physiological signals, anyway each machine learning models has its pros and cons so finding the one that can provide us the best results in our specific case can increase overall performance.

- - We worked with physiological signals that are in time domain, nonetheless extracting features also from frequency domain can be a good further investigation.
  - Regarding feature extraction, another improvement can come by computing also heart feature (e.g., heart rate) and breath pattern – extractable from PPG as explained by Muniyandi and Soni [2017].
  - Physiological signals are computed as time series which means that each value is in a certain way connected to the previous one. With common methods of machine learning this information is totally discarded since to predict a label those models look at features' values in that time moments. This problem can be overcome usually by using deep learning methods as long short-term memory networks (LSTM).
  - We used phone data only to check sleep diaries, an idea could be combined them with wearable data to infer sleep and sleep quality as Martinez et al. [2020] did.

- In our work we had a unique model for any kind of sleep: nightly sleep and nap. Though it seems reasonable that they have probably different patterns that explain sleep quality or sleep/awake segments. An idea can be use different models, one that recognize nap and another one that recognize sleep. However, this approach will require more data, as not all people take naps and not all people that take naps take naps every day in our dataset very few sessions are nap.

- Other variants of our models can be interesting to see:

  - Instead of predict sleep quality we can infer better/worse from the usual user sleep quality. This because for a user this information could be more important to understand how improve his sleep, since even for them an objective sleep quality is not so trivial to understand. Also for this reason a better/worse from the usual could represent better the perceived sleep quality.
  - As explained, we predicted sleep quality in a time window even if our ground-truth is on the overall sleep quality, nap or sleep night. A possible improvements can be to use each one of these window sleep quality predictions during sleep and predict the overall sleep as the majority of the predictions during that whole sleep.
  - Understand how much the performances change if at the very beginning (first day of use) we submit a PSQI questionnaire so during classification we will have a feature with the PSQI score. This will not require too much effort to the user since nowadays pretty much all applications before the first use ask user to answer some questions (e.g., demographics). A similar approach was also used in literature, Can et al. [2020] used a questionnaire in order to cluster user with similar score and apply for each cluster a different model.

- Our model applied filters at the end of the sessions, in this way we could extract better results from this phase since we will have a bigger time series. This means that the prediction will be at the end of the sleep, in spite of everything, a real-time version of this approach could be implemented in which we could make prediction every 10 minutes. Nevertheless, we need to be aware that our definition of storm required a 10-minute window so we cannot go under this time period.

- We prefer to focus on just physiological data, so results were obtained without time information. For this reason, we can state that probably by just adding the knowledge of the time we will improve results.

Chapter 9

Conclusion

The following sections draw the conclusion by summarizing the contributions of this thesis and presenting its implications.

This work aims to understand the feasibility of a robust automatic approach for detecting whether a user is sleeping or is awake and the subjective quality of the sleep. To address our research question we: ran a data collection of 30-days, tested our approach, explained and reasoned on results.

In particular, the best results we found are: a balance accuracy of 90.63% in a 10-minute time window for the sleep/awake user-dependent model (with also storm labels) classification task and a balance accuracy of 63.70% in 1-minute window for the high/low sleep quality user-dependent model.

9.1 Contributions

The contributions of this thesis are presented as follows:

- Explanation of background concepts necessary to understand the thesis but also the literature works presented. These concepts were obtained from both other research and books (chapter 3).
- An in-depth review of existing literature of similar studies with a focus on identify gap in the research (chapter 2).
- Design and carry out a data collection in a real-world setting(chapter 4).
- Dedicated tools to monitor data quality and quantity during data collection (chapter 5).
- Dashboard to visualize collected data and visually inspect it (chapter 5).
- Extension of EDArtifact by adding peak epochs and storms detection with definition based on literature studies (chapter 6).
- A machine learning pipeline to detect sleep/awake and subject sleep quality using electrodermal activity, skin temperature and acceleration data collected with wristbands (chapter 6).

- Evaluation of the model by comparing its performance with: its development as a user-dependent model and a user-independent model, different variants with different sensors and features, different time windows used, three baselines (chapter 7).
- Evaluation of commercial devices (two MiBand, one Fitbit and two Garmin) comparing them with self-reports collected during our study (chapter 7)
- Understand limitation of the current work and suggest future improvements (chapter 8).

9.2 Implications

We consider wearable sensors very promising in future health monitoring systems, especially since their capability of catching physiological signals can open up to thousands of opportunities.

For what we investigated in this work we can conclude that sleep/awake detection system has a good degree of reliability that, maybe in some specific settings, they can be used instead of demanding self-reports. Nevertheless, sleep quality still needs a further investigation, some ideas can be found in chapter 8 but it still missing a strong objective sleep quality definition in literature.

We also noticed that artifacts don't have a big impact of our model, this can be explained because, during sleep, body movements are few and happen in short time. So we may suppose that in our time windows (1, 5 and 10 minutes) there are very few artifacts, an idea for future works can be understand if by looking at artifacts aggregated, during a whole night, we would be able to quantify better their impact.

Reasoning on SHAP results give us an insight to what our model believe is more important to give right predictions, most of the results are double validate looking at how the performance vary after our slightly changes in the sensors and features used. This ensures and proves the reliability of those explanation methods. Anyway, in Table 7.1 and Table 7.4 we added also the results of the models obtained by using only the best 20 features (only with a time window of 10 minutes) according to SHAP, and indeed we obtained similar results to model with all the features (e.g., balance accuracy of SHAP_top_20 user-dependent model: 61.49%, EDA+TEMP+ACC: 62.63%)

To conclude, this work gives evidence to the feasibility of a robust automatic approach in sleep or awake detection in real-world settings. We demonstrated that perceived sleep quality is still no trivial to predict and we provided suggestions to improve that.

Appendix A

Documentation for the study

A.1 Informed Consent Agreement

Informed Consent Agreement

Study: Robust Detection of Sleep Quality Using Mobile and Wearable Sensors

Organisers of the study:
Silvia Santini: silvia.santini@usi.ch (Principal Investigator).
Lidia Alecci, lidia.alecci@usi.ch (Team), Shkurta Gashi: shkurta.gashi@usi.ch (Team), Elena Di Lascio, elena.dilascio@usi.ch (Team), and Maike Debus maike.debus@unine.ch.

Purpose of the research study: The purpose of this study is to use data collected using mobile and wearable devices (wristband and smartphone) to recognize sleep quality and quantity.

What you will do in the study: The study consists of two main phases *pre-study, study* and *post-study phase*. In pre-study phase, you will be asked to fill four questionnaires about your demographics, sleep routine, personality and chronotype. During this phase we will send you a study description and tutorials on how to install the tools needed for the study. We will also arrange a meeting to discuss any issues and questions you might have for the study. This phase will take approximately 90 minutes of your time. During the study, you will be asked to (1) wear the E4 wristband (https://www.empatica.com/research/e4/) every night and for at least 8 hours during the day to gather physiological data; (2) install an Android application that gathers behavioral data in the background; and (3) provide self-reports about the time when you go to sleep, wake up, and the sleep quality when you wake up. To provide self-reports you will choose to use an Android application we developed for the study, a pen-and-paper diary or a Google form accessible with your laptop. You will be reminded through the smartphone application to complete self-reports every day in the morning and evening as well as to charge the devices used for the study. You will be asked also to upload the data from the E4 wristband using the Empatica Manager installed in your laptop.

You can refrain from doing any of the requested tasks and you can stop the study at any time. During the study, the tools used for the study will gather the following data in the background:

- *Physiological data* will be collected from the sensors embedded in E4 device
  - Blood volume pulse
  - Electrodermal activity

    - Acceleration
    - Skin temperature
- *Behavioral data* will be collected from the sensors embedded in your smartphone
    - Time of phone lock/unlock events
    - Time of *screen on/off* events
    - Time and type of *applications* used on the phone
    - Time and application from which a *notification* arrived
    - Time and *proximity* of the phone screen to any surface
    - Time and amount of *ambient light*
    - Smartphone movements
- *Self-reports* collected with smartphone app or diaries
    - Going to sleep time
    - Waking up time
    - Sleep quality score

The sensor data can be used as an objective basis to infer sleep and waking up times as well as sleep quality. The data collected by the Android application, will be anonymized and uploaded to SWITCHdrive (https://www.switch.ch/drive/) automatically every night. SWITCHdrive is a secure academic cloud storage service. The data from the Empatica Manager is anonymized and sent to the Empatica servers. For further information please refer to this link (https://support.empatica.com/hc/en-us/articles/202524239-What-does-Empatica-do-to-protect-end-user-privacy). You can retain access to your data, on request, and you can decide to delete it at any time (see "Study Withdrawal" policy below). After the study, we will send a semi-structured questionnaire with questions regarding the study and experience with the tools.

Time required: The pre-study and post-study phases will require in total approximately 90 minutes of your time. The study phase will require approximately 5 minutes of your time to complete the self-reports every day and to wear the device for 30 days.

Compensation: We will provide a bag of chocolates upon study enrollment. Your participation in the study will be compensated with an initial amount of 20CHF and additionally 1CHF per each day will be provided upon successful collection of the following data 1) physiological data with E4, 2) sleep and wake up time reports and 3) sleep quality answers for 30 days.

Confidentiality: The information and data collected in this study will be stored safely and handled confidentially. Your data will be anonymized through the assignment of an alpha-numerical code and your name will never be mentioned in connection to the data or in any report. Any attempt to deduce your identity from the data is explicitly forbidden by our data analysis policy.

Data sharing agreement: The organizers of this study request your permission to analyze the collected data for research purposes. Your data will be used only in anonymized form and your identity will never be revealed or used in connection with the research efforts. You can also withdraw your permission to use the data at a later stage as indicated in our Study Withdrawal policy. Please indicate whether you agree your *physiological*, *behavioral* and *self-reported* data to be used for research purposes or not.

□ I DO give consent for my data to be used for research purposes.

□ I DO NOT give consent for my data to be used for research purposes.

Study withdrawal: Your participation in the study is completely voluntary. If for any reason you want to withdraw from the study, you can easily do so at any time by uninstalling the application and stop wearing the E4 and using the mobile application. No further data sensing will be performed. You have the right to require data collected prior to your withdrawal to be partially or entirely deleted. To do so please send an e-mail to the organizers of this study (indicated above) with subject "Study Withdrawal" specifying if you want all or part of your data to be permanently deleted. Data that are not removed from the database before May 15, 2021 will be preserved (in anonymized format!) and remain accessible for the authorized researchers for ten years.

□ I participate in this study on a voluntary basis and can withdraw from the study at any time without giving reasons and without any negative consequences.

□ I have been informed orally and in writing about the aims and the procedures of the study, the advantages and disadvantages, as well as potential risks.

□ My questions related to the study have been answered satisfactorily.

□ I have been given a copy of this consent form.

□ I was given enough time to make a decision about participating in the study.

Name: ____

Date: ____

Signature: ____

A.2 Questionnaires

A.2.1 Demographic Questionnaire

Demographics Questionnaire

* Required

1. Username *

2. What is your age? *

3. To which gender do you most identify? *

Mark only one oval.

○ Female
○ Male
○ Non-Binary
○ Prefer not to say
○ Other:

4. Please specify your ethnicity *

Mark only one oval.

○ White
○ Hispanic or Latino
○ Black or African American
○ Native American or American Indian
○ Asian / Pacific Islander
○ Other:

5. Are you currently __ ?

Mark only one oval.

○ Employed for wages

○ Self-employed

○ Out of work and looking for work

○ Out of work but not currently looking for work

○ A homemaker

○ A student

○ Retired

6. What is the highest degree or level of education you have completed? *

Mark only one oval.

○ High School

○ Bachelor's Degree

○ Master's Degree

○ Ph.D. or higher

○ Prefer not to say

7. Do you track your sleep in any way (e.g., using a smartwatch or smartphone application)?

Mark only one oval.

○ Yes

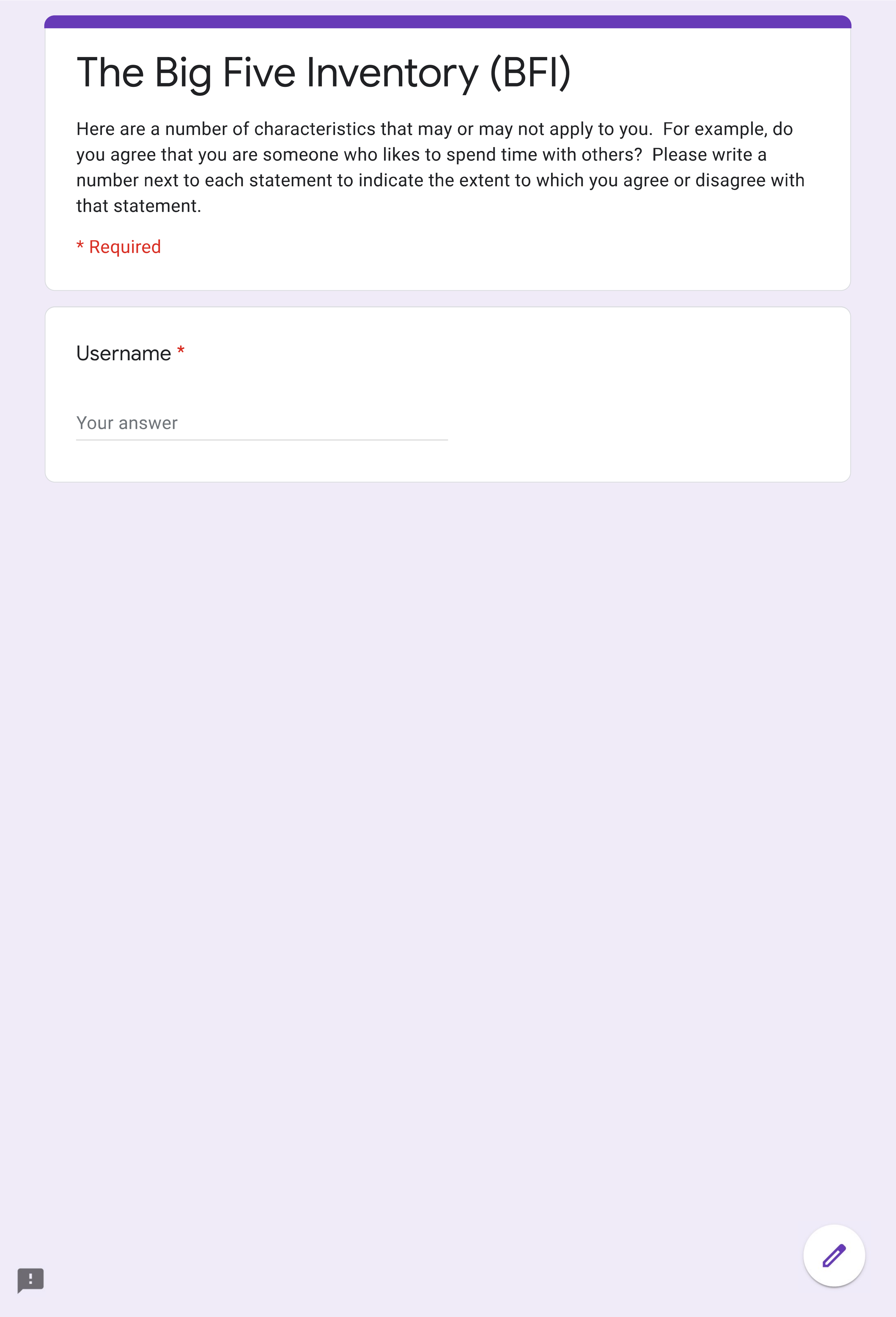
○ No

8. If yes, please explain which tool do you use and how do you track your sleep behavior.

This content is neither created nor endorsed by Google.

