FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Joonas Niemi
Roosa Risto
Saku Salo

DIGITAL SLEEP: EXPERT EVALUATION OF COMMERCIALLY AVAILABLE DIGITAL SLEEP TRACKERS

Bachelor's Thesis
Degree Programme in Computer Science and Engineering
June 2021

Niemi J., Risto R., Salo S. (2021) Digital Sleep: Expert Evaluation of Commercially Available Digital Sleep Trackers. University of Oulu, Degree Programme in Computer Science and Engineering, 57 p.

ABSTRACT

The use of digital technology has become a part of people's life globally. The rapid digitalisation has created opportunities to understand and improve our sleep by monitoring sleep patterns with physiological parameters. In this study, three expert evaluators collect and analyse the sleep data from five different sleep monitoring devices: Dreem2, Fitbit Versa 3, Polar M430, SleepScore Max and Withings Sleep Analyzer. The data consists of three weeks sleep data from all devices, each participant using all devices simultaneously for one week. This study focuses on the device's reliability on sleep monitoring, but also user experience is noticed with surveys filled in one week before assessment nights and one week during. In addition, technology expectation discussions were held to provide user perspective on the device. All devices had some problems recognising night-time wake ups and sleep restfulness. Reliability issues also came up when only two of five devices gave significantly lower sleep score after the user had many servings of alcohol. Devices recorded different amounts of sleep stages and had different algorithms to define sleep efficiency scoring to the user. The collected data and discussions from three expert evaluators, give an overall look on today's marketed sleep tracking equipment compered to each other highlighting some issues that came up using the devices and analysing the sleep data. These results can be used for future work, applying the sleep data gathering timeline into a study of larger group of participants, to improve sleep technology equipment.

Niemi J., Risto R., Salo S. (2021) Digital Sleep: Expert Evaluation of Commercially Available Digital Sleep Trackers. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 57 s.

TIIVISTELMÄ

Digitaalisten laitteiden käyttö on tullut osaksi ihmisten elämää maailmanlaajuisesti. Nopeasti kehittyvä teknologia on tuonut ihmisille mahdollisuuksia ymmärtää ja parantaa untaan tutkimalla sitä erilaisten mitattavien fysiologisten parametrien avulla. Tutkimuksessamme kolme asiantuntija-arvioijaa kerää ja analysoi unidataa hyödyntäen viittä eri unta mittaavaa laitetta: Dreem2, Fitbit Versa 3, Polar M430, SleepScore Max ja Withings Sleep Analyzer. Käsiteltävä unidata koostuu kolmen viikon ajalta kerätystä datasta, jonka aikana jokainen osallistuja on käyttänyt kaikkia laitteita samanaikaisesti yhden viikon ajan. Tutkimus keskittyy laitteiden luotettavuuteen unen mittauksessa, mutta myös käyttäjäkokemus on huomioitu kyselyllä, joka täytetään viikon ajalta ennen unimittauksia ja unimittausviikon aikana. Lisäksi käyttökokemus näkökulma on huomioitu teknologiakeskusteluilla. Jokaisen laitteen kanssa ilmeni joitain ongelmia hereillä olon tunnistamisessa. Luottamusta laitteisiin heikensi myös, kun vain pieni osa laitteista antoi käyttäjälle selvästi alhaisemman arvon unen palauttavuudesta yönä, jota edelsi usea nautittu alkoholiannos. Laitteiden mittaamat univaiheiden määrät vaihtelivat myös merkittävästi ja laitteiden eri menetelmät mitata unen palauttavuutta nousivat esille tutkimuksessa. Kolmen asiantuntija-arvioijan keräämä ja analysoima data sekä teknologiakeskustelut, antavat kokonaiskuvan viidestä nykyaikaisesta unenmittauslaitteesta vertailtuna keskenään, ja nostavat esiin ongelmia laitteiden mittauksissa. Saadut tulokset voidaan hyödyntää tulevaisuudessa tutkimuksessa, jossa unidataa kerätään suuremmalta joukolta käyttäjiä, ja näin ollen kehittää unta mittaavia laitteita.

TABLE OF CONTENTS

1. INTRODUCTION

The world has woken up to rapid digitalisation, but how does it impact our sleep? People all around the world use smart devices and those devices are finding their way to the bedroom. Using a mobile phone in bed is not only limited to children and adolescents, it is also a problem for adults and it is causing a decrease in sleep quality [1]. Link between poor sleep quality and excessive mobile phone usage has been studied often on university students and findings suggest such link exists [2][3]. But could we use the same smart devices that decrease our sleep quality to help better understand our sleeping habits and even improve them?

Nowadays many devices are capable to monitor user's sleep, and they utilise different methods to do it. Some are wearable, like a smart watch or a ring, others might be connected to the bed or sit on a user's nightstand. No matter the technology used, all devices still rely on the same basic functionality. Device collects data to track sleep and collected data is then visualised and presented back to the user in a human-friendly format. Furthermore, same data can be analysed to evaluate the quality and the patterns of sleep.

In this thesis we use expert evaluation on five different sleep tracking devices to compare the sleep data they collect. This study is conducted in three phases: pre-equipment, equipment, and post-equipment. In these phases we use a self-assessment survey and the devices for tracking sleep quality, and a group discussion to talk about our expectations for the different sleep tracking technologies and experiences using the devices.

2. RELATED WORK

2.1. Research Studies

2.1.1. Defining Sleep

It is often discussed how much sleep is a sufficient amount to get the best health benefits and feel fully rested waking up. In this section there will be discussion about what is quality sleep and why is it important to know the sleeping patterns when analysing the data collected from sleep monitoring devices.

Sleep plays an important role in life. It helps maintaining our physical and mental health and also increases day-time productivity and the safety of human being. Low sleep quality can be a factor to increase risk of chronic diseases, depression and accidents. Health care has maximised the capability of Internet of Things (IoT) technology by taking remote monitoring a part of their field. [4]

Sleep consists of cycles that are classified in five different stages: non-rapid eye movement (stages 1-4) and rapid eye movement (stage 5). During non-rapid eye movement (NREM) sleep stages are very light sleep, light sleep, deep sleep and very deep sleep. The final stage rapid eye movement (REM) includes vivid dreams and muscle atony.[5]

Sleep disorders happen when sleep cycles are irregular. To obtain a quality sleep it is important to get the right proportion of every sleep stages and a sufficient amount of sleep cycles [4]. This is why analysing sleep is more complex than just counting the hours of sleep. For getting in to the root of sleep analyse, we need to know the physiological parameters that indicate in which stage of sleep human is at. After defining these parameters it is possible to analyse the data collected by the sleep monitoring devices.

2.1.2. Monitoring Sleep

To gain the insight about the sleep quality of the patent, there are many physiological parameters that can be monitored during sleep [4]. Polysomnography (PSG) is a study that uses collected data of sleep patterns to identify any sleep disorders a patient may have. These physiological parameters are heart rate, eye movement, skeletal muscle activity, brain waves, breathing rate and blood oxygen levels. [6] A PSG sleep study is usually done as an overnight assessment at a sleeping center [7].

Nowadays because of the new IoT technology some of these parameters can be monitored at home by using wearable devices such as a watch or a ring. Although heart rate monitors have been in daily use for some people quite a while, a new group of devices for home sleep monitoring have emerged quite recently. Devices such as a headband use EEG electrodes to measure brain activity and others such as smart sleeping mattresses use movement of the patient to capture more information about the sleep cycles. Because of the fast developing technology, these devices get cheaper and easier to use at home with no expert supervision. This enables longitudinal data

collection, which makes collecting reliable data easier. [7]

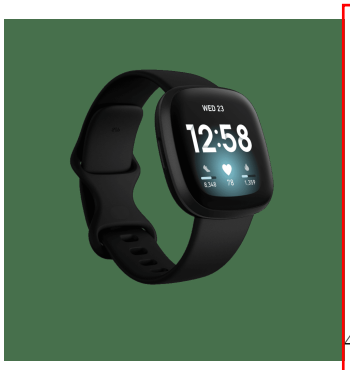
2.2. Industrial Products

For this study we utilise several different wearable and non-wearable devices which measure sleep in various ways. These devices include Polar M430 wrist watch which utilises trademarked feature Polar Sleep Plus™ to measure sleep. Polar Sleep Plus™ bases its objective measurements of sleep quality, timing and duration to wrist-acceleration measurement [8]. Fitbit Versa 3 wrist watch measures sleep stages and score via motion sensor, heart rate sensor and more [9].
Dreem2 headband, a reduced-montage dry-electroencephalographic (EEG) device measures brainwaves via five EEG sensors, heart rate via red-infrared pulse oximeter, movement, position and breathing frequency via 3D accelerometer. The EEG measured with Dreem2 correlates with polysomnographys (PSG) EEG results. [10]
Withings Sleep Analyzer [11] under-mattress sleep tracker measures light, deep and REM sleep cycles and onset, timing and wake up time of sleep. Heart rate, respiratory rate and body movements are measured via pneumatic sensor. Snoring and cessations in breathing are measured via sound sensor. Sleep Analyzer is dissimilar to other devices in our study since it is not a wearable device, instead it is placed under the mattress of the user.
Sleepscore Max sleep tracker is a non-wearable device that can be placed on a bedside table [12]. It uses a ultra-wideband radar sensor to measure sleep quality and quantity while also tracking room temperature and light level[13]. After each night, it provides a SleepScore from 0 to 100 based on six criteria: total sleep duration, time to fall asleep, light sleep, deep sleep, REM sleep, and awakenings throughout the night[14]. Sleepscore Max's most distinct feature is that it does not need to be connected to the user or the user's bed.

(a) Dreem2 Headband [1]

(b) Withings Sleep Analyzer [2]

(c) Sleepscore Max [3]



(d) Fitbit Versa 3 [4]

(e) Polar M430 [5]

Figure 1. Devices used in our study

[1](a) https://shop.dreem.com/europe/themes/dreem/img/dreem/us/Headband.png

[2](b) https://image-cache.withings.com/pages/products/sleep-analyzer/media/product/withings-sleep-analyzer-desktop.jpg?&w=1600&dpr=1

[3](c) cdn.shopify.com/s/files/1/2065/4897/files/Accuracy_you_can_trust_480x480_24d175e7-002f-45a4-b118-cda6b322b4a3_480x480.png?v=1597099575%20(sleepscore)%E2%80%8B

[4](d) https://www.fitbit.com/global/content/dam/fitbit/global/pdp/devices/versa-3/device-360/black/prod0.png

[5](e) https://www.polar.com/sites/default/files/product2/600x600/polar_m430_black_600x600.png

3. STUDY DESIGN

In this chapter we discuss how the data collection and measurement was done in our study. We divided our study into three phases: pre-equipment, equipment and post-equipment. The pre-equipment phase contains all the collected data before using the sleep monitoring devices, equipment phase focuses on the data that is collected with the devices and lastly the post-equipment phase contains our discussion after using the devices. The data was collected from and examined by three expert evaluators as participants (P1, P2 and P3).

3.1. Phase 1: Pre-Equipment

In pre-equipment phase, we focused on collecting self assessment data from participants sleep without using the sleep tracking devices. This was done to get a better knowledge on whether the sleep monitoring equipment had an effect on sleep quality when using them on the equipment phase.
To collect data from the participants sleep quality during the past month, we used the Calculate by QxMDs' Pittsburgh Sleep Quality Index survey [15].
After filling in the survey, we got a numeric value of our overall sleep quality and stored it in a folder as a PDF file. The survey was used at the start of our project and at the end to compare the overall sleep quality before and after using the sleep monitoring devices.
As another method to help us get an overall look on participants sleep quality, we used a daily self-assessment sleep tracking survey for a period of one week before and during the sleep monitoring experiment.
We also took notes to a sleep diary on what in the last day could have affected on participants sleep during both the pre-equipment and the equipment phase. On top of that we discussed and took notes on our own technology expectations regarding to the devices we were going to be using in the second phase.
The daily questions about sleep quality are coded as D1-D8, and are as follows:

- D1: "When did you fall asleep yesterday?": Likert scale 1-5 (1 = earlier than normal, 5=later than normal)
- D2: "When did you wake up and leave the bed this morning?": Likert scale 1-5
- D3: "How long did it take to fall asleep after you went to bed yesterday?": Likert scale 1-5
- D4: "How often do you remember waking up during last night?": Likert scale 1-5
- D5: "In your own opinion, how well did you sleep last night?": Likert scale 1-10 (1 = a lot worse than normal, 10 = much better than normal)
- D6: "In your own opinion, how well rested did you feel after waking up today?": Likert scale 1-10 (1 = more tired than usual, 10 = more refreshed than usual)

- D7: "If you remember, please state when exactly you went to bed yesterday (hour:minute, e.g., 8:30)"
- D8: "If you remember, please state when exactly you got up from bed this morning (hour:minute, e.g., 8:30)"

3.1.1. Technology Expectations

During pre-equipment phase we had a discussion focusing on six different criteria on technology expectations: ease of use, functionality, reliability, usefulness, comfortability and sleep disturbance.
Ease of use refers to how effortless the device is to use. The devices competence, functions and features are discussed under the functionality criteria. Reliability refers to how flawless the manners of the device are and usefulness concludes the users opinions on how beneficial the solutions are for the devices purpose. [16] Comfortability refers to how it feels to wear the device regarding comfort, and sleep disturbance highlights the irritating features such as light or noise disturbance. We used these criteria as a guide to our conversations about the devices and took notes for the evaluation part.

