FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Joose Yrjänäinen
Lauri Sundelin

EFFECT OF SLEEP ON EXPERIENCE SAMPLING METHOD

Bachelor's Thesis
Degree Programme in Computer Science and Engineering
June 2022

Yrjänäinen J., Sundelin L. (2022) Effect of Sleep on Experience Sampling Method. University of Oulu, Degree Programme in Computer Science and Engineering, 49 p.

ABSTRACT

Digital data collection is practically the norm in current research. Countless studies make use of questionnaires and separate data collection methods in aim to gather scientific data from research subjects. Experience Sampling Method (ESM) collects data using stand-alone reports, very much like in traditional diary surveys. Such an approach reduces the risk of errors caused by subjects memory and reconstruction phase of past experiences. By minimising potential points of failure, data can be made more reliable. In addition, this allows data collection to be targeted at main points of the study. Purpose of the study is to explore the impact of ESM surveys on the participants, and further how the methodological questionnaires affect the surveys results. The study makes use of a mobile application used for data collection, a server connection, as well as a database for storing the results. The results did not allow drawing of direct links between ESM responses and sleep quality, albeit importance of timing of the ESM questions was identified, need of a necessary saturation of sleep data was confirmed, and an observation was made that the parameters contributing to the same metric should be placed close together timewise.

Keywords: Experience Sampling Method, Sleep sampling, Sleep efficiency

Yrjänäinen J., Sundelin L. (2022) Otsikko. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 49 s.

TIIVISTELMÄ

Digitaalinen tiedonkeruu on nykytutkimuksen normi. Lukemattomat tutkielmat käyttävät kysymysalustoja, sekä useita erillisiä tiedonkeruutapoja kerätäkseen aineistoa kohdehenkilöiltä. Kokemusnäytteenotossa (ESM) tiedot kerätään itsenäisten raporttien avulla aivan kuten traditionaalisissa päiväkirjatutkimuksissa. Tämä lähestymistapa vähentää virhealttiutta osallistujien muistin- ja menneiden kokemusten rekonstruoimisesta johtuvista seikoista, lisäksi mahdollistaen tiedonkeruun kohdistamisen tapahtumiin, jotka ovat tutkimuksen kohdealueella.

Tämän tutkimuksen tarkoituksena on kartoittaa näiden ESM tutkimusten vaikutusta osanottajiin ja edelleen miten ESM -kokemusnäytteenottomenetelmän kyselylomakkeet vaikuttavat tutkimukseen. Työ koostuu yksinkertaisesta mobiilisovelluksesta, palvelinyhteydestä sekä lyhyestä tutkimuksesta saatujen havaintojen tarkasteluun. Tulokset eivät indikoi suoraa yhteyttä ESM vastauksien ja unenlaadun välillä, vaikkakin ESM kyselyjen ajoituksen tärkeys tunnistettiin, datasaturaation tarpeellinen taso huomioitiin, sekä havainnoitiin samaan metriikkatyökaluun kohdistuvien kysymysten parametrien sijoittaminen ajallisesti lähekkäin.

Avainsanat: kokemusnäytteenottomenetelmä, unen mittaus, unen tehokkuus

TABLE OF CONTENTS

FOREWORD

This Bachelor's thesis is addressed to serve The Center for Ubiquitous Computing (UBICOMP) Oulu by a ExSS study application tool for further research. Thesis is formulated in Applied Computing Project I (ACP1), taught at the University of Oulu in spring 2022.

Oulu, May 20th, 2022

Joose Yrjänäinen, Lauri Sundelin

LIST OF ABBREVIATIONS AND ACRONYMS

ACP1 Applied Computing Project I

APK Android Application Package

CPM Cycles Per Minute

CSV Comma Separated Value

EEG Electroencephalography

ESM Experience Sampling Method

ExSS Experience Sampling Sleep, -study application

HTTP Hypertext Transfer Protocol

JSON JavaScript Object Notation

NREM Non Rapid Eye Movement

PSQI Pittsburgh Sleep Quality Index

RDP Remote Desktop Protocol

REM Rapid Eye Movement

UBICOMP The Center for Ubiquitous Computing

UEQ User Experience Questionnaire

UI User Interface

1. INTRODUCTION

The experience sampling method (ESM) is a fairly commonly used research method to collect data about what people do and think during their daily lives. What makes ESM unique in comparison to other methods is just this ability of capturing life "as is" through questionnaires provided to the subject at random times during their daily lives. Depending on the study, the questions in these questionnaires can include everything from asking about the physical context to thoughts and feelings giving the ability to capture data that could otherwise be hard to collect [1].