Google Forms

A.2.2 The Big Five Inventory (BFI)



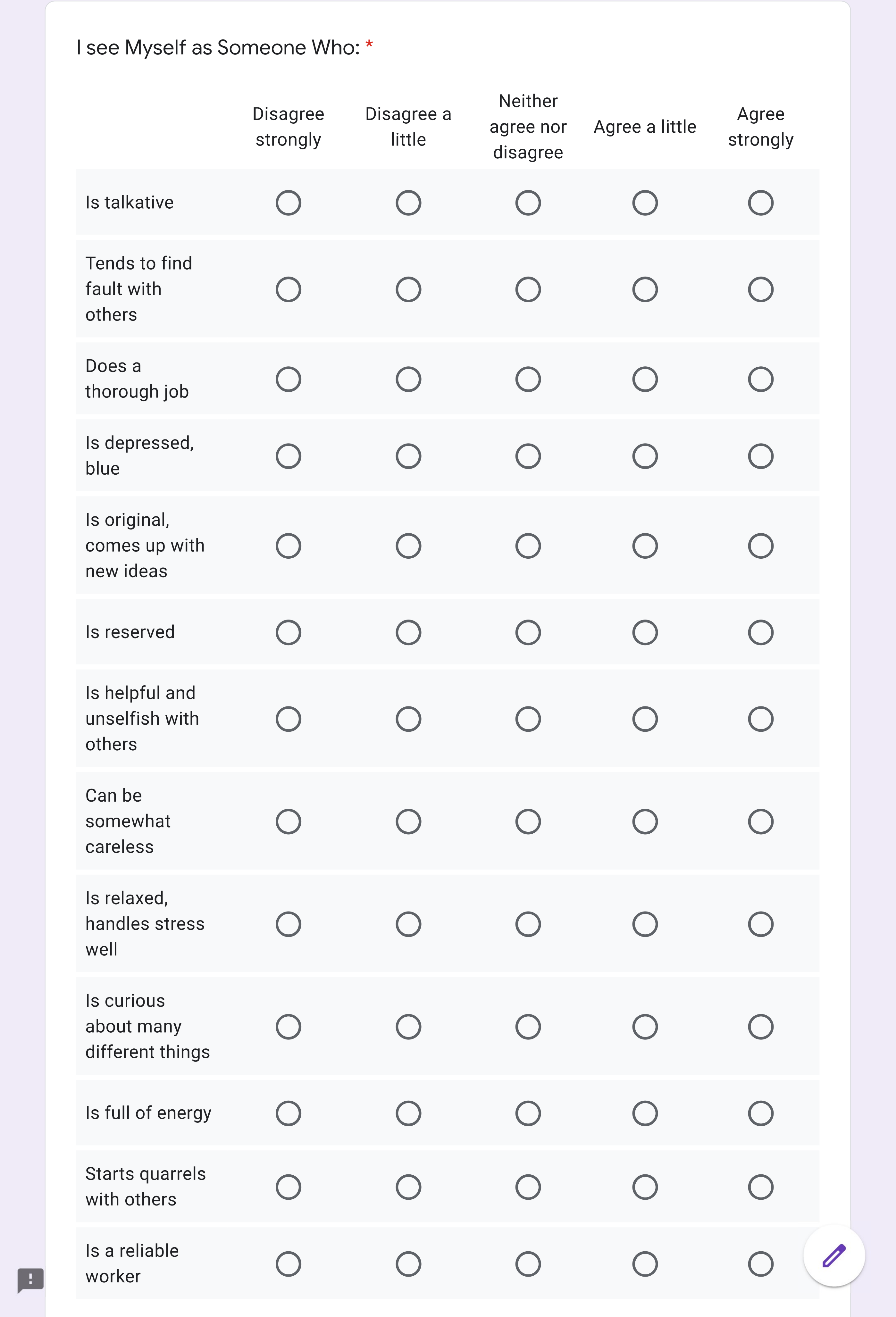
The Big Five Inventory (BFI)

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

* Required

Username *

Your answer



I see Myself as Someone Who: *

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
Is talkative	○	○	○	○	○
Tends to find fault with others	○	○	○	○	○
Does a thorough job	○	○	○	○	○
Is depressed, blue	○	○	○	○	○
Is original, comes up with new ideas	○	○	○	○	○
Is reserved	○	○	○	○	○
Is helpful and unselfish with others	○	○	○	○	○
Can be somewhat careless	○	○	○	○	○
Is relaxed, handles stress well	○	○	○	○	○
Is curious about many different things	○	○	○	○	○
Is full of energy	○	○	○	○	○
Starts quarrels with others	○	○	○	○	○
Is a reliable worker	○	○	○	○	○

Can be tense	○	○	○	○	○
Is ingenious, a deep thinker	○	○	○	○	○
Generates a lot of enthusiasm	○	○	○	○	○
Has a forgiving nature	○	○	○	○	○
Tends to be disorganized	○	○	○	○	○
Worries a lot	○	○	○	○	○
Has an active imagination	○	○	○	○	○
Tends to be quiet	○	○	○	○	○
Is generally trusting	○	○	○	○	○
Tends to be lazy	○	○	○	○	○
Is emotionally stable, not easily upset	○	○	○	○	○
Is inventive	○	○	○	○	○
Has an assertive personality	○	○	○	○	○
Can be cold and aloof	○	○	○	○	○
Perseveres until the task is finished	○	○	○	○	○
Can be moody	○	○	○	○	○

Values artistic, aesthetic experiences	○	○	○	○	○
Is sometimes shy, inhibited	○	○	○	○	○
Is considerate and kind to almost everyone	○	○	○	○	○
Does things efficiently	○	○	○	○	○
Remains calm in tense situations	○	○	○	○	○
Prefers work that is routine	○	○	○	○	○
Is outgoing, sociable	○	○	○	○	○
Is sometimes rude to others	○	○	○	○	○
Makes plans and follows through with them	○	○	○	○	○
Gets nervous easily	○	○	○	○	○
Likes to reflect, play with ideas	○	○	○	○	○
Has few artistic interests	○	○	○	○	○
Likes to cooperate with others	○	○	○	○	○
Is easily distracted	○	○	○	○	○
Is sophisticated	○	○	○	○	○

Is sophisticated in art, music, or literature

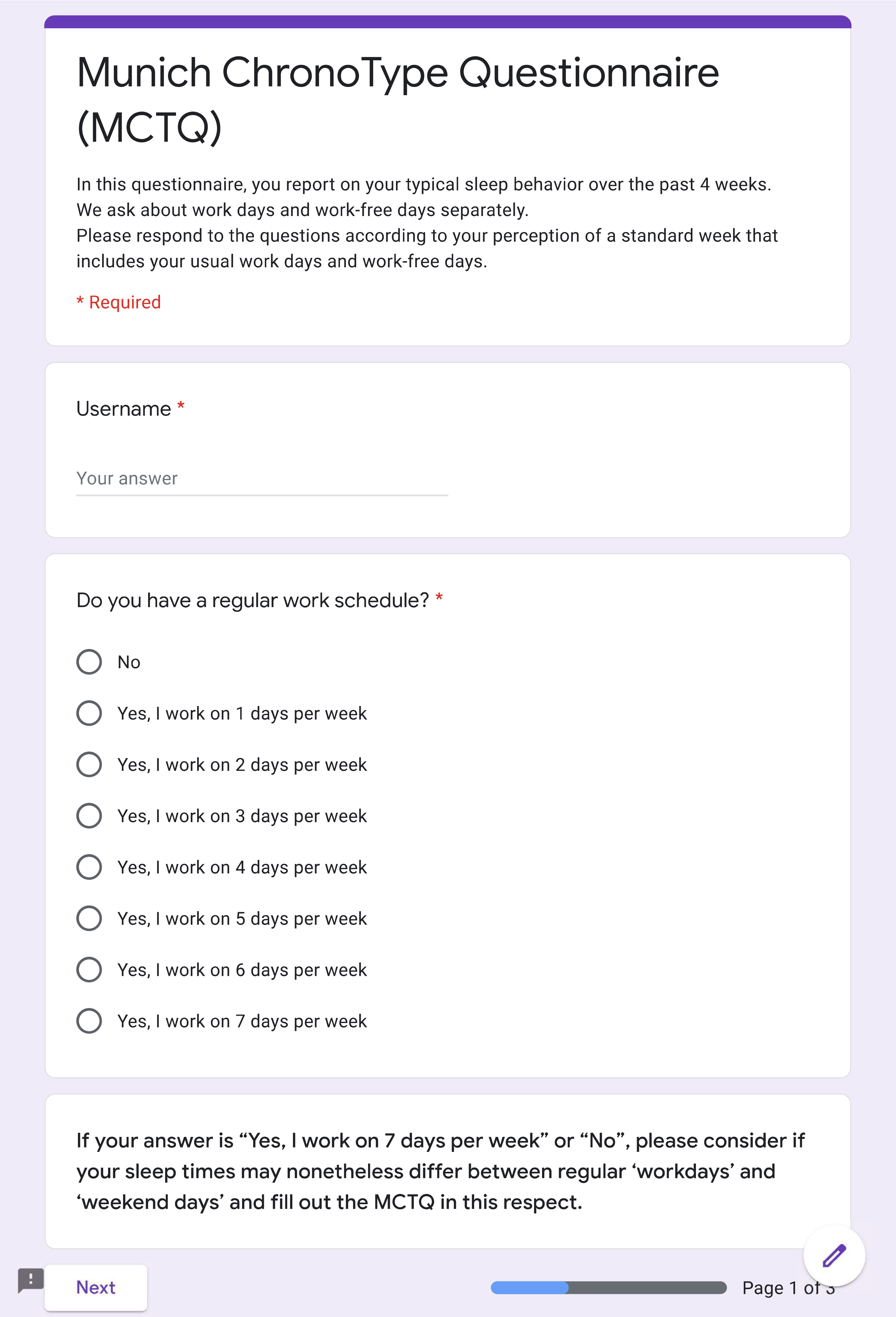
Submit

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Google Forms

A.2.3 Munich ChronoType Questionnaire (MCTQ)



Munich ChronoType Questionnaire (MCTQ)

In this questionnaire, you report on your typical sleep behavior over the past 4 weeks.
We ask about work days and work-free days separately.
Please respond to the questions according to your perception of a standard week that includes your usual work days and work-free days.

* Required

Username *

Your answer

Do you have a regular work schedule? *

○ No

○ Yes, I work on 1 days per week

○ Yes, I work on 2 days per week

○ Yes, I work on 3 days per week

○ Yes, I work on 4 days per week

○ Yes, I work on 5 days per week

○ Yes, I work on 6 days per week

○ Yes, I work on 7 days per week

If your answer is “Yes, I work on 7 days per week” or “No”, please consider if your sleep times may nonetheless differ between regular ‘workdays’ and ‘weekend days’ and fill out the MCTQ in this respect.

Munich ChronoType Questionnaire (MCTQ)

* Required

Workdays

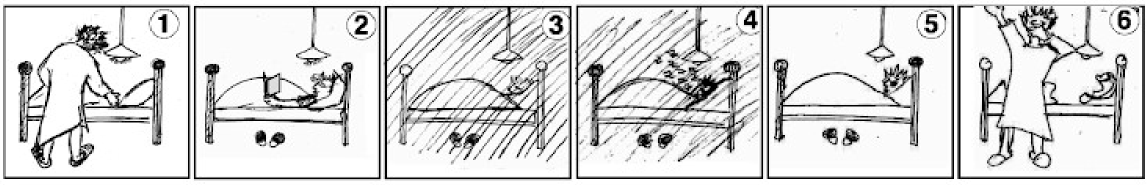


Image 1: I go to bed at _________ o'clock. *

Your answer

Image 2: Note that some people stay awake for some time when in bed!