3.2. Phase 2: Equipment

In the second phase 'Equipment' we focused on tracking participants sleep with cutting-edge sleep tracking devices by monitoring our sleep for one week each with five different devices: Polar M430, Fitbit Versa 3, Dreem2, Withings Sleep Analyzer and Sleepscore Max. At this point we were also filling the same daily sleep tracking survey as in phase 1 without examining the collected data beforehand. It was important to do the self-assessment surveys first page before watching the collected data, to keep the answers true and accurate to the users own opinion on their sleep. After this the data from the sleep tracking devices could be examined and the survey opened up next page with questions about the devices.
After monitoring our sleep daily for one week and filling the self-assessment online survey and our sleep diaries, we collected the CSV-form data from all devices and screenshots from mobile apps including the app name, date and participant number.
The daily questions about use of digital sleep trackers are coded as T1-T4, and are as follows:

- T1: "In your opinion, select the devices that accurately logged your sleep data from last night (including bedtime, wakeup time, and sleep phases):"
- T2: "In your own opinion and assessment, select all sleep-related variables which were inaccurate (or seemed like inaccurate) in last night's sleep data according to each device (including bedtime, wake up time, sleep phases, night-time wake ups or sleep restfulness, heart-rate and related variables, sleep latency and total amount of sleep)"

- T3: "Select all devices which you experienced some form of problem, error or other malfunction during last evening, night, or morning"
- T4: "How well does the device provided sleep metric (sleep score, sleep efficiency, or similar variable) correspond to your feeling of alertness and restfulness in the morning? Rate on a scale of 1-5 (1: does not correspond, 5: corresponds really well)"

3.3. Phase 3: Post-Equipment

After collecting the data from the sleep tracking equipment for one week, we moved on to phase 3 'post-equipment', which examines the user experience of our expert evaluators after using the sleep tracking equipment. Post-equipment phase included filling in once again the Pittsburgh Sleep Quality Index survey and discussing if self-reflection had an impact in sleep quality.
We also repeated the discussion of technology expectations using the same criterias as in pre-equipment phase, this time confirming which expectations were accurate and which were proven irrelevant.

4. STUDY IMPLEMENTATION

In this chapter we present the results of our three data collection phases. Data consists of survey question answers, opinions and measured data of participants P1, P2, P3 in pre-equipment, equipment and post-equipment phases.

4.1. Participant Introduction

P1 is a 25 year old male university student. P2 is a 22 year old female university student. P3 is a 26 year old male university student who works part-time. None of our participants have apparent sleeping disorders. (Table 1)

Participant	Sex	Age	Pre-PSQI	Perceived sleep problems
P1	M	25	7	None
P2	F	22	5	None
P3	M	26	11	None

Table 1. Table of Participants information

4.2. Data Gathering Methods

Pittsburgh Sleep Quality Index

Pittsburgh Sleep Quality Index was created by the researchers at the University of Pittsburgh in 1988. It is a self-rated questionnaire designed to evaluate sleep quality and disturbances over a 1-month period. PSQI consists of 19 individual questions and it creates global score based on seven, separately scored, component categories: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleep medication, and daytime dysfunction. [17]

4.3. Phase 1: Pre-Equipment

In the first phase Pre-equipment, we filled in the PSQI survey to have it to compare at the end of our sleep monitoring week. We had one week lasting informal sleep diary period, where we kept track of what in the last day could have affected our sleep. These factors could have been for example exercising, stress or alcohol. Phase 1 also included technology expectations discussion.

4.4. Phase 2: Equipment

On phase 2 we focused on monitoring our sleep with the sleep tracking equipment for one week each. We also filled in the sleep tracking survey from phase one with the additional questions about the device accuracy and possible problems during the night presented in the design plan.

4.4.1. Data Gathering Timeline

When monitoring sleep with the sleep tracking equipment, we followed a data gathering timeline and did the same steps each expert evaluator.

Night

At night the devises were placed correctly and turned on the record mode if needed. We put the watches Fitbit Versa 3 and Polar M430 on the same wrist to give the devices similar starting point. Dreem2 headband was placed on the head, Withings sleep analyzer under the mattress and SleepScore max on bedside table according to the instructions from the companies websites. With Withings sleep analyzer mattress it was important to place it under your own mattress about the hight of your chest and once plugging it in to wait about 10 minutes for it to calibrate the mattress. SleepScore max was placed towards the sleepers head and not too far away.

Morning

Waking up after the night, we ended the recording of the night if needed and transferred the data into mobile applications on our phones. It was important to remember to put the devices to charge if needed and also not to check the sleep data right away, since we wanted to have a clear view on our own feelings of the sleep before comparing it to the devices data. After every night we tracked the sleep by filling in our sleep diaries, a Google forms sleep survey presented on the design part and taking screenshots from all the mobile applications.

After sleep tracking week

After monitoring our sleep for one week, we filled in the PSQI survey once again, and collected all the data into one Google drive folder each for the data analysis. These folder concluded CSV-formed data collected from all the devices either directly downloading it from the application or manually writing down all the visible data, screenshots from every night and our PSQI surveys filled before and after monitoring our sleep with the sleep tracking equipment.

4.5. Phase 3: Post-Equipment

After the sleep tracking from phase 1 and 2, we moved on to examine the users experience of the devices and their sleep. The experiences were gathered from the expert evaluators. We did a technology expectations discussion and also ranked the devices based on our own user experience and the inter-rater reliability.

5. RESULTS

5.1. Qualitative Results

5.1.1. Technology Expectations

We discussed about expectations of the technology we are going to use to measure our sleep and based the discussion on our six criteria:

Pre-determined themes

Ease-of-use
We expected Sleepscore Max and Withings Sleep Analyzer to be the easiest to use since they do not require to be attached to body while going to sleep and are already in place. Dreem2 would require a little effort when going to bed, because you need to put the headband on every time. P1 thought that Polar M430 and Fitbit Versa 3 would be quite easy too since P1 wears wristwatch on regular basis already, nevertheless adding one more watch on same wrist would require more work when going to sleep even if one would be in full-time use.

Functionality
In the discussion the functionality of Sleepscore Max was doubted since it lacks contact with the participant. Dreem2 was expected to be most functional because it measures brain waves via electroencephalography (EEG). Withings Sleep Analyzer divided opinions, P1 expected it be functional and P2 had doubts about how well it can measure through a thicker mattress and P3 was concerned if the the length of the Sleep Analyzer was sufficent, in case participant is moving during sleep.

Reliability
We expected Dreem2 to be reliable in sleep tracking. In contrast we had little trust in reliability of Withings Sleep Analyzer and Sleepscore Max since they are not attached to the participant and therefore can cause errors in measurements, for example rolling off the Sleep Analyzer or being too far from Sleepscore Max or sleep with a partner. We were a little concerned about battery life on Polar M430, Fitbit Versa 3 and Dreem2 in the event of forgetting to charge the device. We expected Polar M430 to give more reliable data than Fitbit Versa 3 since M430 has specific feature to measure sleep (Polar Sleep Plus™) whereas Versa 3 does not have apparent corresponding feature. P3 thought that if Polar M430 and Fitbit Versa 3 are not weared tight enough it might produce erroneous measurements.

Usefulness
In the discussion we thought that Dreem2 would be most useful in measuring sleep but Withings sleep analyzer would suit everyday life better and still produce useful data. P2 discussed that Polar M430 and Fitbit Versa 3 would not be as useful in sleep measuring as for overall well-being and fitness. We do not expect too much from Sleepscore Max in usefulness.

Emerging themes

We initially chose ease-of-use, functionality, reliability, and usefulness as our technology expectation themes. But before the discussion, we noticed that our pre-determined themes were good for evaluating active and intended user experience, but not very suitable for evaluating passive user experience e.g. attributes that can negatively affect sleep tracking. These could be any type of lights or noises the device produces during use or the device is just uncomfortable to use. To better categorise our expectations, we added comfortability and sleep disturbance to our list of technology expectation themes.

Comfortability

We expected Dreem2 headband to be somewhat uncomfortable to sleep with and Sleepscore Max to be unnoticeable when sleeping since it has no contact with participant. P3 was concerned if tightness required from Polar M430 and Fitbit Versa 3 would be uncomfortable especially to people who do not wear wrist watches. Withings Sleep Analyzer divided opinions, P1 thought that it might be quite comfortable if it is not too thick and P3 thought that it might be uncomfortable while moving during sleep.

Sleep disturbance

For the sake of this work we have to wear Polar M430 and Fitbit Versa 3 in the same hand, we expected that to be quite annoying and that it could make noise if they hit each other. P2 thought that the measuring itself might disturb sleep because of all the devices used together. P3 was concerned of light disturbance from the M430 and Versa 3 during sleep and if the apps will disturb you with notifications.

5.1.2. Realisation of Technology Expectations

Post-equipment phase contained a discussion similar to phase 1, where the expert evaluators would share and compare their thoughts and previous expectations on the devices and log it.

Ease-of-use

	Expectation	Realisation
Dreem2	most effort; takes time to put on	""
Fitbit Versa 3	quite effortless for watch users	quite effortless, but fastening method could have been better
Polar M430	quite effortless for watch users	""
SleepScore Max	effortless	quite effortless; needed manual starting
Withings Sleep Analyser	effortless	quite effortless; 10 min calibration on unpluggin

Table 2. Ease-of-use

In the discussion all expert evaluators agreed that the sleep monitoring with Dreem2 needed manual starting and ending in addition to putting the headband on every night, which needed a little more effort than we initially tought. P3 also brought up that it needed extra effort to *"pull long hair through headband when you have your hair open"* (P3)".
Withings Sleep Analyzer, as expected, was really effortless to use, however if you unplug it after the night as P1 and P2 did, 10 minute calibration had to be waited before going to bed which caused some extra effort.
Sleepscore Max was easy to use and did not need much effort, like Dreem2 the sleep monitoring in Sleepscore Max needed manual start form the app. Also P3 said: *"Most effortless after Withings Sleep Analyzer when you keep devices plugged in and apps open"* (P3).
Fitbit Versa 3 was easy to use, no need for starting tracking manually, wear when going to bed and it will track the night. Nevertheless the straps fastening method was a little hard to use.
Polar M430 was easy after realising that both watches measure sleep automatically.

Functionality

	Expectation	Realisation
Dreem2	most functional; EEG technology	functional; sleep stages + white noise feature
Fitbit Versa 3	no expectation	many functions; sleep stages, SpO2, heart rate (paid feature)
Polar M430	no expectation	worst functions; no sleep stages monitored, adjustable sleep time
SleepScore Max	poor; the lack of contact	functional; sleep stages, ambient temperature and luminosity
Withings Sleep Analyser	quite functional if monitors correctly	fuctional; detects snoring, sleep apnea and sleep stages

Table 3. Functionality

Expert evaluators discussed that Dreem2 is specialized in measuring sleep stages which Dreem2 does via EEG. Dreem2 also was only device which had sleep aiding audio (such as Pink noise and ocean waves) *"but only audio available for assessment period (one week) was ocean waves"* - P2. Dreem2 also had "pleasant smart alarm" which was tested by P1 and P2. Withings Sleep Analyzer in addition to sleep stages detected snoring and sleep apnea. Expert evaluators concluded that functionality was good.

Sleepscore Max tracked sleep stages, however not heart rate. Instead this was only device which measured ambient temperature and luminosity. Also Sleepscore max asked every night how much caffeine and alcohol you have ingested and if you have exercised or stressed during the day.

Fitbit Versa 3 tracks sleep stages, and *"only device I noticed to track blood oxygen saturation (SpO2) during the night"* - P1. However Versa 3 did not track heart rate during sleep, there were paid feature for this though.

Polar M430 did not measure sleep stages at all, only when you were awake or asleep during the night. It did not track heart rate for P2 and P3 "*it tracked heart rate for me in 'exercises'-page but not on 'sleep'-page*" - P1. However Polar had adjustable sleep time so wrong sleep time could be corrected by user. Polar M430 was the only one to give written evaluation of sleep after each night.