While ESM has a lot of advantages, it also has it's practical challenges. One main of the said challenges is the burden that might be placed on the participants. Traditionally this burden could have included things such as having to physically take answers to a laboratory, but even now with all the technological assistance available, having to fill out a questionnaire several times a day might feel cumbersome [2].

In this thesis we are evaluating the correlation of said burden with sleep quality through using different sleep analysis devices in combination with ESM. It has already been proven that poor sleep quality has effects on many aspects from physical health to things like academic performance[3] and so some correlation between sleep quality and having the energy to answer these ESMs can be expected to exist. To test this, we are using different sleep analysis devices evaluated by Joonas Niemi et al. [4] to collect objective data on subjects sleep quality and then comparing it to answers gathered through ESM questionnaires. More precisely the goal is to determine if there is correlation between sleep quality and that how much time and effort test subjects are willing put into these ESM questionnaires. The way we are going to test is that we will conduct a small scale ESM experiment where in addition to collecting the actual ESM responses, we will also collect data on the participants sleep and from the data collected through these two sources, a correlation between the two is attempted to be found. As it turned out, the results obtained did not end up giving any conclusive answers as mainly due to a small amount of participants, the results yielded completely opposite results to what one could have expected based on previous studies on the effects of sleep.

2. RELATED WORK

2.1. Sleep Studies

To understand the motivation for the study conducted in this thesis, it is important to know some background on why sleep is very important topic and what kind of devices, for example, exists for collecting data about it.

2.1.1. Definition of Sleep

Sleep is an crucial factor in an individuals health well-being and performance. and thus it has been studied extensively over the years. There are written papers showing improvements in perceptual- and motor tasks, where subjects have shown performing better right after sleep [5]. While it's clear that getting enough of good quality sleep is necessary for one to function, things like the purpose of dreams has not conclusively been determined.

While the non-fully understood aspects of sleep still exist, a lot is known from the more technical standpoint. For example in [6], the objectively measurable effects of different stages of sleep has been gone over. Basically, the way sleep is divided into categories is that two main categories exists Rapid Eye Movement (REM) sleep and Non Rapid Eye Movement (NREM). These two different sleep phases alternate in cycles through the night. Usually 4 to 6 noted cycles per night measured in adults, where each of these cycles last from 90 to 110 minutes on average. [7]

Furthermore, NREM sleep is also divided into four subcategories: very light sleep, light sleep, deep sleep and very deep sleep. These fore-mentioned categories have been found through monitoring brain activity during sleep and while it might seem for one that duration of sleep is the most important factor in terms of how rested you feel, the time spent in different stages plays substantial role as well. For example in a study from 2005 [8], it is suggested that there is a strong correlation between memory systems and time spent in different stages of sleep.

2.1.2. Collecting Sleep Data

In terms of studying sleep quality and it's effects, there are two main ways to collect data. First and foremost, collecting objective data on sleep with different devices is a good way of getting raw numbers on duration, time spent in different stages and so on (devices used in this thesis briefly introduced in the next subsection). However, while this data is valuable, it does not necessarily contain all the answers as the amount of sleep needed by people on nightly basis ranges from less than six hours to over nine hours [9]. Due to these variations in the population, it is important in sleep related studies to keep in mind that the raw numbers do not always directly correlate with the actual research question.

Amount of sleep different people get every night depends very much on their age, gender, marital status, educational background, occupation, lifestyle [10] and from great many other less significant parameters. This presents far too many variables

in the context of raw data alone, and thus subjective data should also be collected through questions on how the study subject feels about their sleep quality. If the subjective experience is completely ignored, false accusations can be made as the objective numbers should be "adjusted" to the subjects personal sleep habits.