Image 3: I actually get ready to fall asleep at _________ o'clock. *

Your answer

Image 4: I need _________ minutes to fall asleep. *

Your answer

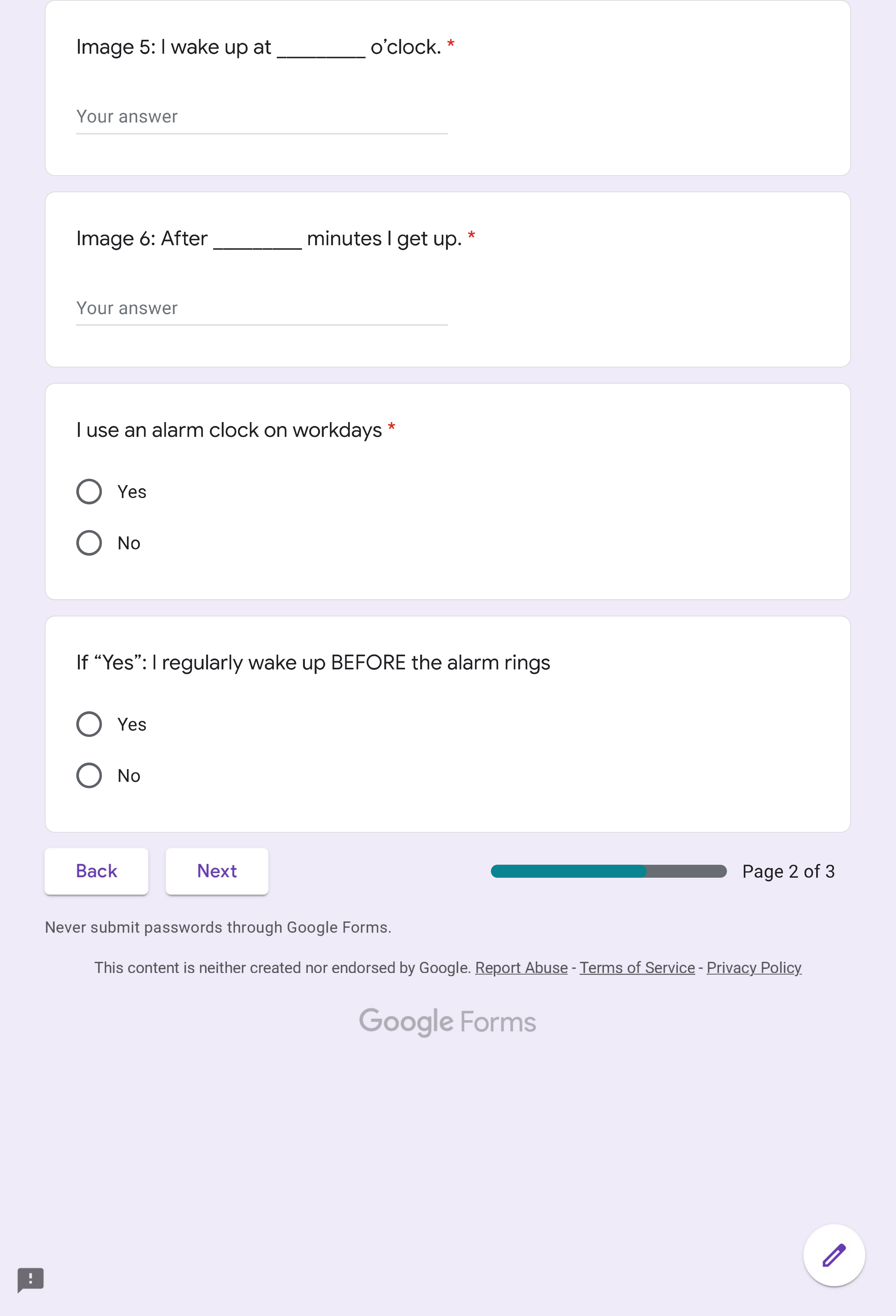


Image 5: I wake up at ________ o'clock. *

Your answer

Image 6: After ________ minutes I get up. *

Your answer

I use an alarm clock on workdays *

○ Yes

○ No

If "Yes": I regularly wake up BEFORE the alarm rings

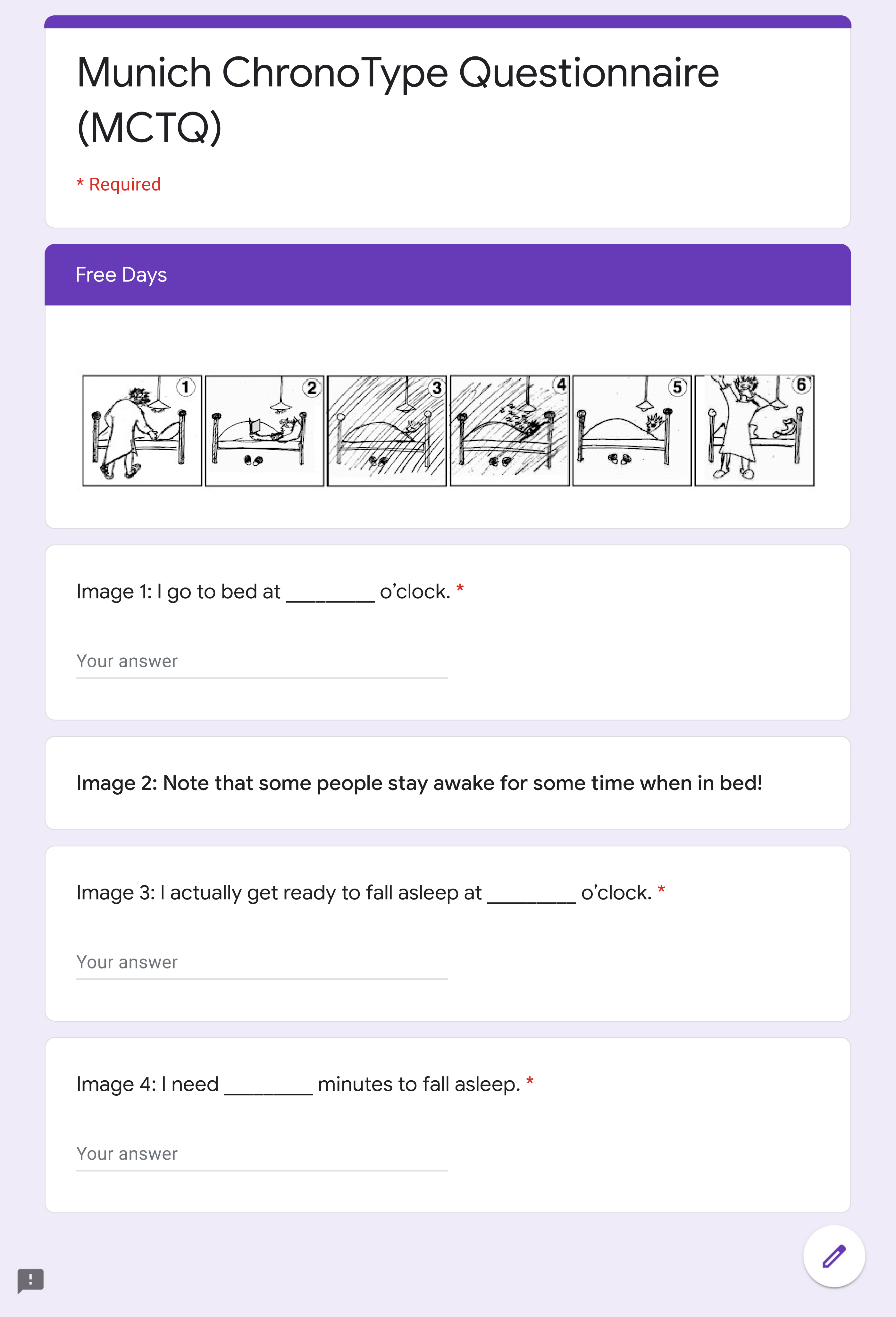
○ Yes

○ No

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Google Forms



Munich ChronoType Questionnaire (MCTQ)

* Required

Free Days

Image 1: I go to bed at ________ o’clock. *

Your answer

Image 2: Note that some people stay awake for some time when in bed!

Image 3: I actually get ready to fall asleep at ________ o’clock. *

Your answer

Image 4: I need ________ minutes to fall asleep. *

Your answer

Image 5: I wake up at ________ o’clock. *

Your answer

Image 6: After ________ minutes I get up. *

Your answer

My wake-up time (Image 5) is due to the use of an alarm clock *

○ Yes

○ No

There are particular reasons why I cannot freely choose my sleep times on free days

○ Yes

○ No

If “Yes”:

○ Child(ren)/pet(s)

○ Hobbies

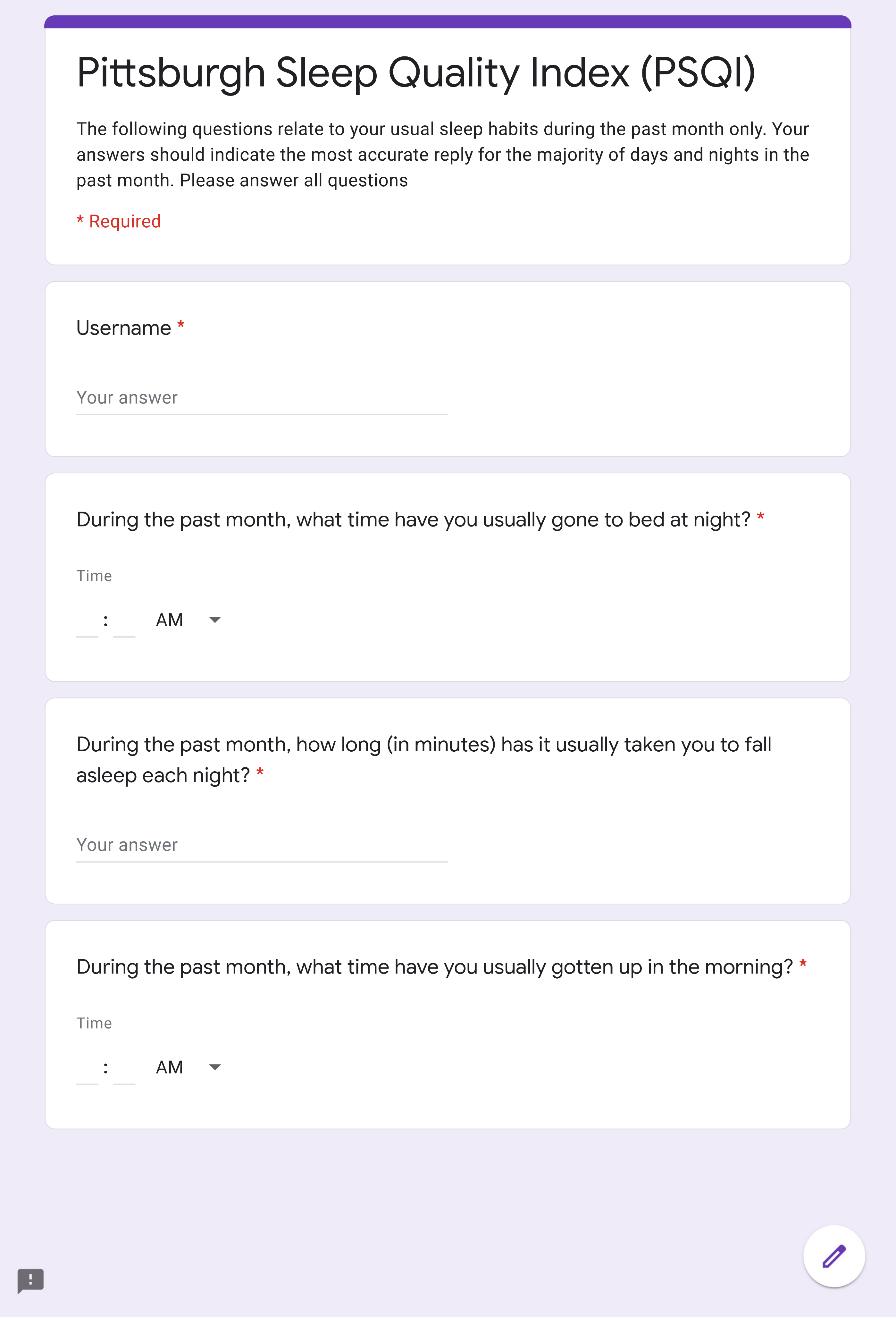
○ Other: ____________________

Back Submit Page 3 of 3

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

A.2.4 Pittsburgh Sleep Quality Index (PSQI)



Pittsburgh Sleep Quality Index (PSQI)

The following questions relate to your usual sleep habits during the past month only. Your answers should indicate the most accurate reply for the majority of days and nights in the past month. Please answer all questions

* Required

Username *

Your answer

During the past month, what time have you usually gone to bed at night? *

Time

__ : __ AM ▾

During the past month, how long (in minutes) has it usually taken you to fall asleep each night? *

Your answer

During the past month, what time have you usually gotten up in the morning? *

Time

__ : __ AM ▾

During the past month, how many hours of actual sleep did you get at night? (This may be different than the number of hours you spent in bed.) *

Your answer

During the past month, how often have you had trouble sleeping because you: *

	Not during the past month	Less than once a week	Once or twice a week	Three or more times a week
Cannot get to sleep within 30 minutes	○	○	○	○
Wake up in the middle of the night or early morning	○	○	○	○
Have to get up to use the bathroom	○	○	○	○
Cannot breathe comfortably	○	○	○	○
Cough or snore loudly	○	○	○	○
Feel too cold	○	○	○	○
Feel too hot	○	○	○	○
Had bad dreams	○	○	○	○
Have pain	○	○	○	○

Other reason(s), please describe

Your answer

If you answer the question above, how often during the past month have you had trouble sleeping because of that other reason(s)?

○ Not during the past month

○ Less than once a week

○ Once or twice a week

○ Three or more times a week

During the past month, how would you rate your sleep quality overall? *

○ Excellent

○ Good

○ Normal

○ Poor

○ Very poor

During the past month, how often have you taken medicine to help you sleep (prescribed or "over the counter")? *

○ Not during the past month

○ Less than once a week

○ Once or twice a week

○ Three or more times a week

During the past month, how often have you had trouble staying awake while driving, eating meals, or engaging in social activity? *

○ Not during the past month

○ Less than once a week

○ Once or twice a week

○ Three or more times a week

During the past month, how much of a problem has it been for you to keep up enough enthusiasm to get things done? *