Reliability

	Expectation	Realisation
Dreem2	reliable; EEG technology	quite reliable; especially with sleep stages, but some connectivity and attachment issues
Fitbit Versa 3	less reliable than Polar M430	not that reliable; much awake time and some recognition issues
Polar M430	more reliable than Fitbit Versa 3; Polar Sleep Plus™	not reliable on sleep measuring; much awake time and no sleep stages
SleepScore Max	not reliable; not attached to the user, distance to the device	more reliable than expected; sometimes scoring seemed too good
Withings Sleep Analyser	not reliable; not attached to the participant, rolling over	quite realible; data seemed reliable but some issues with calibration and sleep recognition

Table 4. Reliability

Dreem2 headbands data was reliable in the expert evaluators assessment, especially sleep stages, however few things came up in discussion: For P3 the headband did not connect to the mobile app one night, and P2 had some troubles with headband coming off from the head during the night and had some doubts if thick hair would interrupt (back) sensors. P1 reported that headbands battery will die if not charged full every day before going to sleep: *"Even about 60% battery charge did not last through whole night"(P1).*

Withings Sleep Analyzer was also thought to be quite reliable in measuring sleep stages, nevertheless there was some issues too: P1 and P2 did not keep Withings Sleep Analyzer plugged in all the time so they reported some issues with calibration, and Withings Sleep Analyzer did not record one night for either one. *"For some reason mattress did not calibrate, and did not track sleep for one night"(P2).* P3 did not have such problem with calibration since mattress was kept plugged in, instead mattress had some sleep recognition issues, *"At times it thought I fell asleep even if I was just lying on the bed"(P3).*

Sleepscore Max was more reliable than expected in expert evaluators' opinion. *"I trusted the data surprisingly much, though sometimes it gave very good scores even if I felt that I slept more poorly" (P1). "I think the Sleepscore Max were more reliable than I expected"(P2). "In my opinion watches and Sleepscore Max were almost on the same level"(P3).*

Fitbit Versa 3 did not quite reach the same reliability as Dreem2 or Withings Sleep Analyzer in our expert evaluators opinion, but was still quite okay. What made Versa 3 more unreliable was the amount of wake time during night for all expert evaluators,

"sometimes even an hour a night, mostly a large amount of short wakings"(P1). P1 and P2 had also issue with sleep recognition *"Fitbit thought I fell asleep for few hours even if I was awake in bed" (P2).*
Conversely to their expectations expert evaluators did not find Polar M430 reliable in sleep tracking. Like Fitbit Versa 3, Polar M430 measured a lot of waketime consisting of short wakenings for all expert evaluators. Also *"Because Polar did not offer much data, not even sleep stages, it did not seem very reliable"(P2).*
There were some things that came up in the discussion that applied to all devices: *"It felt like the devices might have been a little positive in their sleep assessments (sleep scores and -efficiencies)."(P2)* and *"Not one device seemed to notice that when I woke up to first alarm, checked email from phone and went back to sleep on working days"(P3).*

Usefulness

	Expectation	Realisation
Dreem2	most useful data from sleep	
Fitbit Versa 3	quite useful	quite useful; overall look of the night
Polar M430	quite useful	not useful; not much useful or interesting data
SleepScore Max	not very useful	quite useful; some scoring left unexplaned
Withings Sleep Analyser	suits everyday life and produces useful data	

Table 5. Usefulness

The data collection raised few concerns of the data usefulness, since registering to the applications with Apple profile brought up problems with the CSV-data downloading. P1 and P3 registered to the applications with Google and that way all the other CSV-data could be easily collected but Polar M430 did not have this feature at all. *"Polar M430 did not have a feature to download the CSV-formed data, so all the expert evaluators had to write the values by hand." - P1 "Using the Apple profile it was not possible to gather the CSV-data from Withings Sleep Analyser or the Sleepscore Max, since registering with Apple it automatically created a new random email address and the data gathering did not work with that.*" - P2.
All expert evaluators thought that the data from Dreem2 and Withings Sleep Analyser was reliable and useful. Sleepscore Max was following the same tracks but some values and scores were left a bit unexplained. For example the device gave sleep score, body score and mind score all separately, which left the evaluators a bit confused on the meaning of all of these.
Fitbit Versa 3 gave an overall look on the night and its sleep stages but none of the expert evaluators would purchase it for only sleep monitoring purposes. *"I could buy Fitbit Versa 3 for general use but not just to measure sleep" - P1.*
All the expert evaluators thought that Polar M430 was a bit of a wild card to the sleep

tracking field and study. Since it is not created to monitor sleep stages and to give that much data of the sleep, it was not very useful on tracking sleep.

Comfortability

	Expectation	Realisation
Dreem2	somewhat uncomfortable; attached to head	somewhat uncomfortable when being too loose on the head
Fitbit Versa 3	could be uncomfortable to people who do not wear wrist watches	quite comfortable
Polar M430	could be uncomfortable to people who do not wear wrist watches	quite comfortable
SleepScore Max	comfortable; unnoticeable since no attachment to user	comfortable; no attachment to user
Withings Sleep Analyser	devided opinions; comfortable if unnoticeable and quiet while moving	comfortable; no attachment to user, but sometimes made a small sound

Table 6. Comfortability

Comfortability was an emerging theme on the first technology expectations discussion. It was added to the discussion, to highlight the users thoughts on the devices comfort to use, since it is an important factor on thinking which sleep monitoring device to purchase and use.
Dreem2 brought up concerns in all expert evaluators, since it sometimes moved on the head in a way, that it might not monitor the sleep as well. *"I had the smallest extension part but still thought the headband was not tight enough, because it moved and also came off the head multiple times. Sleeping on my stomach, I noticed the headband would always come off during the night.*" - P2 *"At first the headband was too loose so I changed the extension part to a smaller one. After a couple of nights I would say I did not notice the headband anymore.*" - P1 *"There could have been a smaller extension part, since the device moved on the head which caused worries.*" - P3 *"Also with long hair the devices added some discomfort when it would get stuck on the hair, but this happened only once during the week."* - P2.
Withings Sleep Analyser was very comfortable to use. *"Sometimes the mattress made a small creaky sound during the night"* - P3. But other than that all the expert evaluators agreed that the mattress was almost unnoticeable during the night. Like Withings Sleep Analyser also Sleepscore max was comfortable to use since it did not touch the body at all. *"Duration notices were the only thing that bothered.*" - P3. P1 and P2 did not get these notices because they did not have the app locked open on the background all the time.
The comfortability of Fitbit Versa 3 and Polar M430 were compared together and raised some thoughts. All the expert evaluators thought that Fitbit Versa 3 was more comfortable on the wrist during the night than the Polar M430. The only concern

about Fitbit Versa 3 came from P2: *"Adjusting the size of the bracelet was in big steps, and it happened that I was between two sizes, so the watch was either too loose or too tight.*" - P2, P1 and P3 found a right size to them. Polar M430 was big and a little uncomfortable on the hand since it was a bit heavier one, agreed all the expert evalators.

Sleep disturbance

	Expectation	Realisation
Dreem2	no expectation	somewhat disturbing; headband did come off some nights
Fitbit Versa 3	watches together could be disturbing on the wrist	not much sleep disturbance; light could be turned off
Polar M430	watches together could be disturbing on the wrist	not much sleep disturbance
SleepScore Max	no expectation	did not disturb the sleep at all
Withings Sleep Analyser	no expectation	had little affect on sleep; calibration started again at least once

Table 7. Sleep Disturbance

Sleep disturbance was an added parameter that came up during the first technology expectations discussion. It was defined to focus on the sleep time disturbing features of the devices, such as a sound or light feature that would affect the sleep.
Dreem2 headband raised concerned about its attachment to the head. *"During the sleep tracking week, especially at the start of it, I woke up many times checking if the headband was still on my head, since it came off a couple of times and I was worried if the recorded data would suffer."* - P2. Also P1 mentioned that turning in bed did sometimes make the headband to come off the head or turn in a way that the monitors were not completely attached to the head.
On most nights Withings Sleep Analyser did not disturb the sleep at all. *"At least once the calibration started again when I went to bed, which meant I had to get up and wait the 10 minute time period, but other than that Withings sleep analyser did not disturb my sleep at all.*" - P2
Since Sleepscore Max stands on its own on the bedside table without any attachments to the body, expert evaluators did not record any sleep disturbance from it. It did not make any sounds during the night and the green "device connection" -light turned off when the night started, so there were no light disturbance eighter.
Fitbit Versa 3 had a sleep mode, which made the device dark during the night. This way there was not light disturbance when moving the wrist during the night. P2 said both of the watches Fitbit Versa 3 and the Polar M430 were mildly uncomfortable on the hand, but also added that this could be because the watches were placed on the same hand and that the fastening mechanism of Fitbit Versa 3 had a larger scale, so it was harder to get the right tightness adjusted. Polar M340 did not affect the sleep negatively in any other way according to the expert evaluators.

5.1.3. Ranking Devices

	P1	P2	P3
1.	Dreem2	Withings Sleep Analyzer	Dreem2
2.	Withings Sleep Analyzer	Dreem2	SleepScore Max
3.	SleepScore Max	SleepScore Max	Withings Sleep Analyzer
4.	Fitbit Versa 3	Fitbit Versa 3	Fitbit Versa 3
5.	Polar M430	Polar M430	Polar M430

Table 8. Ranking devices

The post-equipment phase ended with the expert evaluators ranking all the devices from best to worst for their own user needs. The ranking was done after the second technology expectations discussion. Participants wrote their answers down and simultaneously sent them to the group chat, so that others answers would have no impact on the ranking. Afterwords the rankings were discussed and reasoned in a group.

Ranking of P1

Expert evaluator 1 had a strong trust on Dreem2 from the start, since its the only device out of the five that uses EEG technology. The Dreem2 headband made it to the first place since its data seemed accurate and responded to participants feelings of most nights. The data was versatile and using the device was comfortable.
The data provided by Withings sleep analyser was similar to Dreem2 and most of the stages matched together with only about half an hour difference. The mattress was also easy and comfortable to use so it got second place on the ranking scale.
Sleepscore Max got third place since it was not as competent as the first two. The data had more variance which made it seem more inaccurate.
Fitbit Versa 3 gave a good overview of the sleep stages and sleep quality, but the data sometimes differed a lot from the other devices.
The last place went to Polar M430 because it did not have the sleep stage feature and thus lacked in the competence of sleep monitoring.

Ranking of P2

Participant 2 ranked Withings Sleep Analyser to the first place, since the mattress was comfortable to use and the data best corresponded to users own feelings of the night. The data was seemingly very comparable to Dreem2.
Dreem2 took the second place since the data seemed very reliable. This device did not get the first place because it did not stay on the head very well with long hair and different sleeping positions.

Sleepscore Max was a nice surprise with pretty accurate seeming data. Some of the scoring was not well explained, for example the body score and mind score classification was confusing.
Fitbit Versa 3 did not correspond well with the second expert evaluators feelings of the nights. There was a lot of small awake times for every night.
Polar M430 got the last place since expert evaluator 2 did not find the data very useful. All the nights were similar and had a lot of short awake times just like Fitbit Versa 3 had as well. Also the sleep stages and heart rate were not monitored, which made it have less features than the other devices.

Ranking of P3

Participant 3 chose Dreem2 headband to be the first device in the ranking. The data seemed believable and there were no problems with the device.
Sleepscore Max was almost as good as Dreem2, which led it to the second place. Expert evaluator 3 thought that the data seemed reliable.
Withings sleep analyzer did not get the highest placements in the ranking because some nights it did not record the sleeping time correctly. Besides that the device was mostly accurate.
Fitbit Versa 3 took the fourth place since it gave a good overview of the night and the sleep stages, but it was not very reliable regarding the awake time during the night. It also only monitored the sleep from the wrist wich made it less reliable according to P3.
Polar M430 was like Fitbit Versa 3 but without all the useful and interesting features. It was only a bit better than just the users own memory of the night, which made the device pretty useless on monitoring sleep.

5.2. Quantitative Results and Data Analysis

5.2.1. Sleep Survey

Self-assessment sleep survey was used to track expert evaluator sleeping habits during the pre-equipment and equipment phase, one week each. One of the tracked metrics was how well did the participant sleep in their own opinion. If we take a look at the Figures 2 - 3, we can see their responses plotted on a graph using a Likert scale from 0 to 10, where 0 means "a lot worse than normal" and 10 means "Much better than normal". No clear trend can be noticed from either of the figures, so it is not possible to evaluate if using the equipment had any effect on their sleep in their own opinion.