○ No problem at all

○ Only a very slight problem

○ Somewhat of a problem

○ A very big problem

Do you have a bed partner or room mate? *

○ No bed partner or room mate

○ Partner/room mate in other room

○ Partner in same room, but not same bed

○ Partner in same bed

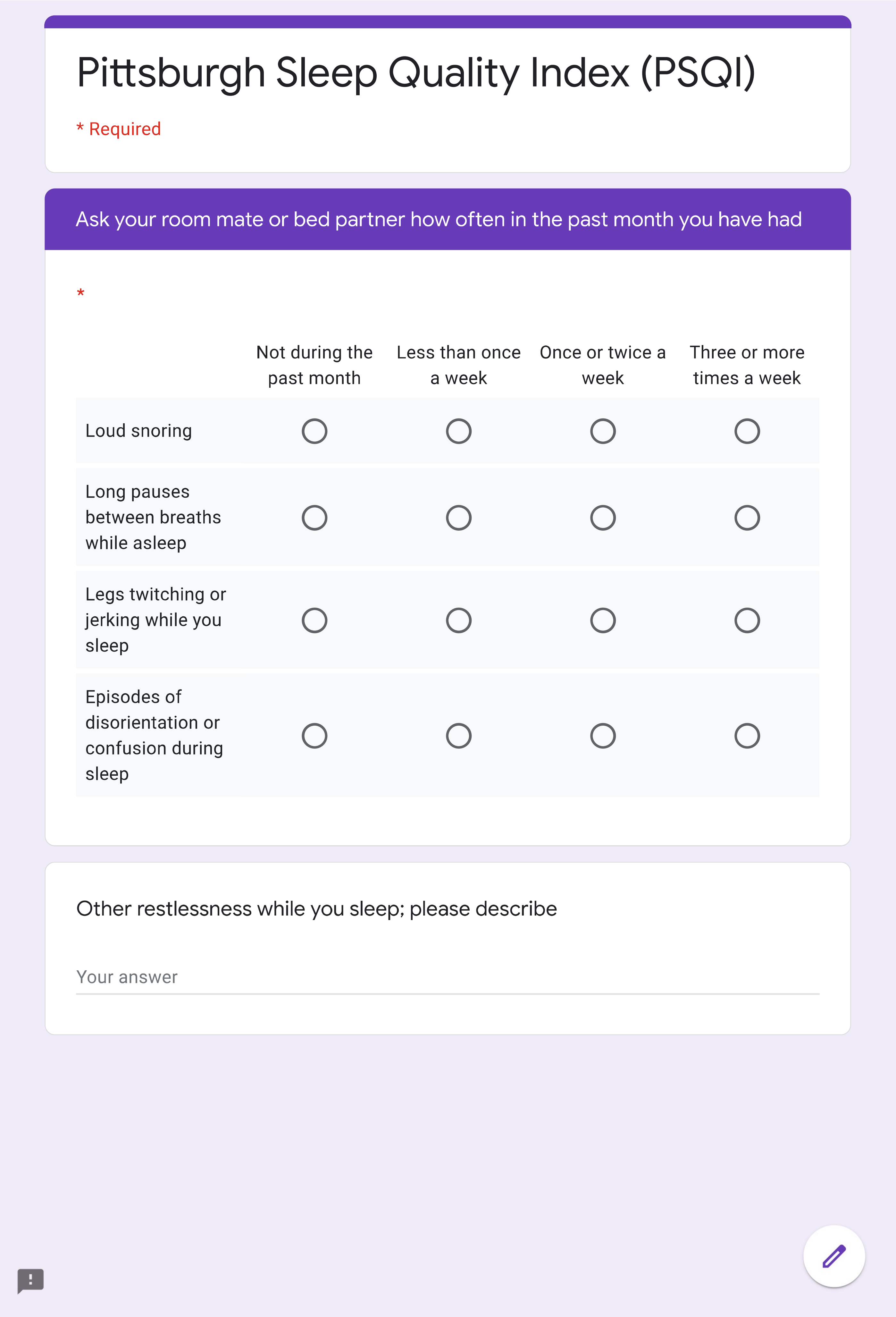
Next

Page 1 of 2

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Google Forms



Pittsburgh Sleep Quality Index (PSQI)

* Required

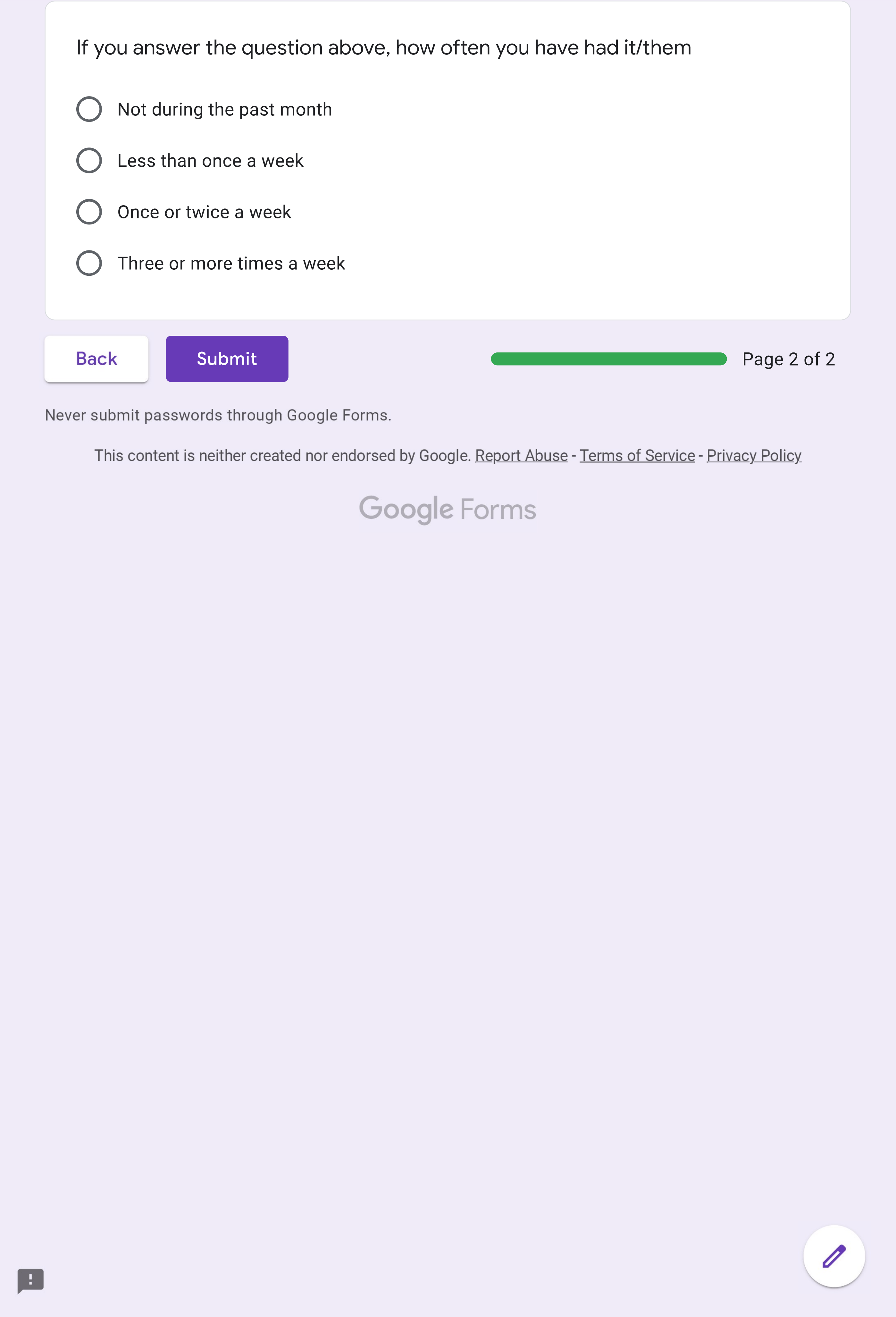
Ask your room mate or bed partner how often in the past month you have had

*

	Not during the past month	Less than once a week	Once or twice a week	Three or more times a week
Loud snoring	○	○	○	○
Long pauses between breaths while asleep	○	○	○	○
Legs twitching or jerking while you sleep	○	○	○	○
Episodes of disorientation or confusion during sleep	○	○	○	○

Other restlessness while you sleep; please describe

Your answer



If you answer the question above, how often you have had it/them

○ Not during the past month

○ Less than once a week

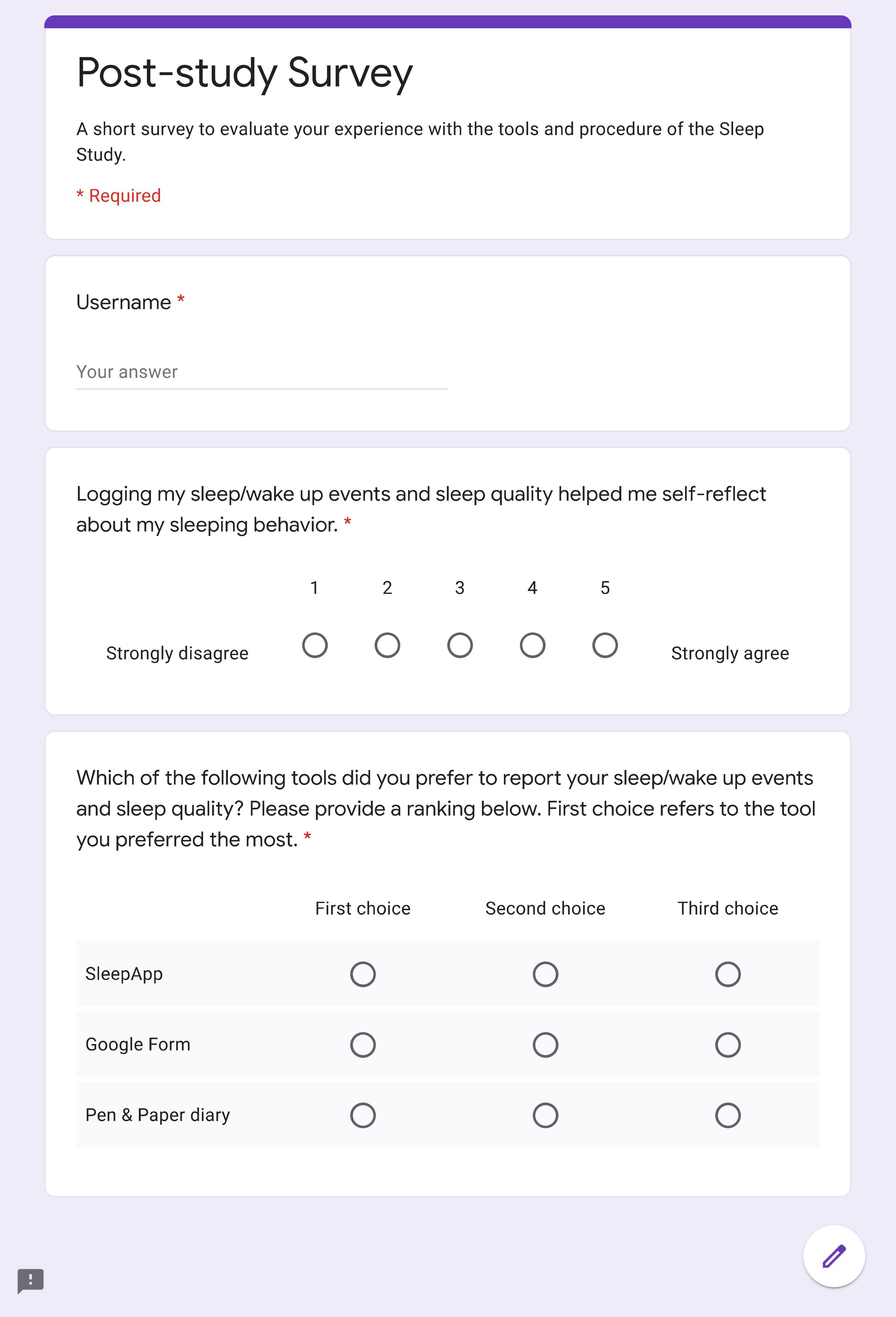
○ Once or twice a week

○ Three or more times a week

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

A.2.5 Post-study Survey



Post-study Survey

A short survey to evaluate your experience with the tools and procedure of the Sleep Study.

* Required

Username *

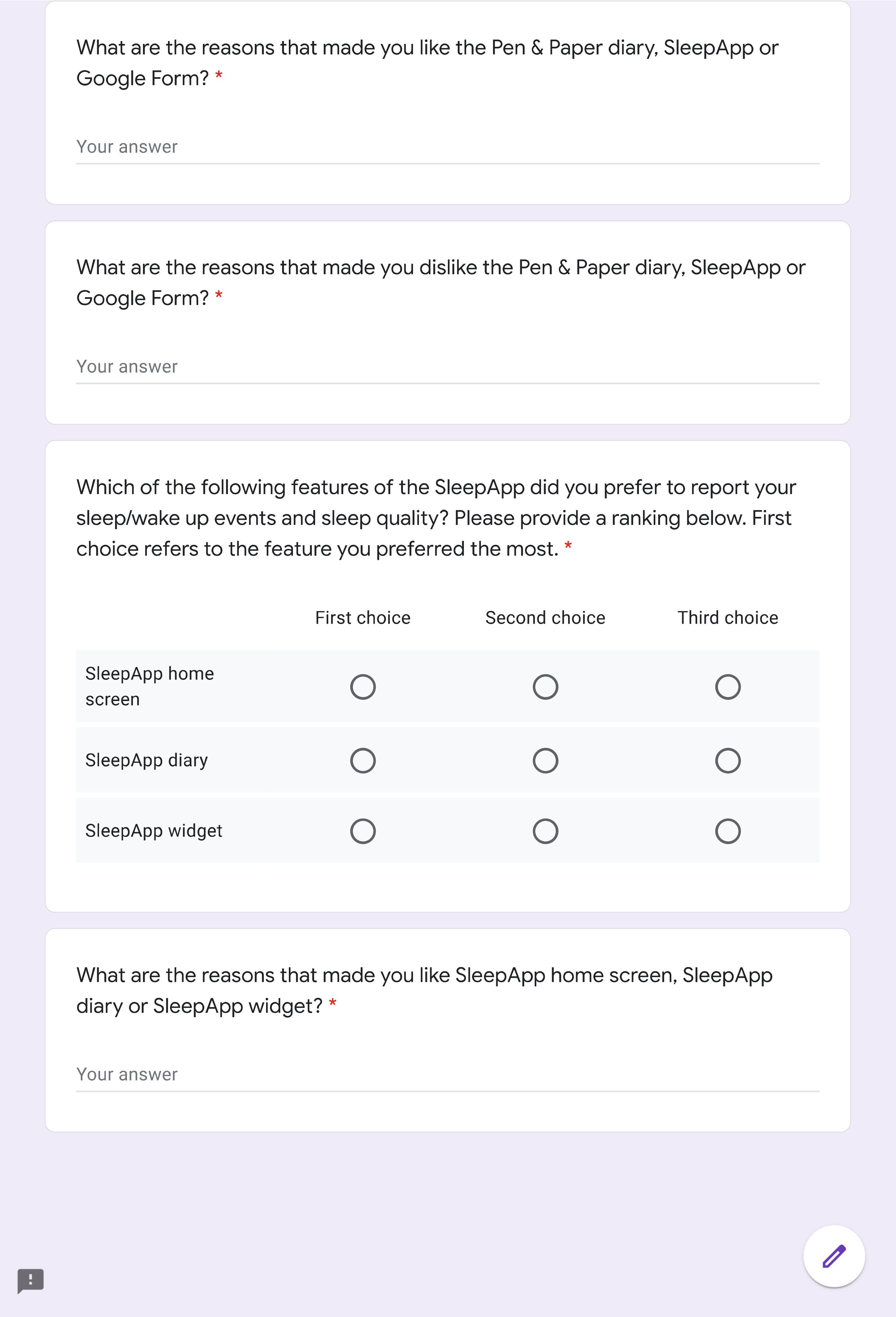
Your answer

Logging my sleep/wake up events and sleep quality helped me self-reflect about my sleeping behavior. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Which of the following tools did you prefer to report your sleep/wake up events and sleep quality? Please provide a ranking below. First choice refers to the tool you preferred the most. *

	First choice	Second choice	Third choice
SleepApp	○	○	○
Google Form	○	○	○
Pen & Paper diary	○	○	○



What are the reasons that made you like the Pen & Paper diary, SleepApp or Google Form? *

Your answer

What are the reasons that made you dislike the Pen & Paper diary, SleepApp or Google Form? *

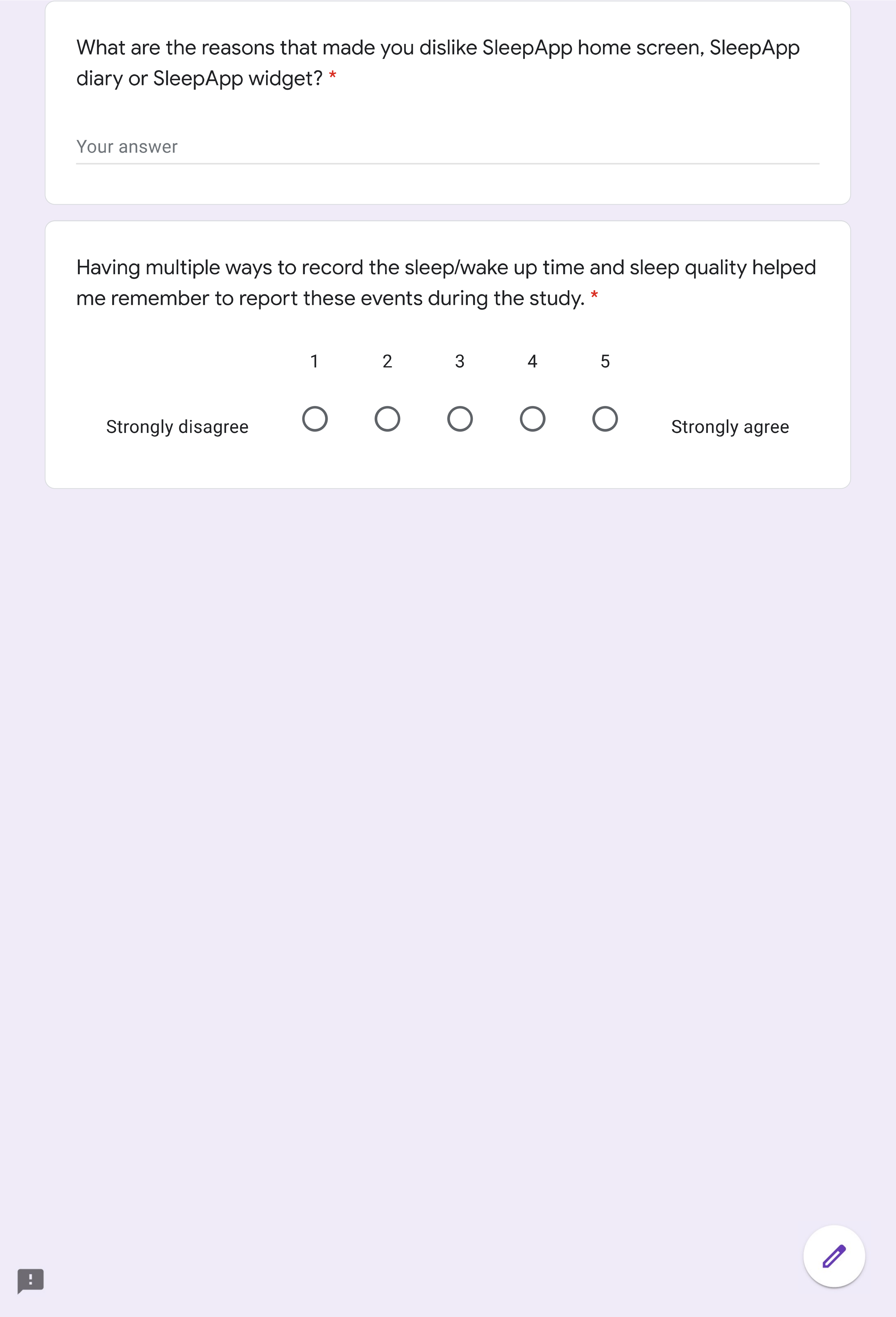
Your answer

Which of the following features of the SleepApp did you prefer to report your sleep/wake up events and sleep quality? Please provide a ranking below. First choice refers to the feature you preferred the most. *

	First choice	Second choice	Third choice
SleepApp home screen	○	○	○
SleepApp diary	○	○	○
SleepApp widget	○	○	○

What are the reasons that made you like SleepApp home screen, SleepApp diary or SleepApp widget? *

Your answer

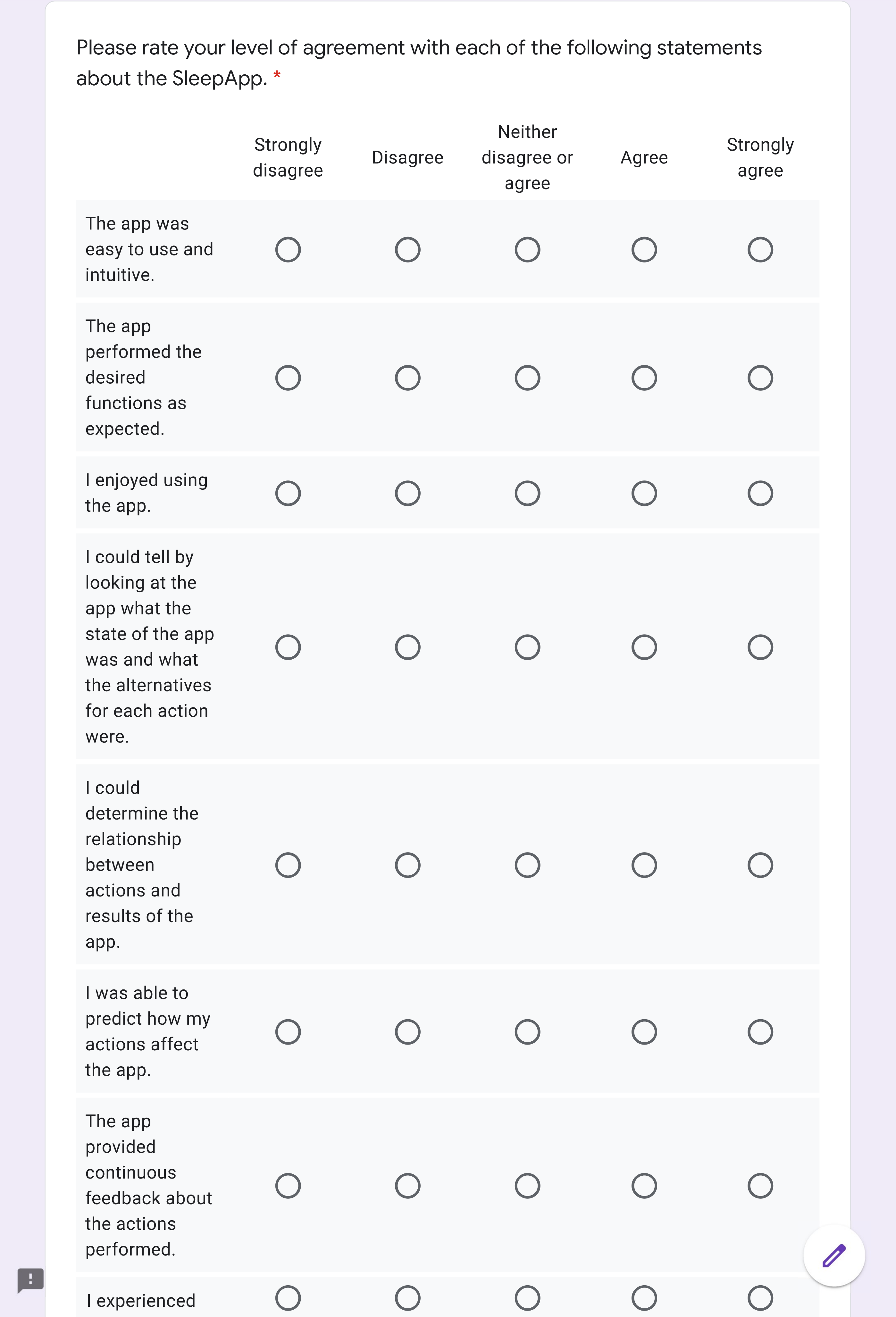


What are the reasons that made you dislike SleepApp home screen, SleepApp diary or SleepApp widget? *

Your answer

Having multiple ways to record the sleep/wake up time and sleep quality helped me remember to report these events during the study. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree



Please rate your level of agreement with each of the following statements about the SleepApp. *

	Strongly disagree	Disagree	Neither disagree or agree	Agree	Strongly agree
The app was easy to use and intuitive.	○	○	○	○	○
The app performed the desired functions as expected.	○	○	○	○	○
I enjoyed using the app.	○	○	○	○	○
I could tell by looking at the app what the state of the app was and what the alternatives for each action were.	○	○	○	○	○
I could determine the relationship between actions and results of the app.	○	○	○	○	○
I was able to predict how my actions affect the app.	○	○	○	○	○
The app provided continuous feedback about the actions performed.	○	○	○	○	○
I experienced	○	○	○	○	○

I experienced technical problems with the app (e.g. crashing, errors).					
I experienced performance problems (e.g., battery consumption, speed) with my phone after installing the app.	○	○	○	○	○