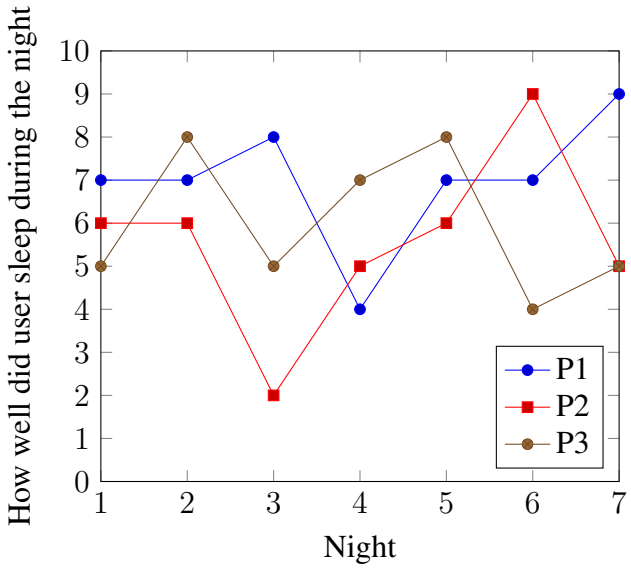


Figure 2. Pre-equipment sleep survey (question D5)

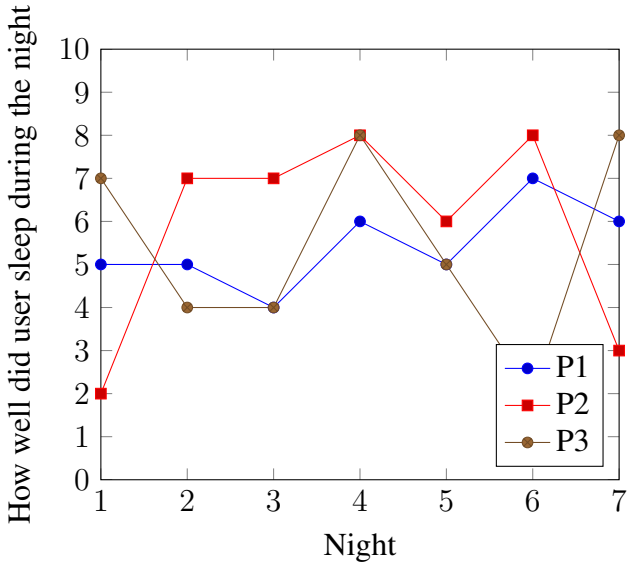


Figure 3. Equipment sleep survey, D5

Equipment

During the equipment phase the self-assessment sleep survey had extra questions about the sleep tracking devices. After every night, each of the three expert evaluators had to choose all the devices that accurately tracked their sleep. That has been charted on Figure 4 where we can see that the three highly ranked devices in the second technology expectations discussion (SleepScore Max, Dreem2, and Withings Sleep Analyzer) have proved to be accurate over 50 percent of the nights with Dreem2 leading the pact with 16/21 accurate nights. In the last place is Polar M430 with three

accurate nights. Both Fitbit Versa 3 and Polar M430 achieved zero percent accuracy with P1.

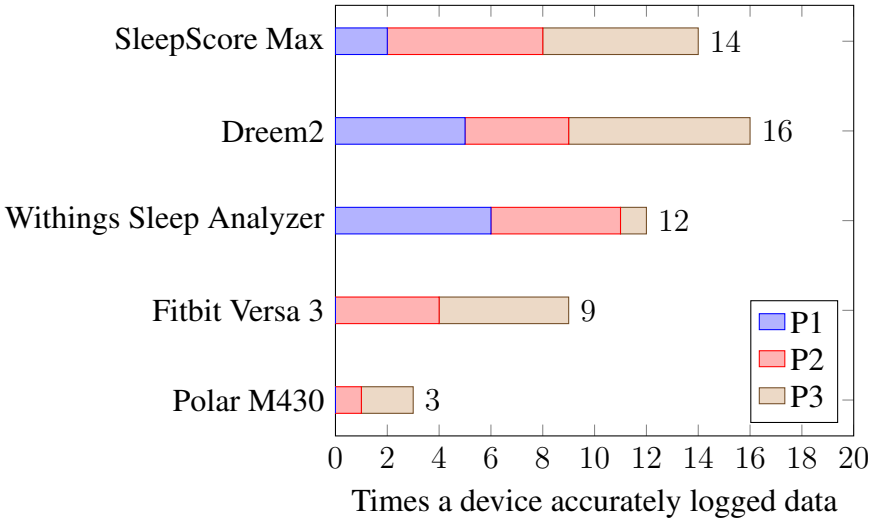


Figure 4. Sleep survey T1

What makes Dreem2 accuracy rating more impressive is when we take a look at Figure 4 which shows how many times our participants encountered a some form of problem, error or other malfunction during last evening, night, or morning. We can see that Dreem2 has the highest number of encountered problems with 9 while the rest of the devices only had between 1 and 3 each.

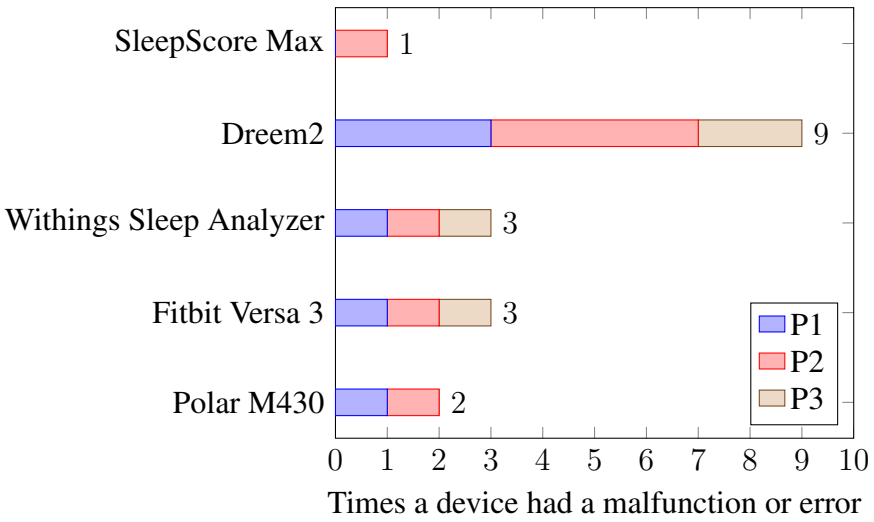


Figure 5. Sleep survey T2

Then when we take a look at Table 9 which contains all the tracked equipment variables in the self-assessment survey and the percentage of how many nights did a device track it incorrectly. In general, all of the 5 devices had problems recognising night-time wake ups or sleep restfulness, with the Dreem2 and SleepScore Max having the lowest score of the 5 with 42,86 percent. Other than that both Dreem2 and SleepScore Max performed well in all the other categories regularly achieving the lowest inaccuracy rate. Polar M430 high score on sleep phases can be explained with

that Polar M430 does not track sleep phases, so our expert evaluators did not have an uniform view on if it should be marked as inaccurate or not.

Device	Bedtime	Wakeup time	Sleep phases	Night-time wakeups or sleep restfulness	Heart-rate & related variables	Sleep latency	Total amount of sleep
Polar M430	33,33	14,29	71,43	52,38	33,33	61,90	28,57
Fitbit Versa 3	38,10	23,81	38,10	47,62	42,86	42,86	42,86
Withings Sleep Analyzer	42,86	19,05	14,29	47,62	9,52	19,05	38,10
Dreem2	9,52	23,81	14,29	42,86	14,29	9,52	28,57
SleepScore Max	14,29	23,81	14,29	42,86	14,29	9,52	28,57

Table 9. Times user reported a sleep-related variable to be inaccurate (%)

5.2.2. Analysis of Data

Sleep times

All 5 devices in our study measured bedtime, wake up time, sleep duration, and time in bed in one way or another. Dreem2 headband, SleepScore Max, and Withings Sleep Analyzer do not report bedtime in their CSV dataset and it had to be calculated from the time user started using the device and the time it took for the user to fall asleep.

Looking at the Table 10 we can see that for P1 all the devices that were ranked high after the second technology expectations discussion (Dreem2 headband, SleepScore Max, and Withings Sleep Analyzer) were close to each other in terms of sleep duration. While the two fitness tracker wrist watches, Polar M430 and Fitbit Versa 3, are tens of minutes away from the aforementioned trio. P2 experienced problems with the Dreem2 headband, mainly issues with the headband not staying properly on their head during sleep, which directly affects its sleep duration average. But if we remove one outlier night from the average, the new average is 07.20.15, which is more in the same line with the other devices. P3 did not any malfunctions with the devices, which could explain why the devices are tightly grouped, the exception being Withings Sleep Analyzer. P3 was the only one to leave it plugged in during the week, unlike the other devices which they turned on only after deciding to sleep. So it is likely that Withings Sleep Analyzer got confused when the bed was used for relaxation, internet surfing with a mobile phone or sleeping.

Average ± Standard Devition

	P1		P2		P3	
Device	**Sleep duration**	**Time in bed**	**Sleep duration**	**Time in bed**	**Sleep duration**	**Time in bed**
Dreem2	06.44±55	07.01±59	06.42±103	07.54±88	06.46±68	07.03±72
Fitbit Versa 3	07.09±46	07.54±44	07.42±45	08.40±50	06.42±79	07.25±90
Polar M430	06.30±26	06.49±28	07.51±30	08.14±31	06.37±67	06.52±71
SleepScore Max	06.54±32	07.23±40	07.33±53	08.24±38	06.23±71	06.54±73
Withings Sleep Analyzer	06.52±27	07.15±30	08.07±31	08.07±31	7.55±55	09.20±56

Table 10. Average time (h.m.s) and standard deviation (min) for sleep duration and time in bed

Participant 1

Taking a closer look at the Tables 11 - 14, which show the bedtime, wake up time, sleep duration, and time in bed for P1 per device per night. The results highlight three malfunctions: Dreem2 night 1, Fitbit Versa 3 night 7, and Withings Sleep Analyzer night 7. Comparing Fitbit Versa 3 and Polar M430 to other devices, we can see some inconsistencies with bedtime and wake up time. This leads to unreliable values for sleep duration, giving confirmation for their low positions on the ranking chart. Other notable outlier is Withings Sleep Analyzer night 5 where bedtime is over 20 minutes later compared to Dreem2 or SleepScore Max.

Bedtime - P1 Sleep times

Night	**Dreem2**	**Fitbit Versa 3**	**Polar M430**	**SleepScore Max**	**Withings Sleep Analyzer**
Night 1	N/A	00.49.00	00.50.00	00.49.00	01.01.00
Night 2	00.25.58	00.23.00	00.25.00	00.25.00	00.28.00
Night 3	02.37.40	01.36.00	02.32.00	02.34.00	02.37.00
Night 4	01.01.39	00.53.00	01.02.00	01.13.00	01.07.00
Night 5	23.44.01	23.46.00	00.30.00	23.49.00	01.13.00
Night 6	00.47.54	00.42.00	01.04.00	01.02.00	01.26.00
Night 7	01.14.28	N/A	01.14.00	01.13.00	N/A

Table 11. Fall asleep time, P1 (h.m.s)

Wakeup time - P1 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	N/A	07.12.00	07.20.00	07.32.00	07.54.00
Night 2	08.23.28	07.44.00	06.08.00	08.00.00	08.13.00
Night 3	07.49.10	08.07.00	08.46.00	08.48.00	09.07.00
Night 4	07.57.09	07.30.00	07.44.00	07.57.00	07.53.00
Night 5	06.13.01	07.25.00	07.37.00	07.31.00	07.55.00
Night 6	07.57.54	09.02.00	07.44.00	07.45.00	08.04.00
Night 7	07.56.28	N/A	07.45.00	07.51.00	N/A

Table 12. Wake up time, P1 (h.m.s)

Sleep duration - P1 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	N/A	06.23.00	06.30.00	06.43.00	06.53.00
Night 2	7:57:30	07.21.00	05.43.00	07.35.00	07.45.00
Night 3	5:11:30	06.31.00	06.14.00	06.14.00	06.30.00
Night 4	6:55:30	06.37.00	06.42.00	06.44.00	06.46.00
Night 5	6:29:00	07.39.00	07.07.00	07.42.00	06.42.00
Night 6	7:10:00	08.20.00	06.40.00	06.43.00	06.38.00
Night 7	6:42:00	N/A	06.31.00	06.38.00	N/A

Table 13. Total sleep duration, P1 (h.m.s)

Time in Bed - P1 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	N/A	07.06.00	07.00.00	07.05.00	07.14.00
Night 2	08.27.15	08.01.00	05.56.00	08.17.00	08.09.00
Night 3	05.24.28	07.37.00	06.40.00	06.43.00	06.45.00
Night 4	07.05.40	07.16.00	06.58.00	07.01.00	07.02.00
Night 5	06.50.47	08.23.00	07.28.00	08.21.00	07.25.00
Night 6	07.22.22	09.03.00	06.58.00	07.24.00	06.52.00
Night 7	06.54.30	N/A	06.44.00	06.52.00	N/A

Table 14. Total time in bed, P1 (h.m.s)

Participant 2

Examining the Tables 15 - 18 gives us the per device per night statistics for P2, and from that we can tell that there were two malfunctions: SleepScore Max night 4 and Withings Sleep Analyzer night 1. Bedtime and wake up time data is very inconsistent across all devices for P2. Few outliers in Dreem2 data, nights 2 and 5, are most likely

explained with the issue on the headband not staying in place. Fitbit Versa 3 and Polar M430 again performed poorly, especially on nights 6 and 7, with inconsistencies with bedtime and wake up time. Fitbit's night 3 should also be noted because its bedtime is two hours earlier than other devices.