Where did you put your phone during the study while you were sleeping? *

○ On the bed

○ Near the bed face down

○ Near the bed face up

○ Far away from the bed

○ Other: ____________________

Where did you put your phone before the study while you were sleeping? *

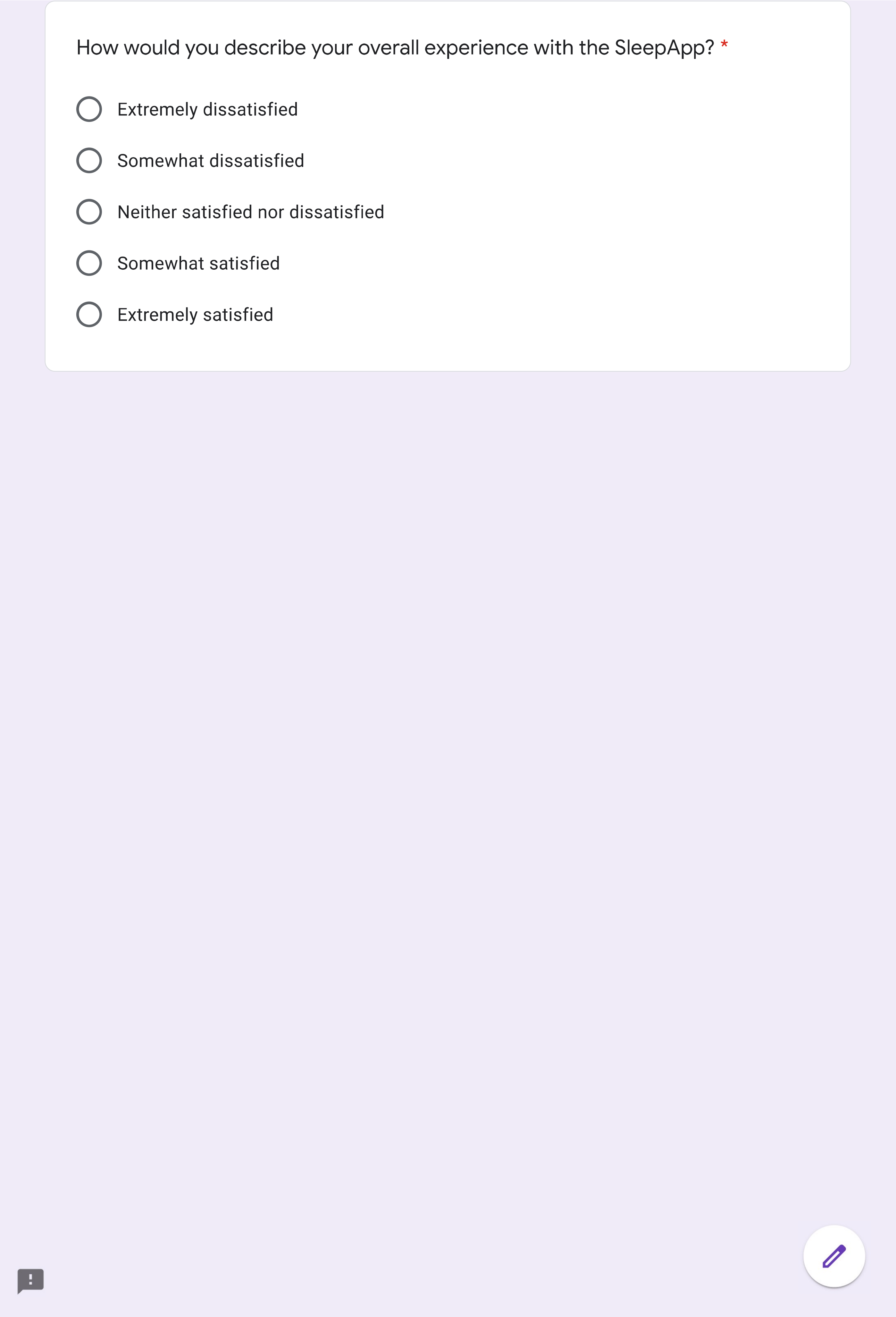
○ On the bed

○ Near the bed face down

○ Near the bed face up

○ Far away from the bed

○ Other: ____________________



How would you describe your overall experience with the SleepApp? *

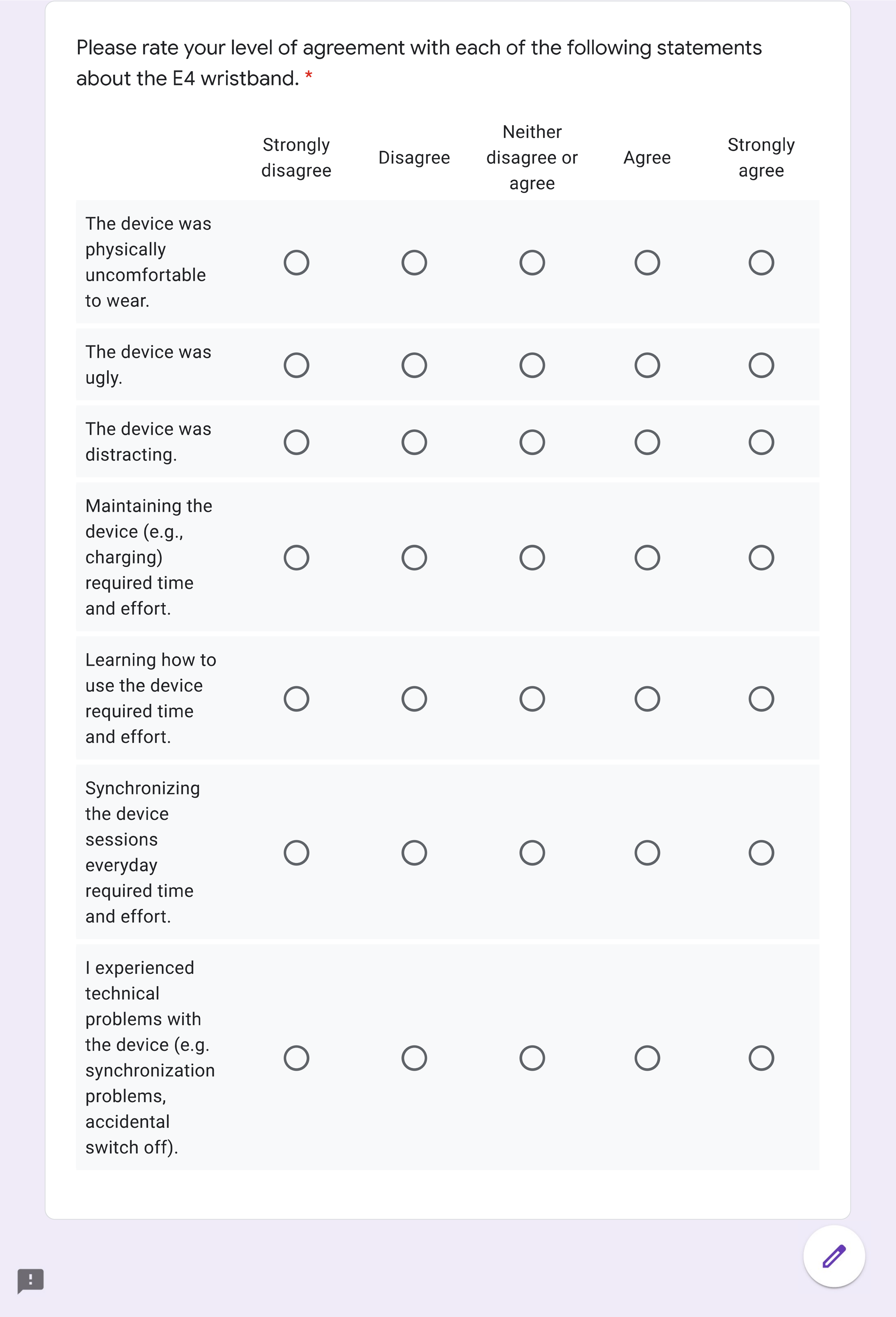
○ Extremely dissatisfied

○ Somewhat dissatisfied

○ Neither satisfied nor dissatisfied

○ Somewhat satisfied

○ Extremely satisfied



Please rate your level of agreement with each of the following statements about the E4 wristband. *

	Strongly disagree	Disagree	Neither disagree or agree	Agree	Strongly agree
The device was physically uncomfortable to wear.	○	○	○	○	○
The device was ugly.	○	○	○	○	○
The device was distracting.	○	○	○	○	○
Maintaining the device (e.g., charging) required time and effort.	○	○	○	○	○
Learning how to use the device required time and effort.	○	○	○	○	○
Synchronizing the device sessions everyday required time and effort.	○	○	○	○	○
I experienced technical problems with the device (e.g. synchronization problems, accidental switch off).	○	○	○	○	○

How would you describe your overall experience with the Empatica E4 wristband? *

○ Extremely dissatisfied

○ Somewhat dissatisfied

○ Neither satisfied nor dissatisfied

○ Somewhat satisfied

○ Extremely satisfied

How much would you be interested in wearing the E4 wristband in your everyday life if it would be able to provide feedback of your sleep and how to improve it? *

○ Not interested at all

○ Not very interested

○ Neither interested nor uninterested

○ Somewhat interested

○ Very interested

How much interested are you in knowing your physiological data (e.g., heart rate, body temperature, etc.) throughout the day and night? *

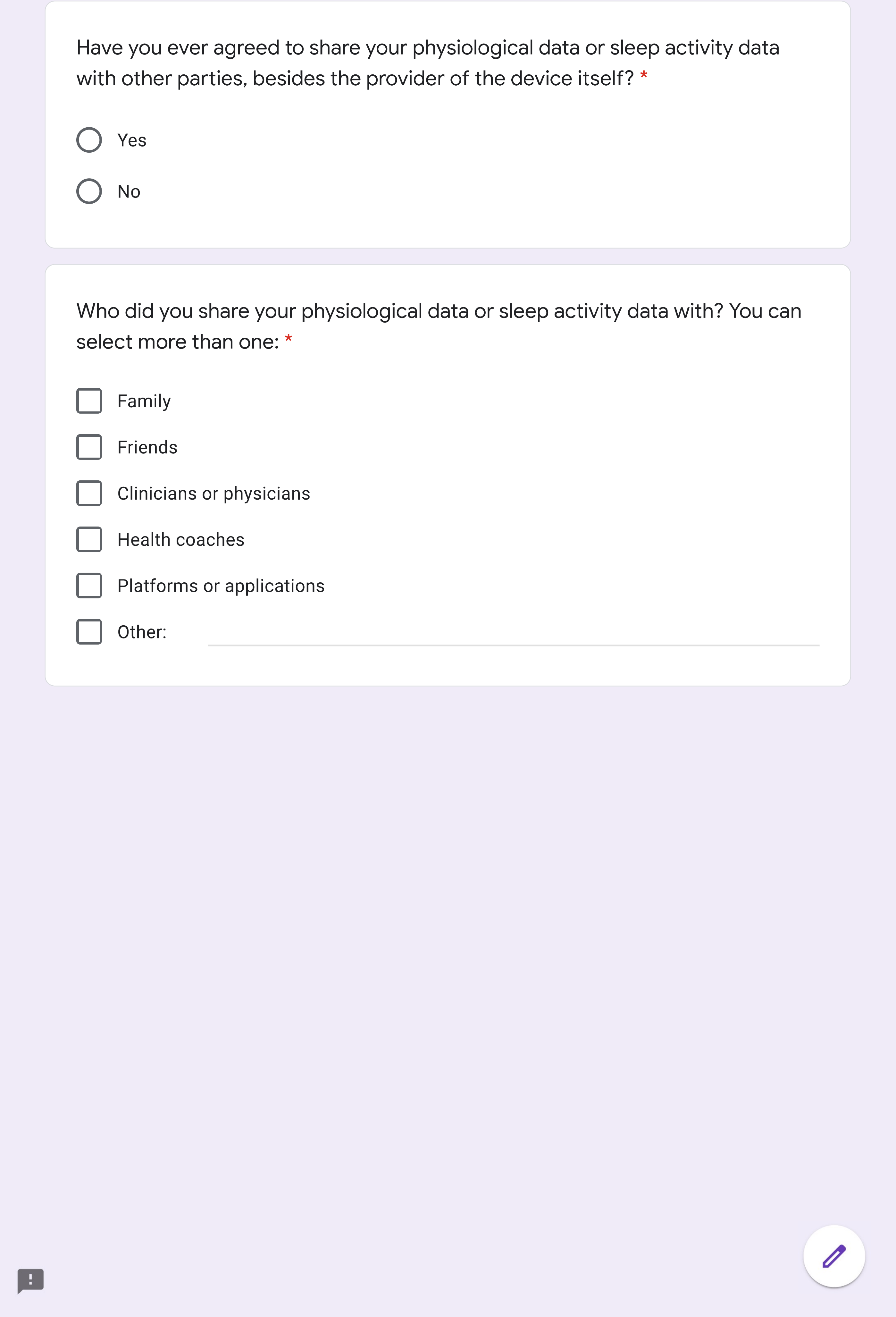
○ Not interested at all

○ Not very interested

○ Neither interested nor uninterested

○ Somewhat interested

○ Very interested



Have you ever agreed to share your physiological data or sleep activity data with other parties, besides the provider of the device itself? *

○ Yes

○ No

Who did you share your physiological data or sleep activity data with? You can select more than one: *

□ Family

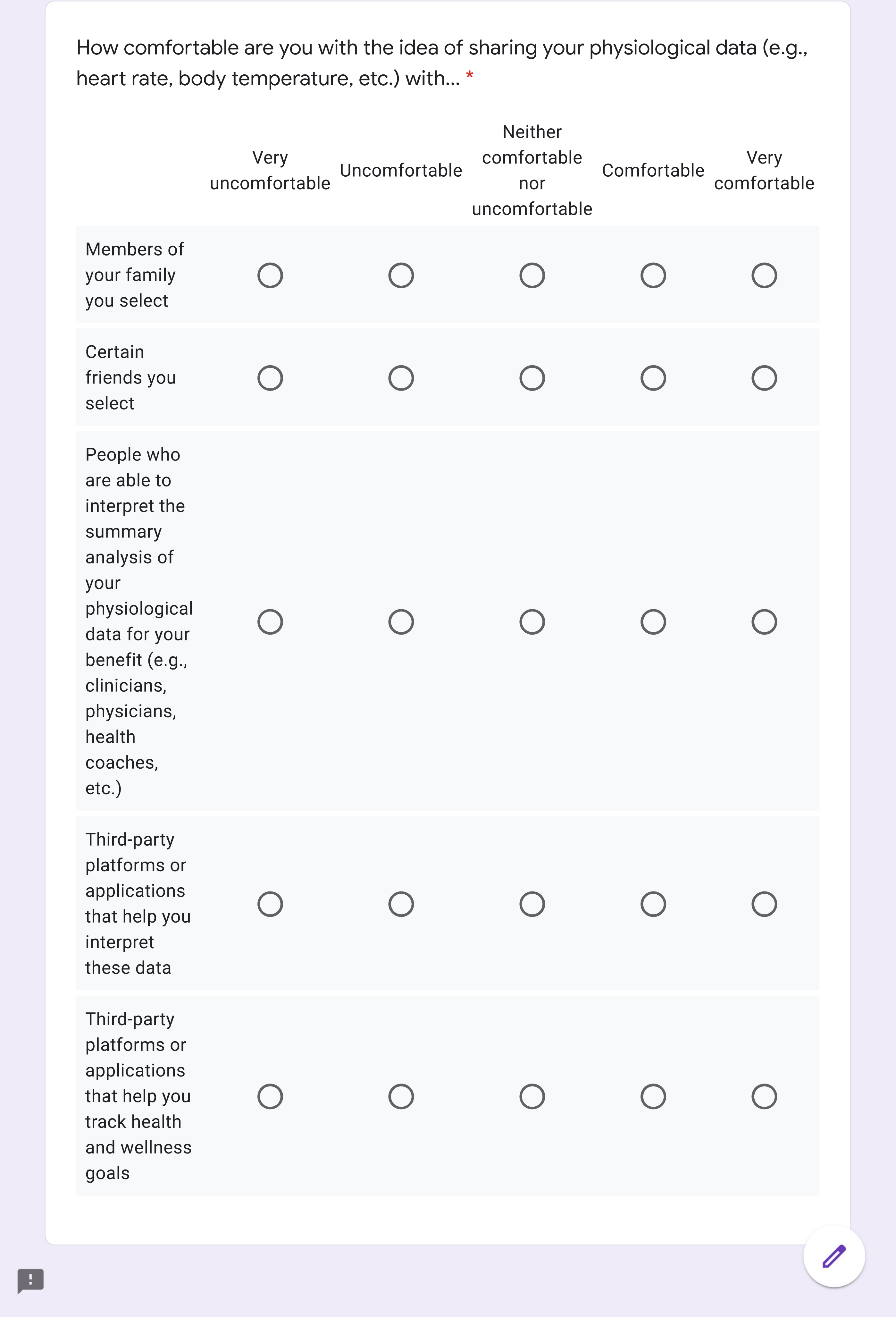
□ Friends

□ Clinicians or physicians

□ Health coaches

□ Platforms or applications

□ Other: ____________________



How comfortable are you with the idea of sharing your physiological data (e.g., heart rate, body temperature, etc.) with... *

	Very uncomfortable	Uncomfortable	Neither comfortable nor uncomfortable	Comfortable	Very comfortable
Members of your family you select	○	○	○	○	○
Certain friends you select	○	○	○	○	○
People who are able to interpret the summary analysis of your physiological data for your benefit (e.g., clinicians, physicians, health coaches, etc.)	○	○	○	○	○
Third-party platforms or applications that help you interpret these data	○	○	○	○	○
Third-party platforms or applications that help you track health and wellness goals	○	○	○	○	○

If you were given an opportunity to pick a wearable device you could use to track your sleeping behavior, which factors would influence more your choice of the devices? Please choose at most 3 factors: *

□ The device is easy to learn how to use.

□ The device brings some benefits to me.

□ The device is comfortable to wear physically.

□ The device has a reasonable price.

□ The device always provides feedback about what it is doing or measuring (i.e., I know what the device is doing at all times).

□ Wearing the device does not make me feel self-conscious (i.e., other people will not judge me for wearing the device).