Bedtime - P2 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	23.59.06	23.49.00	23.48.00	00.14.00	N/A
Night 2	00.44.11	23.25.00	23.24.00	23.47.00	00.24.00
Night 3	23.41.15	21.30.00	23.46.00	23.46.00	23.57.00
Night 4	23.30.36	23.34.00	23.30.00	N/A	23.58.00
Night 5	23.43.25	23.26.00	23.58.00	00.11.00	23.37.00
Night 6	23.24.07	00.07.00	00.09.00	23.25.00	23.54.00
Night 7	01.27.17	00.13.00	00.17.00	01.17.00	23.31.00

Table 15. Fall asleep time, P2 (h.m.s)

Wakeup time - P2 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	07.19.36	08.03.00	08.02.00	07.25.00	N/A
Night 2	07.31.41	07.40.00	08.10.00	07.29.00	07.49.00
Night 3	07.44.45	06.13.00	07.23.00	07.42.00	08.02.00
Night 4	07.11.06	06.51.00	07.25.00	N/A	08.03.00
Night 5	02.38.25	07.05.00	07.37.00	07.12.00	07.48.00
Night 6	06.43.37	07.16.00	07.38.00	08.25.00	07.48.00
Night 7	08.25.17	06.49.00	07.34.00	07.44.00	08.31.00

Table 16. Wake up time, P2 (h.m.s)

Sleep duration - P2 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	7:20:30	08.14.00	08.14.00	07.11.00	N/A
Night 2	6:47:30	08.15.00	08.46.00	07.42.00	07.25.00
Night 3	8:03:30	08.43.00	07.37.00	07.56.00	08.05.00
Night 4	7:40:30	07.17.00	07.55.00	N/A	08.05.00
Night 5	2:55:00	07.39.00	07.39.00	07.01.00	08.11.00
Night 6	7:19:30	07.09.00	07.29.00	09.00.00	07.54.00
Night 7	6:58:00	06.36.00	07.17.00	06.27.00	09.00.00

Table 17. Total sleep duration, P2 (h.m.s)

Time in Bed - P2 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	08:06:18	08.43.00	08.43.00	08.09.00	N/A
Night 2	08:39:48	09.09.00	09.07.00	08.35.00	07.25.00
Night 3	08:44:03	10.13.00	08.00.00	08.38.00	08.05.00
Night 4	08:21:32	08.12.00	08.15.00	N/A	08.05.00
Night 5	04:49:47	08.37.00	08.02.00	08.30.00	08.11.00
Night 6	09:13:41	08.05.00	07.53.00	09.14.00	07.54.00
Night 7	07:21:34	07.40.00	07.36.00	07.20.00	09.00.00

Table 18. Total time in bed, P2 (h.m.s)

Participant 3

Compared to P1 and P2, P3 has no malfunctions on any of nights for any of the devices, as seen from the Tables 19 - 22. P3 reported that Withings Sleep Analyzer had an issue confusing distinguishing when they were actually trying to sleep and when they were just very still while using a mobile device. The effects of this can be seen in the bedtime and wake up time for Withings', especially on nights 3, 6, and 7. Fitbit Versa 3 and Polar M430 performed well compared to others on bedtime, but wake up times have major inconsistencies to both ways, which explains why their average times are comparable to Dreem2 and SleepScore Max.

Bedtime - P3 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	00.56.47	00.48.00	00.58.00	00.59.00	00.21.00
Night 2	01.20.45	01.14.00	01.17.00	01.27.00	01.23.00
Night 3	00.42.52	00.27.00	00.31.00	00.58.00	23.02.00
Night 4	02.35.26	02.15.00	02.30.00	02.25.00	02.26.00
Night 5	05.45.24	05.31.00	05.41.00	05.50.00	04.29.00
Night 6	04.29.16	04.27.00	04.28.00	04.27.00	04.15.00
Night 7	04.15.52	03.59.00	04.14.00	04.13.00	04.20.00

Table 19. Fall asleep time, P3 (h.m.s)

Wakeup time - P3 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	07.55.04	07.42.00	07.54.00	07.26.00	07.50.00
Night 2	08.01.15	09.19.00	07.53.00	07.37.00	08.05.00
Night 3	07.26.52	06.42.00	07.12.00	07.05.00	06.41.00
Night 4	10.43.56	10.06.00	10.30.00	10.44.00	10.46.00
Night 5	11.36.24	11.12.00	11.29.00	11.20.00	12.10.00
Night 6	09.22.16	08.54.00	09.09.00	09.09.00	13.56.00
Night 7	12.11.22	11.40.00	11.54.00	11.37.00	12.10.00

Table 20. Wake up time, P3 (h.m.s)

Sleep duration - P3 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	7:08:30	06.54.00	06.56.00	06.27.00	07.29.00
Night 2	6:40:30	08.05.00	06.36.00	06.10.00	06.42.00
Night 3	6:44:00	06.15.00	06.41.00	06.07.00	07.39.00
Night 4	8:08:30	07.51.00	08.00.00	08.19.00	08.20.00
Night 5	5:51:00	05.41.00	05.48.00	05.30.00	07.41.00
Night 6	4:53:00	04.27.00	04.41.00	04.42.00	09.41.00
Night 7	7:55:30	07.41.00	07.40.00	07.24.00	07.50.00

Table 21. Total sleep duration, P3 (h.m.s)

Time in Bed - P3 Sleep times

Night	Dreem2	Fitbit Versa 3	Polar M430	SleepScore Max	Withings Sleep Analyzer
Night 1	07:24:10	07.29.00	07.13.00	06.52.00	07.53.00
Night 2	07:01:10	09.20.00	06.51.00	06.59.00	08.22.00
Night 3	07:03:53	07.03.00	06.57.00	06.49.00	09.09.00
Night 4	08:30:25	08.34.00	08.19.00	08.28.00	09.59.00
Night 5	06:03:24	06.08.00	05.57.00	05.55.00	09.35.00
Night 6	05:02:38	04.58.00	04.50.00	05.02.00	10.33.00
Night 7	08:15:59	08.20.00	07.58.00	08.16.00	09.51.00

Table 22. Total time in bed, P3 (h.m.s)

Sleep Stages

In our study we had 4 devices that measured sleep stages these are Dreem2 headband, Fitbit Versa 3, SleepScore Max and Withings sleep Analyzer. Polar M430 unfortunately did not measure sleep stages and therefore does not have any data in this section.

Average ± Standard Deviation

P1				
Device	**Light**	**Deep**	**REM**	**Awake**
Dreem2	3.05.50±20	1.20.15±19	2.22.10±15	0.10.30±4
Fitbit Versa 3	3.50.20±35	1.14.50±11	2.03.20±1	0.45.50±9
SleepScore Max	3.39.51±30	1.32.43±20	1.40.51±20	0.29.09±11
Withings Sleep Analyzer	3.13.40±21	1.25.20±13	2.13.20±17	0.22.10±10
Average of Devices	3.27.25±18	1.23.17±7	2.04.55±15	0.26.55±13
P2				
Device	**Light**	**Deep**	**REM**	**Awake**
Dreem2	3.03.21±52	1.07.51±7	2.36.34±56	0.34.43±21
Fitbit Versa 3	5.06.20±26	0.56.20±17	1.33.50±20	0.58.00±17
SleepScore Max	4.39.50±51	1.02.10±10	1.50.10±19	0.23.20±13
Withings Sleep Analyzer	3.18.30±33	2.24.30±34	2.23.40±23	N/A
Average of Devices	4.02.00±52	1.22.43±36	2.06.04±25	0.38.41±14
P3				
Device	**Light**	**Deep**	**REM**	**Awake**
Dreem2	3.07.47±24	1.20.43±20	2.22.21±51	0.11.51±4
Fitbit Versa 3	4.24.43±48	0.57.51±19	1.19.26±20	0.42.34±15
SleepScore Max	3.14.17±41	1.50.51±25	1.16.26±29	0.31.43±15
Withings Sleep Analyzer	3.07.34±66	2.43.00±31	2.04.00±12	1.35.00±35
Average of Devices	3.28.35±33	1.43.06±39	1.45.33±28	0.45.17±31

Table 23. Average time (h.m.s) and standard deviation (min) in sleep stages

Dreem2 - P1 Sleep Stages

Night	Light	Deep	REM	Awake
Night 1	N/A	N/A	N/A	N/A
Night 2	3.23.00	1.56.30	2.40.30	0.17.00
Night 3	2.23.30	0.56.00	1.57.00	0.08.00
Night 4	3.13.30	1.10.00	2.34.30	0.07.00
Night 5	3.08.00	1.12.30	2.13.30	0.16.00
Night 6	3.24.30	1.22.00	2.29.30	0.06.00
Night 7	3.02.30	1.24.30	2.18.00	0.09.00

Table 24. Dreem2 Sleep Stages, P1 (h.m.s)

Dreem2 - P2 Sleep Stages

Night	Light	Deep	REM	Awake
Night 1	2.44.30	1.08.00	3.31.30	0.44.00
Night 2	2.52.00	1.00.30	3.05.00	1.17.00
Night 3	3.27.00	1.08.00	3.30.30	0.38.00
Night 4	4.21.00	1.18.00	2.05.00	0.16.00
Night 5	1.18.30	1.05.00	0.37.00	0.36.00
Night 6	3.31.00	1.16.30	2.34.30	0.14.00
Night 7	3.09.30	0.59.00	2.52.30	0.18.00

Table 25. Dreem2 Sleep Stages, P2 (h.m.s)

Dreem2 - P3 Sleep Stages

Night	Light	Deep	REM	Awake
Night 1	3.11.30	1.46.30	2.10.30	0.16.00
Night 2	3.28.00	1.15.00	2.03.00	0.15.00
Night 3	3.01.30	1.44.30	2.01.00	0.16.00
Night 4	3.40.30	1.02.30	3.36.30	0.10.00
Night 5	2.19.00	1.13.30	2.19.30	0.11.00
Night 6	3.13.00	0.48.00	0.56.00	0.05.00
Night 7	3.01.00	1.35.00	3.30.00	0.10.00

Table 26. Dreem2 Sleep Stages, P3 (h.m.s)

Withings Sleep Analyzer - P1 Sleep Stages

Device	Light	Deep	REM	Awake
Night 1	3.36.00	1.08.00	2.09.00	0.21.00
Night 2	3.30.00	1.33.00	2.42.00	0.24.00
Night 3	2.58.00	1.11.00	2.21.00	0.15.00
Night 4	3.29.00	1.21.00	1.56.00	0.16.00
Night 5	3.12.00	1.37.00	1.53.00	0.43.00
Night 6	2.37.00	1.42.00	2.19.00	0.14.00
Night 7	N/A	N/A	N/A	N/A

Table 27. Withings Sleep Analyzer Sleep Stages, P1 (h.m.s)

Withings Sleep Analyzer - P2 Sleep Stages

Device	Light	Deep	REM	Awake
Night 1	N/A	N/A	N/A	N/A
Night 2	2.59.00	2.35.00	1.51.00	N/A
Night 3	2.19.00	3.27.00	2.19.00	N/A
Night 4	3.31.00	1.49.00	2.45.00	N/A
Night 5	3.46.00	2.28.00	1.57.00	N/A
Night 6	3.18.00	1.43.00	2.53.00	N/A
Night 7	3.58.00	2.25.00	2.37.00	N/A

Table 28. Withings Sleep Analyzer Sleep Stages, P2 (h.m.s)

Withings Sleep Analyzer - P3 Sleep Stages

Device	Light	Deep	REM	Awake
Night 1	2.40.00	2.45.00	2.04.00	0.33.00
Night 2	1.58.00	2.59.00	1.45.00	1.54.00
Night 3	2.33.00	3.04.00	2.02.00	1.47.00
Night 4	3.16.00	3.09.00	1.55.00	1.50.00
Night 5	3.13.00	2.23.00	2.05.00	1.54.00
Night 6	5.37.00	1.35.00	2.29.00	0.52.00
Night 7	2.36.00	3.06.00	2.08.00	2.15.00

Table 29. Withings Sleep Analyzer Sleep Stages, P3 (h.m.s)

SleepScore Max - P1 Sleep Stages

Night	Light	Deep	REM	Awake
Night 1	3.26.00	1.52.00	1.25.00	0.22.00
Night 2	4.49.00	0.57.00	1.48.00	0.42.00
Night 3	3.13.00	1.15.00	1.45.00	0.29.00
Night 4	3.31.00	1.48.00	1.25.00	0.17.00
Night 5	3.45.00	1.32.00	2.24.00	0.39.00
Night 6	3.25.00	1.56.00	1.21.00	0.41.00
Night 7	3.30.00	1.29.00	1.38.00	0.14.00