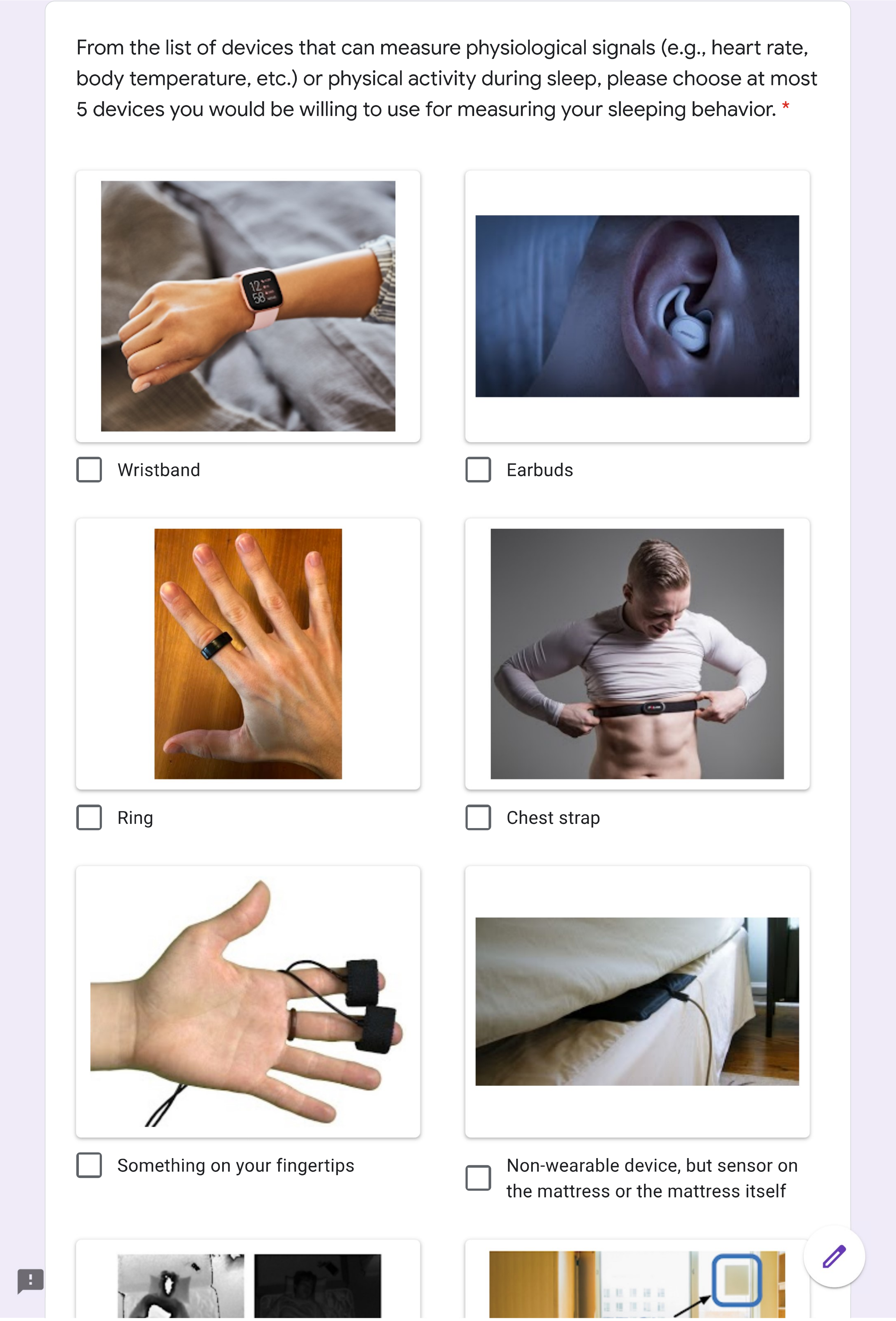
□ The device does not limit the way in which I like to perform my activity (i.e., not intrusive).

□ The device has a nice appearance.

□ The device does not require much effort to maintain (e.g., charging).

Are there other factors you consider important for the choice of the device? If yes, please describe below.

Your answer



From the list of devices that can measure physiological signals (e.g., heart rate, body temperature, etc.) or physical activity during sleep, please choose at most 5 devices you would be willing to use for measuring your sleeping behavior. *

□ Wristband

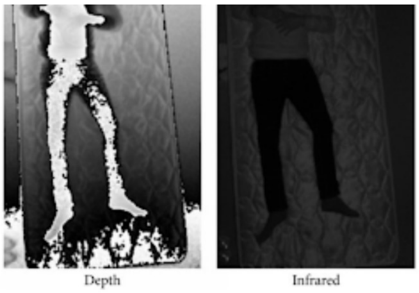
□ Earbuds

□ Ring

□ Chest strap

□ Something on your fingertips

□ Non-wearable device, but sensor on the mattress or the mattress itself

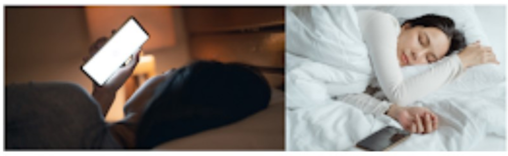


□ Depth or infrared Kinect sensor

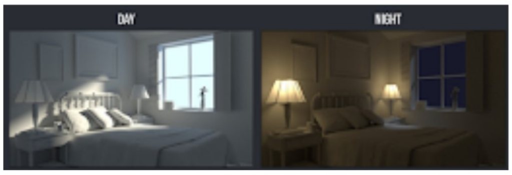
□ Contactless devices that use radio signals

□ Polysomnography (records your brain waves, heart rate, breathing, eye and leg movements through electrodes attached to your body and head)

□ Audio sensor collected from the smartphone

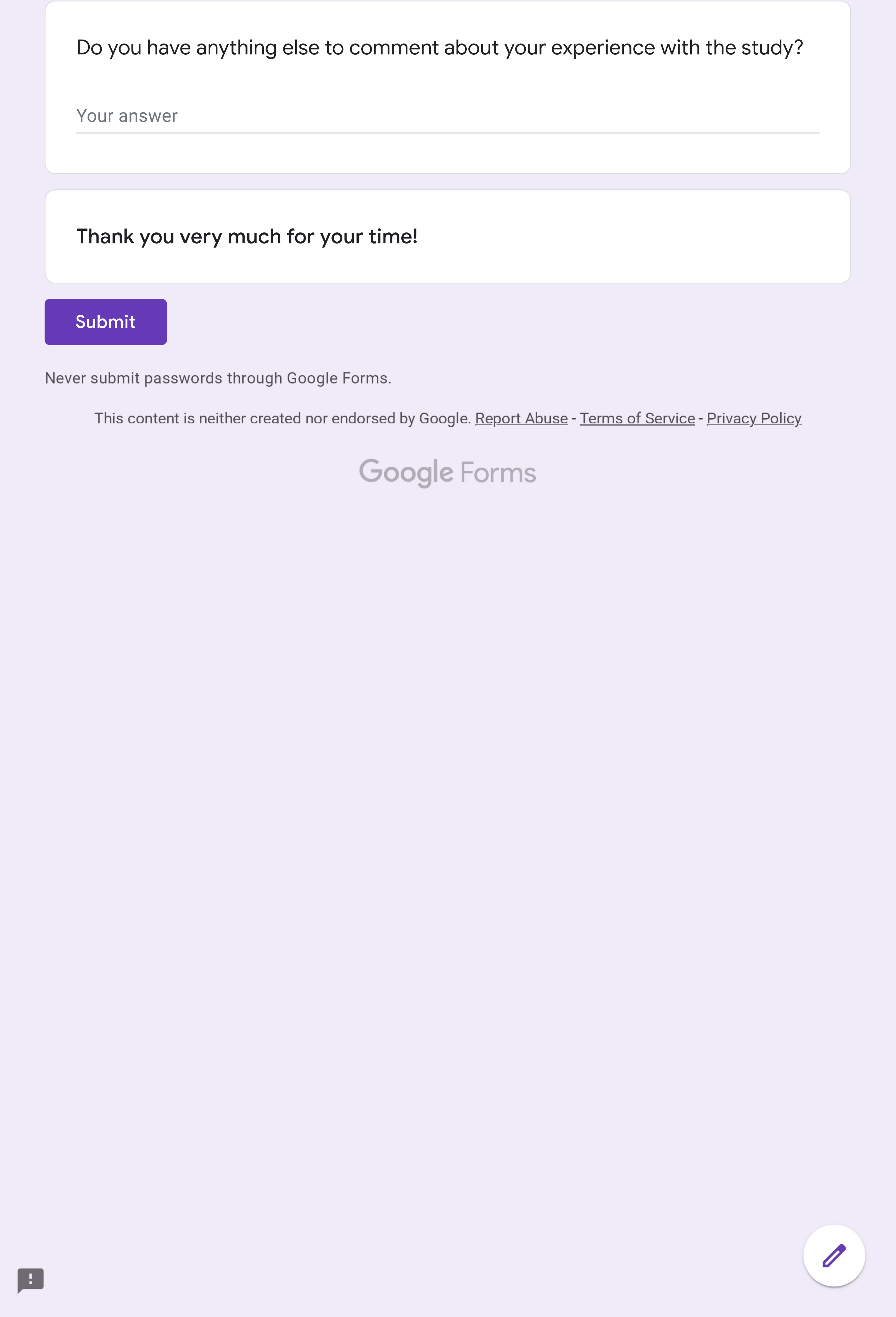


□ Smartphone usage (e.g., screen on/off, phone lock/unlock)



□ Environment sensors (e.g., amount of light)

□ Other:



Do you have anything else to comment about your experience with the study?

Your answer

Thank you very much for your time!

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Acronyms

AASM american academy of sleep medicine. 9

ACC 3-axis acceleration. 17, 38–40, 45, 48, 53, 54, 56, 61–63, 68

ANS autonomic nervous system. 1

BFI big five inventory. 14

BVP blood volume pulse. 17

ECG electrocardiography. 6, 113

EDA electrodermal activity. iii–v, vii, 2, 3, 10–12, 17, 33, 38–40, 45, 48, 53–55, 61, 63, 68, 113, 114

EEG electroencephalography. 6, 113

EMG electromyography. 6, 113

EOG electrooculography. 6, 113

HR heart rate. 1

HRV heart rate variability. 1

LSTM long short-term memory networks. 64

MCTQ munich chronotype questionnaire. 14, 15

NREM non rapid eye movement. 9, 10, 114

PII personally identifiable information. 21

PPG photoplethysmography. 17, 64

PSG polysomnography. 5, 6, 10, 113

PSQI pittsburgh sleep quality index. 14, 15, 42, 46, 49, 56, 58, 64

REM rapid eye movement. 9, 10, 113

SC skin conductance. 113

SCL skin conductance level. 113

SCR skin conductance response. 11, 113, 114

SE sleep efficiency. 8

SHAP shapley additive explanations). 53, 113

SNS sympathetic nervous system. 10, 114

SRMD sleep-related rhythmic movement disorder. 12

ST skin temperature. 17, 38–40, 63

SWS slow wave sleep. 9, 11, 114

TST total sleep time. 8

TWT total wake time. 8

USI università della svizzera italiana. 19

Glossary

SHAP (SHapley Additive exPlanations) Is an explanation model that use game theoretic approach (Shapley values) to understand the internal structure of machine learning methods. 53

Actigraphy Is a method of measuring sleep parameters and motor activity based on recording movements. 8

Artifact "changes in the recorded biosignal which do not stem from the signal source in question"Boucsein [2013]. 11

Electrocardiography (ECG) Is a technique for measuring the electrical function of the heart. 6

Electrodermal Activity (EDA) Observable changes on the skin. Also known as "Galvanic Skin Response" or "Skin Conductance (SC)". EDA has two main components: the skin conductance level (SCL) and the skin conductance response (SCR). 10

Electroencephalography (EEG) Is an electrophysiological monitoring method to record electrical activity of the brain using electrodes. 6

Electromyography (EMG) Is an electrodiagnostic medicine technique for evaluating and recording the electrical activity produced by skeletal muscles using electrodes. Two types of EMG: surface EMG and intramuscular EMG. 6

Electrooculography (EOG) Is a method for measuring the corneo-retinal standing potential that exists between the front and the back of the human eye. Electrooculography on the left eye (LEOG) and on right eye (REOG). 6

Polysomnography (PSG) Is a method of studying sleep based on brain activity (EEG), eye movements (EOG), muscle activity (EMG) and hearth rythm (ECG). The term "polysomnography" means readout (graph) of sleep (somnus) that is made up of multiple signals (poly). It is considered as the current gold standard for measuring sleep. 6

Rapid Eye Movement (REM) Is a period of sleep in which eyes rapidly dart from side to side underneath the lids, the brain activity is almost identical to the one that can be found in awake people. This sleep phase is strongly connected to dreaming, in fact, this phase is often describe as dream sleep or paradoxical sleep (because brain seems awake but body is clearly asleep). 9

Shapley value Concept taken from game theory, can be described as the quantification of the contribution that each player brings to the game . 53

Skin Conductance Response (SCR) Also called phasic component, SCR refers to peaks in the EDA signal. 11

Slow Wave Sleep (SWS) Is the deepest phase of non-rapid eye movement (NREM) sleep, phase 3. 9

Sympathetic Nervous System (SNS) Its main function is to provide energy by increasing a number of physiological parameters. It is mainly associated with the "fight or flight" response. 10

Bibliography

Alp Baran and Ronald Chervin. Approach to the patient with sleep complaints. *Seminars in neurology*, 29:297–304, 10 2009. doi: 10.1055/s-0029-1237116.

Wolfram Boucsein. *Electrodermal activity*. 01 1992. ISBN 978-1-4757-5095-9. doi: 10.1007/978-1-4757-5093-5.

Wolfram Boucsein. *Electrodermal activity: Second edition*. 08 2013. ISBN 978-1-4614-1125-3. doi: 10.1007/978-1-4614-1126-0.

A. Bujari and Bogdan Licar. Movement pattern recognition through smartphone's accelerometer. 01 2012. doi: 10.1109/CCNC.2012.6181029.

Orfeu Buxton and Enrico Marcelli. Short and long sleep are positively associated with obesity, diabetes, hypertension, and cardiovascular disease among adults in the united states. *Social science & medicine (1982)*, 71:1027–36, 09 2010. doi: 10.1016/j.socscimed.2010.05.041.

Daniel J. Buysse, Charles F. Reynolds, Timothy H. Monk, Susan R. Berman, and David J. Kupfer. The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry research*, 28(2):193 – 213, 1989. ISSN 0165-1781. doi: https://doi.org/10.1016/0165-1781(89)90047-4. URL `http://www.sciencedirect.com/science/article/pii/0165178189900474`.

John Cacioppo, Louis Tassinary, and Gary Berntson. *Handbook of psychophysiology*. 01 2007. doi: 10.13140/2.1.2871.1369.

Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, Javier Fernandez-Alvarez, Giuseppe Riva, and Cem Ersoy. Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches. *IEEE Access*, 8:38146–38163, 2020. doi: 10.1109/ACCESS.2020.2975351.

Mary Carskadon and William Dement. Normal human sleep: an overview. principles and practice of sleep medicine. m.h. kryger (ed.). *W.B. Saunders, Philadelphia*, pages 3–13, 01 1989.

Leandro Casiraghi, Ignacio Spiousas, Gideon P. Dunster, Kaitlyn McGlothlen, Eduardo Fernández-Duque, Claudia Valeggia, and Horacio O. de la Iglesia. Moonstruck sleep: synchronization of human sleep with the moon cycle under field conditions. *Science advances*, 7(5), 2021. doi: 10.1126/sciadv.abe0465. URL `https://advances.sciencemag.org/content/7/5/eabe0465`.

Jean-Philippe Chaput, Caroline Dutil, and Hugues Sampasa-Kanyinga. Sleeping hours: What is the ideal number and how does age impact this? *Nature and Science of Sleep*, Volume 10: 421–430, 11 2018. doi: 10.2147/NSS.S163071.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016. doi: 10.1145/2939672.2939785.

Zhenyu Chen, Mu Lin, Fanglin Chen, N.D. Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and A.T. Campbell. Unobtrusive sleep monitoring using smartphones. pages 145–152, 01 2013. ISBN 978-1-4799-0296-5. doi: 10.4108/icst.pervasivehealth.2013.252148.