Table 30. SleepScore Max Sleep Stages, P1 (h.m.s)

SleepScore Max - P2 Sleep Stages

Night	Light	Deep	REM	Awake
Night 1	4.31.00	1.14.00	1.25.00	0.24.00
Night 2	4.05.00	1.15.00	2.21.00	0.30.00
Night 3	5.00.00	0.51.00	2.04.00	0.11.00
Night 4	N/A	N/A	N/A	N/A
Night 5	4.02.00	1.01.00	1.58.00	0.29.00
Night 6	6.23.00	1.03.00	1.33.00	0.04.00
Night 7	3.58.00	0.49.00	1.40.00	0.42.00

Table 31. SleepScore Max Sleep Stages, P2 (h.m.s)

SleepScore Max - P3 Sleep Stages

Nights	Light	Deep	REM	Awake
Night 1	3.16.00	2.15.00	0.55.00	0.25.00
Night 2	3.07.00	1.18.00	1.44.00	0.49.00
Night 3	2.54.00	1.28.00	1.44.00	0.42.00
Night 4	4.21.00	2.11.00	1.46.00	0.09.00
Night 5	2.14.00	2.24.00	0.51.00	0.25.00
Night 6	2.46.00	1.26.00	0.30.00	0.20.00
Night 7	4.02.00	1.54.00	1.25.00	0.52.00

Table 32. SleepScore Max Sleep Stages, P3 (h.m.s)

Fitbit Versa 3 - P1 Sleep Stages

Device	Light	Deep	REM	Awake
Night 1	3.33.00	1.10.00	1.40.00	0.43.00
Night 2	3.30.00	1.28.00	2.23.00	0.40.00
Night 3	3.07.00	1.21.00	2.03.00	1.06.00
Night 4	3.53.00	0.57.00	1.47.00	0.39.00
Night 5	4.00.00	1.07.00	2.32.00	0.44.00
Night 6	4.59.00	1.26.00	1.55.00	0.43.00
Night 7	N/A	N/A	N/A	N/A

Table 33. Fitbit Versa 3 Sleep Stages, P1 (h.m.s)

Fitbit Versa 3 - P2 Sleep Stages

Device	Light	Deep	REM	Awake
Night 1	4.22.00	1.06.00	1.08.00	1.04.00
Night 2	5.08.00	0.46.00	1.15.00	0.56.00
Night 3	5.23.00	0.28.00	1.48.00	0.58.00
Night 4	5.02.00	0.54.00	1.21.00	0.55.00
Night 5	5.47.00	1.00.00	1.56.00	1.30.00
Night 6	4.56.00	1.24.00	1.55.00	0.54.00
Night 7	N/A	N/A	N/A	0.29.00

Table 34. Fitbit Versa 3 Sleep Stages, P2 (h.m.s)

Fitbit Versa 3 - P3 Sleep Stages

Device	Light	Deep	REM	Awake
Night 1	4.17.00	0.53.00	1.05.00	0.48.00
Night 2	5.11.00	1.12.00	1.28.00	0.43.00
Night 3	4.03.00	0.22.00	1.16.00	0.27.00
Night 4	2.55.00	0.43.00	0.49.00	0.31.00
Night 5	4.46.00	1.07.00	1.48.00	0.39.00
Night 6	4.05.00	1.06.00	1.43.00	0.35.00
Night 7	5.36.00	1.22.00	1.07.00	1.15.00

Table 35. Fitbit Versa 3 Sleep Stages, P3 (h.m.s)

Average Sleep Stages and Standard Deviation of the Week

On one weeks average sleep stages (Table 23) Fitbit Versa 3 measured the most light sleep for all participants and Dreem2 measured the least light sleep for P1 and P2 and Withings Sleep Analyzer for P3.

Fitbit Versa 3 measured the least deep sleep for all participants and Withings Sleep Analyzer the most deep sleep for P2 and P3, SleepScore Max measured the most deep sleep for P1. Devices measured deep sleep quite steadily for P1, the biggest difference between devices average being only about 18 minutes. P2 and P3 had more variance in deep sleep results biggest difference being about 1,5 hours for both.

Dreem2 measured the most REM sleep for all participants and Fitbit Versa 3 measured the least REM sleep for P1 and P2. For P3 the least REM sleep was measured by SleepScore Max which had only 2,5 minutes less REM sleep than Fitbit.

Withings Sleep Analyzer had most (about 26 minutes) and Fitbit Versa 3 had least (about 21 minutes) standard deviation (SD) in average of all participants and sleep stages. SleepScore Max had about 23 minutes SD and Dreem2 had about 24 minutes SD.

As presented in Table 23, most SD was in Light sleep where Almost all devices had near 30 minutes or more SD, especially for P2 and P3. Withings Sleep Analyzer had especially much SD for P3, Dreem2 had 52 minutes SD in light and 56 in REM sleep and SleepScore Max had 51 minutes SD in Light sleep for P2. Fitbit Versa 3 had most SD for P1 in light sleep but very little SD in REM sleep. SleepScore Max also had large SD especially in light sleep.

Heart rate

Dreem 2

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
1	56	62	59	57	61	62	59
2	68		75	73		72	78
3	45	46	46	46	47	59	47

Table 36. Dreem2 - Average heart rates (bpm)

Withings Sleep Analyser

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
1	58	61	59	57		61	57
2		68	69	73	72	71	70
3	45	46	47	46	46	59	47

Table 37. Withings Sleep Analyser - Average heart rates (bpm)

Sleep tracking devices often monitor heart rate, since it can tell a lot of the users sleep quality. Sleeping and relaxing in general typically drops the heart beat to about 50-60 beats per minute. Low heart rate is a sign of recovering after the days activities, and it mirrors to the health of heart and circulation system. The heart rate can also rise due to for example exercising or stress. [18] There are also other factors that affect the average heart rate. Women have usually higher heart rate due to heart size. After puberty, men have about 15-30% greater absolute heart mass than woman. [19]

In our study 4/5 devices use heart rate value as one of the parameters to monitor the users sleep quality and the sleep stages. These devices are Fitbit Versa 3, Polar M430, Dreem2 and Withings Sleep Analyser. Fitbit Versa 3 is not included in the heart rate value analysis, since the data is not free to download. Also Polar M430 does not display the heart rate values to the user after the night.

Dreem2 did not record the average heart rate for P2 for two nights and both P1 and P2 did not get the average heart rate data on one night out of the seven recorded with Withings Sleep Analyser. These nights the devices had a malfunction and did not record any data.

The average heart rate values for the whole week were similar with both of the devices. Dreem2 gave P1 59 bpm, P2 73 bpm and P3 48 bpm, whereas Withings Sleep Analyser gave P1 59 bpm, P2 71 bpm and P3 48 bpm. P2 got 2 bpm higher result with Dreem2 but other participants got the exact same results on the whole week average heart rate. P2 had also two malfunctions and therefore two days data loss which

affects the results.
In conclusion the data seems reliable looking at P1 and P3 results, and overall average heart rates of the whole week, but P2 had more variation between the devices. For example on third night, heart rate was 75 bpm with Dreem2 and 69 bpm with Withings Sleep Analyser. This makes the value range 6 bpm which raises concerns of which device was inaccurate. These chances in data might be because of the headband coming off and being loose on the participants head, since this information came up in the technology expectations realisation discussion. This would explain why the data had more variance than the other participants data.
P1 had a malfunction with Dreem2 on the first night, since the battery died around 3 am in the morning. This could affect to the heart rate being 56 which is the lowest recorded heart rate value on the participants half with Dreem2. Withings Sleep Analyser recorded the nights average bpm to be 58, which seems more reliable here. The overall values are so similar that the one nights error with device using did not make the whole weeks average heart rate value to differ from Withings Sleep Analysers data.
P3 answered in the Daily sleep tracking surveys question: "In your own opinion, how well rested did you feel after waking up today?" the personal highest answer 7 after the first and the last night. On both of those nights the heart rate was below the participants whole weeks average heart rate 48, since on the first night the average bpm was 45 and the last night it was 47 according to both devices. Lower average heart rate could indicate to better rested night [18].

Position Changes

Dreem 2

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
P1		24	21	31	40	30	21
P2	36	42	50	35	15	33	41
P3	53	46	39	57	27	30	53

Table 38. Position Changes, Dreem2

Different sleeping positions may help reducing certain health problems. Some people may struggle to find the best sleep position for them due to ailments, for example back pain or age-related issues. It has been shown in the results of the Great British Sleep Survey that the older we get the more likely we are to have our sleep being disrupted by physical discomfort. Although changing one's sleeping position is unlikely to resolve sleeping problems all at once, changing the sleep positions multiple times at night may leave the sleeper more or less frustrated. [20]
Position changes during sleep have received little attention compared to other aspects of sleep study. It has been studied that during one night, subjects change their positions an average of 20-40 times and stay in the same position for about 15 min. In the same

study one of the main observations was that each subject seemed to have their own repertoire of positions which repeated quite constantly during the night. [21]
In our study the only device that records sleep positions is Dreem2 and the devices web page did not give the information how the sleep positions are recorded. In the CSV-data we got results of the number of sleep position changes per night, but in the application there is also displayed the sleeping positions in three ways: sleeping on back, sleeping on left side and sleeping on right side. Sleeping on stomach is not recorded separately from the sleeping on back data.
P1 had one night of missing data, since Dreem2 did not record on the first night due to battery loss which was noticed in the participants sleep diary. The average sleep position changes during the whole week were: P1 28, P2 36, and P3 44, which are mostly in the range of the average 20-40 position changes per night. Lowest amount of sleep position changes recorded was participant 2 night 5, when there was only 15 position changes counted. In the Daily Sleep Tracking Survey P2 said that in that night Dreem2 had some form of problem, error or other malfunction.
Even though positions changes may not affect the sleep hugely, they can cause more errors in all of the resulted data, since the device moving may affect on the monitoring system. This can happen when the device is not getting the best contact to the subject.

Respiration

Dreem 2

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
P1		15	15	15	14	15	14
P2	16	16	16	16	17	16	17
P3	15	15	15	15	16	16	16

Table 39. Respiration (breaths per minute), Dreem2

Withings Sleep Analyzer

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
P1	14	14	14	14	14	14	13
P2							
P3	15	15	15	15	16	16	15

Table 40. Respiration (breaths per minute), Withings Sleep Analyzer

Respiration rate is the number of breaths taken during one minute. Typically adults have 12 to 16 breaths per minute. Minor respiratory changes can lead to sleep apnea

or other sleep disorders. [22]
In our study Dreem2 and Withings Sleep Analyser are the only devices detecting breathing irregularities giving an average respiratory rate for every night. In Dreem2 support web page, there is a response on how the respiration rate is being collected. Dreem2 uses real-time audio recording to detect the frequency of a user's breathing irregularities. When irregular breathing is detected, the built-in microphone activates to collect the audio. However, due to confidentiality, the audio analysis is carried out in real-time in the headband and its never stored in any server. The Dreem2 support also adds that sleeping with a partner, the headband also pics up the breathing irregularities of the person you are sleeping with. [23] Withings Sleep Analyzer has an algorithm that analyzes the interruptions in breathing patterns during the night. It monitors the intensity of breathing through the mattress. [24]
All the other nights respiration rates were recorded with Dreem2 except for P1 first night. At this night the battery died during the night because it was not completely loaded. Dreem2 support gave the answer on why the battery should be fully loaded at every night. When enabling the feature of irregular breathing detection, the headband only has a 9-hour battery life per night [23]. The other participants did not have the same problems, because they were informed about this by P1. Participant 2 did not get the CSV-data out of Withings Sleep Analyzer due to an account malfunction.
The average whole week respiration rates were P1 15, P2 16 and P3 15 with Dreem2 and P1 14 and P3 15 with Withings Sleep Analyzer. All of these values are similar to the adults average value 12 to 16 breaths per minute. P1 and P2 slept all of the nights next to someone and P3 slept 2/7 nights next to someone. This is why our respiration rate data is probably affected by other people's breathing too.

Sleep efficiency

It is typical for sleep tracking devices to offer a value of the slept night describing how effective it was. This value is uniquely calculated by each device and usually goes from 1 to 100, worst sleep efficiency to best sleep efficiency. Devices have their own algorithms to use defining the sleep efficiency with different parameters such as personal need of sleep, time spend in bed, time spend asleep and sleep stage sufficiency.