Jongyoon Choi, Beena Ahmed, and Ricardo Gutierrez-Osuna. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 16:279–86, 09 2011. doi: 10.1109/TITB.2011.2169804.

Francois Chollet. *Deep learning with Python*. Manning Publications Co., USA, 1st edition, 2017. ISBN 1617294438.

Praveena Devi. A review on insomnia: The sleep disorder. pages 227–230, 12 2018.

Elena Di Lascio, Shkurta Gashi, and Silvia Santini. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2:1–21, 09 2018. doi: 10.1145/3264913.

Jessilyn Dunn, Ryan Runge, and Michael Snyder. Wearables and the medical revolution. *Personalized medicine*, 15(5):429–448, 2018. doi: 10.2217/pme-2018-0044. URL `https://doi.org/10.2217/pme-2018-0044`. PMID: 30259801.

Stephen Fairclough. Physiological data must remain confidential. *Nature*, 505:263, 01 2014. doi: 10.1038/505263a.

Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W. Picard, and Simone Tognetti. Empatica e3 — a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, pages 39–42, 2014. doi: 10.1109/MOBIHEALTH.2014.7015904.

Shkurta Gashi, Elena Di Lascio, Bianca Stancu, Vedant Das Swain, Varun Mishra, Martin Gjoreski, and Silvia Santini. Detection of artifacts in ambulatory electrodermal activity data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(2), June 2020. doi: 10.1145/3397316. URL `https://doi.org/10.1145/3397316`.

GlobalData. $54bn wearable tech industry offers immense opportunity for the construction industry by 2023. `https://www.globaldata.com/54bn-wearable-tech-industry-offers-immense-opportunity-for-the-construction-industry-by-2023/`, 2019. Accessed: 2021-04-09.

Alberto Greco, Gaetano Valenza, Antonio lanatà, Enzo Scilingo, and Luca Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE transactions on biomedical engineering*, 2016:797–804, 04 2016. doi: 10.1109/TBME.2015.2474131.

Aurlien Gron. *Hands-On machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc., 1st edition, 2017. ISBN 1491962291.

Li Guo. Quantified-self 2.0: using context-aware services for promoting gradual behaviour change. 10 2016.

Tian Hao, Guoliang Xing, and Gang Zhou. Isleep: Unobtrusive sleep quality monitoring using smartphones. 11 2013. doi: 10.1145/2517351.2517359.

Paul Hibbing, Samuel Lamunion, Andrew Kaplan, and Scott Crouter. Estimating energy expenditure with actigraph gt9x inertial measurement unit. *Medicine and science in sports and exercise*, 50, 12 2017. doi: 10.1249/MSS.0000000000001532.

Vanessa Ibáñez, Josep Silva, and Omar Cauli. A survey on sleep assessment methods. *PeerJ*, 6: e4849, 05 2018. doi: 10.7717/peerj.4849.

Somayeh Imani, Amay Bandodkar, Vinu Mohan, Rajan Kumar, Shengfei Yu, Joseph Wang, and Patrick Mercier. A wearable chemical–electrophysiological hybrid biosensing system for real-time health and fitness monitoring. *Nature communications*, 7:11650, 05 2016. doi: 10.1038/ncomms11650.

Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33:1–21, 07 2019. doi: 10.1080/08839514.2019.1637138.

O. P. John, E. M. Donahue, and R. L. Kentle. The big five inventory. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1991.

A. Kales, A. Rechtschaffen, Los Angeles. Brain Information Service University of California, and NINDB Neurological Information Network (U.S.). *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects: Allan Rechtschaffen and Anthony Kales, Editors*. NIH publication. U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, 1968. URL `https://books.google.it/books?id=Z41IvQEACAAJ`.

Daniel Kay, Jared Tanner, and Dawn Bowers. Sleep disturbances and depression severity in patients with parkinson's disease. *Brain and behavior*, 8, 03 2018. doi: 10.1002/brb3.967.

Kristen Knutson, Karine Spiegel, Plamen Penev, and Eve Van Cauter. The metabolic consequences of sleep deprivation. *Sleep medicine reviews*, 11:163–78, 07 2007. doi: 10.1016/j.smrv.2007.01.002.

Sylvia Kreibig. Autonomic nervous system activity in emotion: a review. *Biological psychology*, 84:394–421, 04 2010. doi: 10.1016/j.biopsycho.2010.03.010.

Kurt Kräuchi, Christian Cajochen, Esther Werth, and Anna Wirz-Justice. Functional link between distal vasodilation and sleep-onset latency? *American journal of physiology. Regulatory, integrative and comparative physiology*, 278:R741–8, 04 2000. doi: 10.1152/ajpregu.2000.278.3.R741.

Kurt Kräuchi, Christian Cajochen, and Anna Wirz-Justice. Waking up properly: is there a role of thermoregulation in sleep inertia? *Journal of sleep research*, 13:121–7, 07 2004. doi: 10.1111/j.1365-2869.2004.00398.x.

Antonio Lanatà, Gaetano Valenza, Alberto Greco, Claudio Gentili, Riccardo Bartolozzi, Francesco Bucchi, Francesco Frendo, and Enzo Pasquale Scilingo. How the autonomic nervous system and driving style change with incremental stressing conditions during simulated driving. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1505–1517, 2015. doi: 10.1109/TITS.2014.2365681.

Teofilo L Lee-Chiong. *Sleep: a comprehensive handbook*. John Wiley & Sons, 2005.

Julian Lim and David Dinges. Sleep deprivation and vigilant attention. *Annals of the New York Academy of Sciences*, 1129:305–22, 05 2008. doi: 10.1196/annals.1417.002.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran associates, Inc., 2017. URL `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

Gonzalo Martinez, Stephen Mattingly, Jessica Young, Louis Faust, Anind Dey, Andrew Campbell, Munmun Choudhury, Shayan Mirjafari, Subigya Nepal, Pablo Robles-Granda, Koustuv Saha, and Aaron Striegel. Improved sleep detection through the fusion of phone agent and wearable data streams. 03 2020. doi: 10.1109/PerComWorkshops48775.2020.9156211.

A. Carskadon Mary and C. Dement William. Distribution of REM sleep on a 90 minute sleep-wake schedule. *Sleep*, 2(3):309–317, 09 1979. ISSN 0161-8105. doi: 10.1093/sleep/2.3.309. URL `https://doi.org/10.1093/sleep/2.3.309`.

Luca Menghini, Dilara Yüksel, Aimée Goldstone, Fiona Baker, and Massimiliano de Zambotti. Performance of fitbit charge 3 against polysomnography in measuring sleep in adolescent boys and girls. *Chronobiology international*, 38:1–13, 04 2021. doi: 10.1080/07420528.2021.1903481.

Michelle Miller and Francesco Cappuccio. Inflammation, sleep, obesity and cardiovascular disease. *Current vascular pharmacology*, 5:93–102, 05 2007. doi: 10.2174/157016107780368280.

Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason Hong. Toss 'n' turn: smartphone as sleep and sleep quality detector. *Conference on human factors in computing systems - Proceedings*, 04 2014. doi: 10.1145/2556288.2557220.

Manivannan Muniyandi and Rahul Soni. Breath rate variability (brv) - a novel measure to study the meditation effects. *International journal of yoga*, Accepted, 01 2017. doi: 10.4103/ijoy.IJOY_27_17.

Ignacio Perez-Pozuelo, Bing Zhai, João Palotti, Raghvendra Mall, Michael Aupetit, Juan Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque. The future of sleep health: a data-driven revolution in sleep science and medicine. *npj Digital medicine*, 3, 12 2020. doi: 10.1038/s41746-020-0244-4.

Philips. World sleep day 2021. `https://www.usa.philips.com/c-e/smartsleep/campaign/world-sleep-day.html`, 2021. Accessed: 2021-04-29.

Andrew Phillips, William Clerx, Conor O'Brien, Akane Sano, Laura Barger, Rosalind Picard, Steven Lockley, Elizabeth Klerman, and Charles Czeisler. Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Scientific reports*, 7, 06 2017. doi: 10.1038/s41598-017-03171-4.

Rosalind Picard, Szymon Fedor, and Yadid Ayzenberg. Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Review*, 8, 03 2015. doi: 10.1177/1754073914565517.

Rana el Kaliouby Picard Rosalind W., Akane Sano. Palmar vs. forearm eda during natural sleep at home. *Media Laboratory*.

Emanuela Piciucco, Elena Di Lascio, Emanuele Maiorana, Silvia Santini, and Patrizio Campisi. Biometric recognition using wearable devices in real-life settings. *Pattern Recognition Letters*, 146:260–266, 2021. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2021.03.020. URL `https://www.sciencedirect.com/science/article/pii/S0167865521001070`.

Thomas Ploetz. Applying machine learning for sensor data analysis in interactive systems: common pitfalls of pragmatic use and ways to avoid them. *ACM Computing Surveys*, 54: 1–25, 07 2021. doi: 10.1145/3459666.

Xiaoli Qiang, Huangrong Chen, Xiucai Ye, Ran Su, and Leyi Wei. M6amrfs: Robust prediction of n6-methyladenosine sites with sequence-based features in multiple species. *Frontiers in genetics*, 9:495, 10 2018. doi: 10.3389/fgene.2018.00495.

Ivan Riobo, Agata Rozga, Gregory Abowd, Rosalind Picard, and Javier Hernandez. Using electrodermal activity to recognize ease of engagement in children during social interactions. 09 2014.

A. Roebuck, V. Monasterio, E. Gederi, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. Clifford. A review of signals used in sleep analysis. *Physiological measurement*, 35 1:R1–57, 2014.

Till Roenneberg and Martha Merrow. Munich chronotype questionnaire (mctq). `https://thewep.org/documentations/mctq`, 2003. Accessed: 2021-07-04.

Reza Sadeghi, Tanvi Banerjee, Jennifer Hughes, and Larry Lawhorne. Sleep quality prediction in caregivers using physiological signals. *Computers in Biology and Medicine*, 05 2019. doi: 10.1016/j.compbiomed.2019.05.010.

Sohrab Saeb, Ted Cybulski, Konrad Kording, and David Mohr. Scalable passive sleep monitoring using mobile phones: opportunities and obstacles. *Journal of Medical Internet Research*, 19: e118, 04 2017. doi: 10.2196/jmir.6821.

Akane Sano and Rosalind Picard. Toward a taxonomy of autonomic sleep patterns with electrodermal activity. volume 2011, pages 777–80, 08 2011. doi: 10.1109/IEMBS.2011.6090178.

Akane Sano and Rosalind Picard. Comparison of sleep-wake classification using electroencephalogram and wrist-worn multi-modal sensor data. volume 2014, pages 930–3, 08 2014. doi: 10.1109/EMBC.2014.6943744.

Akane Sano and Rosalind W Picard. Quantitative analysis of electrodermal activity during sleep. *Sleep*, 1(2Q):3Q, 2012.

Akane Sano, Andrew Phillips, Amy Yu, Andrew Mchill, Sara Taylor, Natasha Jaques, Charles Czeisler, Elizabeth Klerman, and Rosalind Picard. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. pages 1–6, 06 2015. doi: 10.1109/BSN.2015.7299420.

Akane Sano, Andrew J. Phillips, Sara Taylor, Andrew W. McHill, Conor O'Brien, Justin Buie, Cesar A. Hidalgo, Laura Barger, Charles A. Czeisler, Elizabeth B. Klerman, and Rosalind Picard. Influence of weekly sleep regularity on self-reported wellbeing. *Media Laboratory*, 2016.

Akane Sano, Weixuan Chen, Daniel Martinez, Sara Taylor, and Rosalind Picard. Multimodal ambulatory sleep detection using lstm recurrent neural networks. *IEEE journal of biomedical and health informatics*, PP:1–1, 08 2018. doi: 10.1109/JBHI.2018.2867619.

Sleepiz. Sleepiz. `https://sleepiz.com/`, 2021. Accessed: 2021-05-03.

Michael Smith, Christina Mccrae, Joseph Cheung, Christopher Harrod, Jonathan Heald, and Kelly Carden. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: An american academy of sleep medicine clinical practice guideline. *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, 14, 07 2018. doi: 10.5664/jcsm.7230.

Adriane Soehner, Kathy Kennedy, and Timothy Monk. Personality correlates with sleep-wake variables. *Chronobiology international*, 24:889–903, 02 2007. doi: 10.1080/07420520701648317.

Adriane Soehner, Kathy Kennedy, and Timothy Monk. Circadian preference and sleep-wake regularity: associations with self-report sleep parameters in daytime-working adults. *Chronobiology international*, 28:802–9, 11 2011. doi: 10.3109/07420528.2011.613137.

Flavia Sparacino. The museum wearable: real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences. 01 2021.

J. Stone, Lauren E Rentz, J. Forsey, J. Ramadan, R. Markwald, Victor S. Finomore, S. Galster, A. Rezai, and Joshua A Hagen. Evaluations of commercial sleep technologies for objective monitoring during routine sleeping conditions. *Nature and science of sleep*, 12:821 – 842, 2020.

Theresa Tanenbaum, Karen Tanenbaum, Katherine Isbister, Kaho Abe, Anne Sullivan, and Luigi Anzivino. Costumes and wearables as game controllers. pages 477–480, 01 2015. doi: 10.1145/2677199.2683584.

Sara Taylor, Natasha Jaques, Weixuan Chen, Szymon Fedor, Akane Sano, and Rosalind Picard. Automatic identification of artifacts in electrodermal activity data. volume 2015, pages 1934–1937, 08 2015. doi: 10.1109/EMBC.2015.7318762.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew Campbell. *StudentLife: using smartphones to assess mental health and academic performance of college students*, pages 7–33. 07 2017. ISBN 978-3-319-51393-5. doi: 10.1007/978-3-319-51394-2_2.

Rui Wang, Weichen Wang, Alex daSilva, Jeremy Huckins, William Kelley, Todd Heatherton, and Andrew Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2:1–26, 03 2018. doi: 10.1145/3191775.

Bing Zhai, Ignacio Perez-Pozuelo, Emma Clifton, João Palotti, and Yu Guan. Making sense of sleep: multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4:1–33, 06 2020. doi: 10.1145/3397325.

Hao Zhao, Jie-Yun Yin, Wanshui Yang, Qin Qin, Ting-Ting Li, Yun Shi, Qin Deng, Sheng Wei, li Liu, and Shao-Fa Nie. Sleep duration and cancer risk: a systematic review and meta-analysis of prospective studies. *Asian Pacific journal of cancer prevention : APJCP*, 14:7509–15, 12 2013. doi: 10.7314/APJCP.2013.14.12.7509.