Dreem 2

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
P1		96	96	98	95	97	97
P2	90	78	92	92	60	79	95
P3	96	95	95	96	97	97	96

Table 41. Sleep Efficiency (1-100), Dreem2

In Dreem2 the sleep efficiency is counted by dividing the time spent asleep by the time between the moment user wishes to sleep and the end of the record resulting the

time spent in bed asleep. The time user wishes to sleep is referred to the time when user stays still with their eyes closed and not moving. The higher user spends time in bed sleeping the higher the sleep efficiency is. [25]

Sleep efficiency	Score attributed	Comment
+ 90%	5	Excellent
85-90%	4	Very good
80-85%	3	Good
75-80%	2	Average
- 75%	1	Poor

Table 42. Dreem2 Sleep Efficiency scoring [25]

The average sleep efficiencies for the whole week were P1 97, P2 84, and P3 96. Participant 2 has sleep efficiency 60 on the night 5 which is significantly lower compered to other sleep efficiencies. At technology expectations discussion realisation P2 mentioned that the headband would come off the head multiple times and that would cause waking up during the night. At that night P2 also had a malfunctions mentioned with Dreem2 in the Daily sleep tracking survey, which could lead for the sleep efficiency to be lower scored. Otherwise the sleep efficiencies are evenly distributed.

Fitbit Versa 3

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
P1		81	75	81	80	84	
P2		78	79	81	82	89	
P3		81	78	84	77	62	85

Table 43. Sleep Score (1-100), Fitbit Versa 3

Fitbit Versa 3 provides sleep score to help the user see the trends of the personal sleep patterns. The overall nightly sleep score is based on the users heart rate, the time spend awake or restless and the sleep stages. It is calculated by summing up the users individual sleep duration scores, sleep quality and restoration resulting a value between 1 and 100. [26]

Excellent	90-100
Good	80-89
Fair	60-79
Poor	<60

Table 44. Fitbit Versa 3 Sleep score [27]

Fitbit Versa 3 does not provide sleep score of the first night, since it needs to have some personal information of the sleep tracked before defining the sleep score. The average sleep scores of the remaining six nights were P1 80, P2 82, P3 78. P1 and P2 both had also a data loss on the last night, which might affect the results. P3 had the lowest sleep score on night 6 and wrote to the sleep diary of having many servings of alcohol before the night, which is probably the reason for lower score. Overall the sleep scores were quite similar and did not vary much.

SleepScore Max

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
P1	90,84,89	97,99,90	97,99,90	91,81,94	88,96,85	93,91,77	91,81,95
P2	92,90,84	95,95,85	93,97,74		87,96,79	91,92,80	84,90,73
P3	87,87,91	63,49,82	75,71,96	95,92,94	79,83,83	78,88,80	82,76,94

Table 45. SleepScore, MindScore and BodyScore (1-100), SleepScore Max

SleepScore Max was the only device to calculate MindScore and BodyScore separately from the SleepScore. SleepScore is calculated using total sleep duration, time taken to fall asleep, time in different sleep stages and the number and duration of awakenings during the night. After tracking these variables are measured against known sleep averages for the users age and gender and then put into an algorithm to create users specific nightly Sleep score rated from 0-100. [28]
SleepScore Max uses sleep stages to define MindScore and BodyScore. MindScore is mainly based on how much REM sleep the user gets each night since REM sleep is the most important sleep stage regarding to the memory storage and brain regeneration. It is made to show the user how mentally restorative the slept night was. BodyScore uses mostly the time spent in deep sleep to tell the user how physically restorative the slept night was. [28]
The whole week average SleepScore, MindScore and BodyScores were P1: 92, 90, 89, P2: 90, 93, 79, and P3: 79, 78, 89. The only data loss happened with participant 2 on the night 4. MindScores and BodyScores have a bit more variance which is based on different amount of the participants sleep stages. P1 has almost the same values on personal average MindScore and BodyScore but P2 has way better personal average MindScore than BodyScore and P3 has way better BodyScore than MindScore. This is probably because P2 sleeps more REM sleep and P3 more deep sleep compered to the whole nights sleep stages.

Withings Sleep Analyzer

PX	night 1	night 2	night 3	night 4	night 5	night 6	night 7
P1	87	95	81	86		81	85
P2	95	96	97	97	98	100	
P3	90	74	78	87	77	44	83

Table 46. Sleep Score, (1-100), Withings Sleep Analyzer

Withings Sleep Analyzer has created their Sleep score rating to provide a simple and intuitive way to display the user how well the night was slept. For every night's sleep, the user gets a score out of 100 points based on 6 key points: duration, depth, regularity, interruptions, time to sleep and time to get up. [27]
The whole week average Sleep scores are P1 86, P2 97, P3 76. P1 and P2 had both one night of data loss. P3 had the most variance in Sleep scores and P2 had the most similar values. P2 had better Sleep score each night, which could mean that the the user got used to the devices during the sleep tracking week and could sleep better with them.

Polar M430

Polar M430 does not provide a Sleep score of the night but it does give a short description of how effective the night was.
For P1 Polar gave 5/7 nights the same message: *"Shorter sleep than your goal, a lot of wakefulness"*, on the third night: *"Shorter sleep than your goal, poor quality"*, and on the seventh night: *"Shorter sleep than your goal, good overall quality"*.
P2 had the same message after 6/7 nights: "Enough sleep, a lot of wakefulness", but on the first night the message was: *"Shorter sleep than your goal, a lot of wakefulness"*.
P3 had most variance in the Polar sleep feedback section. The first nights message was: *"Shorter sleep than your goal, a lot of wakefulness"*, fourth and seventh nights message was: *"Enough sleep, good quality"*, and the rest of the nights message was: *"Shorter sleep than your goal, good overall quality"*.
The data provided by Polar M430 can be interesting to the user in the morning, but since it is not numeric data, it is not very comparable to the other devices data.

6. DISCUSSION

6.1. Key Findings

6.1.1. Reliability of Sleep Monitoring

In this section there will be discussed some possible reliability issues found during the study. Errors such as data loss or device not staying on are not included here, since there is a clear connection between the results and what affected them.
P3 had written in the sleep diary of having many servings of alcohol before night 6, which showed in Fitbit Versa 3 and in Withings Sleep Analyzer of significantly lower Sleep score. Other devices seemed to not have had an impact of this in the scoring although consuming alcohol is known to have a negative impact on sleep. This could due to inadequate parameters used in sleep efficiency monitoring, since Withings Sleep Analyzer and Fitbit Versa 3 had multiple parameters monitoring sleep efficiency, but for example Dreem2 used only time asleep in the calculating process [28]. This would not explain why SleepScore Max did not display the alcohol usage in any significant way in sleep scoring, since it uses multiple parameters to calculate the sleep including sleep stages [31].
All the participants ranked the device that gave them the best sleep efficiency scores to be the best device. There could be a link to people wanting to know that they sleep well and having a more positive feeling on the device that gives them positive feedback. The rankings were also reasoned with the device's comfortability and other factors mentioned on the technology expectations discussions, so the sleep scores are just one of the factors affecting the ranking. However Sleep score is an important factor in the device ranking, since it is the most reasoned data that the device gives to the user. Fitbit Versa 3 gave overall lower Sleep scores than the other devices and all the expert evaluators were not impressed with the devices sleep tracking in the technology expectation results discussion. The lower sleep scores could be due to high amounts of night-time wake ups and sleep restfulness tracked by the device.

6.1.2. Sleep Stages

Devices had some clear differences in sleep stage measurements which either might be correct for some of the devices or conversely measured too much or too little amount of some individual sleep stage. Dreem2 measured clearly the most REM sleep, which is probably the most accurate result from our devices, and it is the most accurate sleep stage measured by Dreem2 [10].
Fitbit Versa 3 measured most light sleep and least deep sleep by average for all participants, especially big difference between light and deep sleep was for P2 and P3 who have said that they do not use watches on regular basis.
SleepScore Max is wireless sleep tracking device and might thereby be more vulnerable to outside factors, such as sleeping with partner, albeit SleepScore Labs informs in their website that *"Two sleepers in the same bed have been shown to not affect accuracy but note that only one partner may track using SleepScore at a time."*

[29], but no references to back this up were given. This gave reason to doubt that multiple persons in bed could possibly have some effect in sleep measurements. P1 and P2 slept every night with partner, for P1 results were mostly consistent and P2 had some variance due to probable difference in sleeping habits, but since there were no nights that P1 or P2 slept alone, so their nights are incomparable in this scenario. Reasoning for doubts is found in P3's measurements, where Light sleep is clearly highest during the nights 4 and 7 (32), when P3 reported to have slept with partner, other devices did not have similar differences on these nights, except for Dreem2 which measured most REM sleep during these nights. Other probable reason for this could be that P3's sleep cycle was affected by other person on these nights.
Alcohol usage shortens REM sleep, elongates non-REM sleep and causes more sleep disruptions. [30] This is visible on P3's night 6, when P3 had multiple servings of alcohol and also on P2's night 7 when P2 had only few servings of alcohol. During P3's night 6, Dreem2 (26) measured least REM and deep sleep, SleepScore Max (32) measured least REM sleep of all nights and Withings Sleep Analyzer measured a lot more light sleep on night 6 than any other night for P3 (29). Fitbit Versa 3, it did not measure big differences in sleep stages on night 6 for P3 (35). During P2's night 7 Dreem2 (25) and SleepScore Max (31) measured least deep of all nights and Withings Sleep Analyzer measured most light sleep (28). These results imply that Dreem2, Withings Sleep Analyzer and SleepScore Max measured the expected effects of alcohol consumption somewhat correctly.

6.1.3. Devices' Error Sensitivity

We had two devices, Dreem2 and SleepScore Max, that required the user to manually start the sleep tracking, all the other devices started tracking automatically. In the case of Withings Sleep Analyzer, P1 and P2 unplugged the device after every night and therefore needed to plug it back in, for it to start tracking but this is a personal choice so we still count Withings' sleep tracking to be automatic. Both Dreem2 and SleepScore Max recorded the time they were started, so we have reliable timestamp on when the user was still awake and decided to try to fall asleep.

Comparing other devices bedtimes against the Dreem2's start up time, both Fitbit Versa 3 and Polar M430 reported on multiple nights that the user had fallen asleep before they started the Dreem2's sleep tracking. As none of our expert evaluators reported falling asleep before setting up all the devices, we can conclude that both Fitbit and Polar have incorrectly measured user's bedtime. This was also noticed by our participants, as both devices were reported to give inaccurate bedtimes in one third of the nights.

6.1.4. Technology Expectations and Realisations

During the technology expectation and realisation discussions a few unexpected points emerged, but most of our expectations were altogether correct.
We expected wearable devices (i.e. Dreem2, Fitbit and Polar) to have some problems with battery life, which was unexpectedly short for Dreem2, lasting only 9 hours on full charge when breathing irregulation detection was enabled. Dreem2 also unexpectedly came off the head of P2 during sleep. Nevertheless we expected Dreem2 to be most functional for sleep measurements and its data to be the most useful.
Polar M430 was expected to be better of the two watches because it uses trademarked Polar Sleep Plus function which was mentioned in devices user manual [31], which left impression to our expert evaluators, that it should be able to measure sleep stages, but unexpectedly it did not measure them at all, only sleep time and wake time. This was mentioned in the Polar Sleep Plus white paper's limitations [8], nevertheless in the same paragraph it is mentioned that they released new function Polar Sleep Plus Stages that is able to measure sleep stages, however this is not available for model M430 which may lead to misunderstandings.
SleepScore Max measured more reliable data and was more functional, closer to Dreem2 and Withings Sleep Analyzer, than previously expected.
Both SleepScore Max and Withings Sleep Analyzer were expected to be most comfortable since they are not attached to user, which was confirmed after using devices. These might be best devices to measure sleep unobstructively in everyday use.

6.1.5. Comparing Device Features

Devices included in this study had a good variance of features, some more useful than others. The study was also affected by the devices different sleep efficiency monitoring algorithms, since it made the scoring not totally comparable with each other.
The promised Polar Sleep Plus feature, was not as competitive than the other devices sleep features, since it did not provide data of the sleep stages. Dreem2 had special features such as white noise and it also counted and displayed the sleep position changes to the user. SleepScore Max measured the temperature and light level of the room. These features were nice addition but that interesting or useful overall.
Since sleep efficiency is calculated using each device's own methods and algorithms, the data cannot be compared with each other invariably. The comparing is also impacted by SleepScore Max being the only device that provides MindScores and BodyScores and Polar M430 not providing a numeric scoring of the night. Expert evaluators did not find SleepScore Max's and Polar M430's special sleep scoring methods as useful as the overall sleep score as numeric data.
Different sleep efficiency counting methods made some scoring seem more reliable than others. For example Dreem2 counting the sleep efficiency by only the time spent asleep compared to the time spend awake trying to sleep [28], does not seem as particular as with Fitbit Versa 3 where the sleep score is calculated with heart rate, the time spend awake or restless and the sleep stages [29]. This does not mean that the

scoring of Dreem2 would be automatically less reliable, since the monitoring methods are different, one measuring on the head and one on the wrist.

6.2. Reflection on Related Work

Sleep quality can be measured from many different variables which in turn makes it complicated. In our study the devices we tested used different technologies which caused them to measure different amounts of total sleep time and sleep stages. Our findings are inline with a different study, where they compared sleep-tracking devices against medical device and found conflicts in total sleep time, sleep stage ratios, and number of awakenings during the night[32].

We also pondered on the possibility that user might rank a device more highly if the device gives them a more positive sleep quality score. Looking at the related work where they found out that people favoured the device that gave results more inline with their expectations and feelings over a possibly more accurate medical device[32].

Our results show that in terms of bedtime, wake up time, total amount of sleep, Polar M430 was deemed to be more accurate than Fitbit Versa 3, but still all participants ranked Fitbit more highly than Polar, which indicates that user's trust in the device is not solely based on accuracy. In our study biggest reason for Polar's bad rank was the lack of features, mainly no sleep stage monitoring. Study have shown that user might trust a device less, because they do not trust the way it measures sleep, even if it provides similar results, as was the case when Fitbit Charge 2 was tested against a medical device[32].

6.3. Future Work

There is future work to be done with comparing different sleep tracking technologies. We propose a two different ways of extending this study.

The first is to create a better sleep diary for participants. Our study gave free rein to the expert evaluators to write down things they thought might have affected their sleep. This meant they might have forgotten to mention something they thought to be unimportant. An improved sleep diary could contain questions about the amount of coffee or alcohol consumed during the day, amount and type of exercise done during the day, and did the user use a mobile phone in bed before going to sleep. It should also contain a better method for self-tracking sleep times to better be able to determine most accurate device.

The second is to study the effects of following a device's suggestions on bettering one's sleep. Some devices give tips and suggestions on how to make the user sleep better. We propose a study where first, a week is spent using the devices normally like in our study and then a second week is spent following the suggestions the devices give out. Results could then be compared to see if the suggestion cause a statistically meaningful difference.

6.4. Limitations

Main limitations for our study is small sample size consisting only 3 expert evaluators and quite short measurement period of 7 days for each participant, which leads to small amount data, which does not show real trends for devices measuring capabilities, especially since some of the devices were used during the device's assessment period where all features were not available.

7. SUMMARY

Sleep monitoring devices have generalised in everyday use due to rapid digitalisation. In this thesis three expert evaluators collected and analysed the data of five different sleep monitoring equipment. The study included two wrist watches Fitbit Versa 3 and Polar M430, one EEG measuring headband Dreem2, one mattress Withings Sleep Anlayzer and non-wearable device SleepScore Max placed on the user's bedside table. The study was conducted in three phases: pre-equipment, equipment, and post-equipment. First phase focusing on participants sleeping habits before using the devices and expectations on the devices, second phase focusing on using the devices and tracking the experience down after every night and third phase focusing on the aftermath of using the devices concluding discussions on the devices and the analysis of the collected data. Before and after using the devices, expert evaluators went through their thoughts on the devices with technology expectations discussions in categories ease-of-use, functionality, reliability, usefulness, comfortability, and sleep disturbance. The devices were ranked by each participant individually, raising Dreem2 and Withings Sleep Analyzer to the top rankings and leaving Polar M430 and Fitbit Versa 3 to lower rankings.

The data analysis brought up five major key points regarding to the devices and the study. The reliability of sleep monitoring was questioned when only 2/5 devices displayed a significantly lower sleep efficiency rating after a night where participant had many servings of alcohol. The changes between devices sleep scoring are probably due to different algorithms the devices use to define the sleep efficiency. While analysing sleep stages Dreem2 turned out to measure clearly the most REM sleep out of all the devices and Fitbit Versa 3 most light sleep and least deep sleep. Dreem2 uses EEG measuring, which is why expert evaluators think that Dreem2 might be the most accurate on monitoring REM sleep. Fitbit Versa 3 displayed much more awakenings during the night compared to other devices, which could lead to false light sleep tracking. SleepScore Max raised concerns on whether it also monitors the sleep of the persons sleeping next to you, because of the changes in the amounts of light sleep.

Technology expectations were mostly correct, but some exceptions were found. Dreem2 came off P2's head many times, which affected the results and Polar M430 did not record sleep stages and there for was not as competent as the other devices. The devices had multiple different features, Dreem2 having the most unique ones such as it displayed white noise and counted position changes. Also SleepScore Max had some unique features like counting MindScore and BodyScore separately from SleepScore. All the devices used different algorithms to count the sleep efficiency. Polar M430 was the only one to not give the user a numeric scoring on it. Analyzing the data, some irregularities were found in the sleeping time, for example Withings Sleep Analyzer did not recognise sleeping from just laying down in bed for P3 at one night.

This thesis is a research on digital sleep tracking devices that highlights the user perspective. Our sleep data gathering timeline and practises can be applied in future work regarding digital sleep.

8. CONTRIBUTIONS

8.1. Contributions of Roosa Risto

- Project management, Communication with supervisor
- Monitoring sleep with devices, data extraction from devices, filling in sleep surveys, having technology expectations discussions
- Abstract, Tiivistelmä
- Related work: Defining Sleep, Monitoring Sleep
- Study Design: Intro, Phase 1 - Phase 3 (All but Technology expectations and Sleep Survey Questions)
- Study Implementation: Phase 1 - Phase 3
- Qualitative Results: Technology Expectations and Realisation of Technology Expectations (Usefulness, Comfortability, Sleep Disturbance), Ranking Devices, Tables
- Data cleaning and analysis (Heart rate, Position Changes, Respiration, Sleep Efficiency)
- Quantitative Results: Heart rate, Position Changes, Respiration, Sleep Efficiency
- Discussion: Reliability of Sleep Monitoring, Compairing Device Features
- Summary

8.2. Contributions of Joonas Niemi

- Monitoring sleep with devices, data extraction from devices, filling in sleep surveys and having technology expectations discussions
- Related Work: Industrial Products (Dreem2, Fitbit Versa 3, Withings Sleep Analyzer, Polar M430)
- Study Design (Technology expectations, Daily Sleep Tracking Survey Questions)
- Qualitative Results: Technology Expectations and Realisation of Technology Expectations (Ease-of-use, Functionality, Reliability)
- Data cleaning and analysis (Sleep Stages)
- Quantitative Results: Sleep Stages
- Discussion: Sleep Stages, Technology Expectations and Realisation, Limitations

8.3. Contributions of Saku Salo

- Monitoring sleep with devices, data extraction from devices, filling in sleep surveys, having technology expectations discussions
- Introduction
- Related Work: Industrial Products (SleepScore Max)
- Study Implementation: Pittsburgh Sleep Quality Index
- Data cleaning and analysis (Sleep Survey, Sleep Times)
- Qualitative Results: Defining Emerging Themes
- Quantitative Results: Sleep Survey, Sleep Times
- Discussion: Devices' Error Sensitivity, Reflection on Related Work, Future Work

9. REFERENCES

[1] Exelmans L. & Van den Bulck J. (2016) Bedtime mobile phone use and sleep in adults. Social Science Medicine 148, pp. 93–101. URL: https://www.sciencedirect.com/science/article/pii/S0277953615302458.

[2] Ibrahim N.K., Baharoon B.S., Banjar W.F., Jar A.A., Ashor R.M., Aman A.A. & Al-Ahmadi J.R. (2018) Mobile phone addiction and its relationship to sleep quality and academic achievement of medical students at king abdulaziz university, jeddah, saudi arabia. Journal of research in health sciences 18, p. e00420.

[3] Sahin S., Ozdemir K., Unsal A. & Temiz N. (2013) Evaluation of mobile phone addiction level and sleep quality in university students. Pakistan journal of medical sciences 29, p. 913.

[4] Surantha N., Kusuma G.P. & Isa S.M. (2016) Internet of things for sleep quality monitoring system: A survey. In: 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), pp. 1–6.

[5] Carskadon M.A. .D.W. (2011) Principles and practice of sleep medicine. Monitoring and staging human sleep. In M.H. Kryger, T. Roth, & W.C. Dement (Eds.)St. Louis: Elsevier Saunders. 5, pp. 16–26. DOI: http://apsychoserver.psych.arizona.edu/JJBAReprints/PSYC501A/Readings/Carskadon%20Dement%202011.pdf.

[6] Giorgi A. (2020) Polysomnography. Healthline DOI: https://www.healthline.com/health/polysomnography.

[7] Pierrick J Arnal Valentin Thorey E.D.M.E.B.A.B.H.A.G.H.J.M.H.M.G.P.V.B.M.C.F.S. (2020) The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. Sleep 43. DOI: https://doi.org/10.1093/sleep/zsaa097.

[8] Polar Electro Oy R. & Technology, White paper, polar sleep plus. URL: https://www.polar.com/sites/default/files/static/science/white-papers/polar-sleep-plus-white-paper.pdf, page 3. Accessed 5.2.2021.

[9] Fitbit Inc., Fitbit sleep technology (2021). URL: https://www.fitbit.com/global/us/technology/sleep. Accessed 5.2.2021.

[10] Arnal P.J., Thorey V., Debellemaniere E., Ballard M.E., Bou Hernandez A., Guillot A., Jourde H., Harris M., Guillard M., Van Beers P., Chennaoui M. & Sauvet F. (2020) The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. Sleep 43. URL: https://doi.org/10.1093/sleep/zsaa097, zsaa097.

[11] Withings, Withings under-mattress sleep tracker - sleep analyzer. URL: https://www.withings.com/fi/en/sleep-analyzer. Accessed 5.2.2021.

[12] SleepScore Labs, SleepScore Max sleep tracker. URL: https://www.sleepscore.com/sleepscore-max-sleep-tracker/. Accessed 8.3.2021.

[13] SleepScore Labs, SleepScore Technology. URL: https://www.sleepscore.com/the-science/. Accessed 8.3.2021.

[14] SleepScore Labs, SleepScore - What is a SleepScore. URL: https://www.sleepscore.com/sleepscore-max-sleep-tracker/faq/. Accessed 8.3.2021.

[15] Buysse D.J. R.C.M.T.B.S.K.D. (1989), Pittsburgh sleep quality index. DOI: https://www.sleep.pitt.edu/wp-content/uploads/Study_Instruments_Measures/PSQI-Instrument.pdf.

[16] Simo Hosio Jaro Karppinen N.v.B.J.O.J.G. (2018) COVER FEATURE MOBILE COACHING IN HEALTHCARE , pp. 34–43DOI: https://nielsvanberkel.com/files/publications/computer2018a.pdf.

[17] Buysse D.J., Reynolds III C.F., Monk T.H., Berman S.R. & Kupfer D.J. (1989) The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry research 28, pp. 193–213.

[18] Nurmi P., Mitä sydämen syke voi kertoa unestasi? URL: https://nyxo.app/fi/lesson/what-can-heart-rate-tell-about-your-sleep/. Accessed 24.4.2021.

[19] A. L.L., Sex is a potent modifier of the cardiovascular system. the journal of clinical investigation. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC166308/. Accessed 24.4.2021.

[20] Best sleeping position. URL: https://www.sleepio.com/articles/sleep-tips/best-sleeping-position/. Accessed 7.5.2021.

[21] De Koninck J., Gagnon P. & Lallier S. (1983) Sleep Positions in the Young Adult and Their Relationship with the Subjective Quality of Sleep. Sleep 6, pp. 52–59. URL: https://doi.org/10.1093/sleep/6.1.52.

[22] What is a normal respiratory rate for kids and adults? URL: https://www.healthline.com/health/normal-respiratory-rate. Accessed 7.5.2021.

[23] How does dreem detect breathing irregularities? [beta]. URL: https://support.dreem.com/hc/en-gb/articles/360010666819-How-does-Dreem-detect-breathing-irregularities-BETA-. Accessed 7.5.2021.

[24] Detect breathing disturbances. URL: https://www.withings.com/mx/en/sleep. Accessed 7.5.2021.

[25] What is the sleep efficiency? URL: https://support.dreem.com/hc/en-us/articles/360030345492-What-is-the-sleep-efficiency-. Accessed 10.5.2021.

[26] What's sleep score in the fitbit app? URL: https://help.fitbit.com/articles/en_US/Help_article/2439.htm. Accessed 10.5.2021.

[27] Sleep / sleep analyzer - what is the sleep score? URL: https://support.withings.com/hc/en-us/articles/360000158287-Sleep-Sleep-Analyzer-What-is-the-Sleep-score-. Accessed 10.5.2021.

[28] What is a sleep score and how is it calculated? URL: sleepscore.com/how-is-your-sleepscore-calculated/. Accessed 10.5.2021.

[29] What if two people are in the room at the same time? URL: https://www.sleepscore.com/sleepscore-app/faq/. Accessed 25.5.2021.

[30] Pacheco D., Alcohol and sleep. URL: https://www.sleepfoundation.org/nutrition/alcohol-and-sleep. Accessed 26.5.2021.

[31] Polar m430 user manual: Sleep tracking. URL: https://support.polar.com/e_manuals/M430/Polar_M430_user_manual_English/Content/Sleep-tracking.htm. Accessed 27.5.2021.

[32] Liang Z. & Ploderer B. (2020) How does fitbit measure brainwaves: A qualitative study into the credibility of sleep-tracking technologies. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4, pp. 1–29.