2.2. Commercial Products

In this subsection we will briefly introduce devices that were used to collect the objective data for the study conducted for this thesis. Further description and detailed process flow of the data collected from said devices is presented in the subsection 4.1.2. Also, a general overview of the data possible to collect with both of the devices introduced can be found in Table 1

2.2.1. Dreem2

The Dreem2 is a wireless head-worn device used during the night to record and analyze sleep. The device records three types of physiological signals. Brain activity via five Electroencephalography (EEG) sensors, heart rate and oxygen saturation via a pulse oximeter, as well as, user movement via accelerometer

Headband connects to a mobile device via Bluetooth to store and represent information gathered in a dedicated application. There the sensor data can be monitored or compared with longer-term sleep information. Application is essential as the data of sleep duration, sleep latency, sleep efficiency, etc. is collected through it.

The device also includes two bone conduction transducers, used to create sound stimuli for relaxation purposes. However, these are out of scope of this study.

2.2.2. SleepScore Max

Unlike Dreem2, SleepScore Max is a device that aims to collect accurate data on sleep without being noticeable to the user. This means that instead of wearing the device, for example, it's next to the users bed and it does it's measurements based on movement of the users body. To achieve this, the device uses a very low power radio waves in a similar way to how echolocation works; it emits them and then senses what is happening based on sensor data.

From the data, the device analyses the users sleep and gives it a score between 0-100. This score is based on things like sleep duration, time spent in different stages of sleep and so on. The algorithm that counts the score was apparently created by analysing more than six million nights of sleep. In addition to this score, you also have the access to the raw data on how much you slept, how long it took for you to fall asleep, how much time was spent in light/deep/rem sleep and how much wake time there was during the night. This data is given to you in two formats: total amounts of each category as a "summary" as well as a timeline where you can see in which state you were at which point in time by the accuracy of minute.

Parameter	**Dreem2**	**SleepScore Max**
Sleep Duration	x	x
Sleep Onset Duration	x	x
Light Sleep Duration	x	x
Deep Sleep Duration	x	x
REM Sleep Duration	x	x
Wake After Sleep Onset Duration	x	
Number of awakenings	x	x
Time in bed		x
Score	Efficiency	SleepScore, MindScore, BodyScore
Mean heart rate	x	
Mean Respiration Cycles Per Minute (CPM)	x	
Start Time	x	x
Stop Time	x	Resolvable
Position Changes	x	
Hypnogram	x	
Wake time		x
Number of Stimulants	Overal number	Caffeine, Alcohol, Cigarettes
Sleepiness		x
Stress		x
Energy		x
Focus		x
Mood		x
Exercise		x

Table 1. Data provided by the hardware

Figure 1. Sleep tracking devices, Dreem2 (left), SleepScore Max (right)

3. STUDY DESIGN

The most demanding constraints on the scale this study was the limitations imposed by the fact that only two hardware devices were available for use. These devices are only suitable for collecting data from one person at a time, and therefore this was the single biggest factor limiting the scope of this study. All subsequent decisions are based on this constraint in an attempt to take into account both scalability, and relevance in the configuration which the study was conducted.

3.1. Study Framework

In essence, the study consists of three parts, namely the preparation of the research application, the study period, culminating in the harvest of the results in the final analysis and the conclusion phase. The total execution time for the project was roughly twelve weeks, out of which, about four weeks were reserved for the development of the research application stack (Experience Sampling Sleep, -study application (ExSS) + back-end) and four weeks for the final Evaluation and assessment. Study period length was two weeks excluding one week calibration and testing phase. The remaining time of the mentioned twelve weeks could be considered to have been spent on general miscellaneous things like planning and fixing faced issues.

3.1.1. Production Phase

Application development was one of the most time-critical tasks, which was carried out in four-week production phase, where the ExSS was implemented from start to finish. As the application development was based on ease of implementation scalability, thematic applicability and the necessary amount of data required (data saturation) to compensate for the small sample size of the study. More information on the technical features and, for example, the application screenshots can be found in the chapter chapter 4. Finally, as said the testing phase was carried out in parallel with the P1 Dreem2 calibration week.

3.1.2. Study Period

First week of the two-week study period was spent preparing the subjects, the application stack, and hardware. P1's research instrument Dreem2 required a week-long calibration period in with the headband was self-calibrated for the user. Due to time constraints imposed, this time was also used to test the ExSS, in particular to verify the triggering of the notifications and to ensure the functionality of the response-fields, which are critical part of the response gathering and thus high in priority.

When entering the survey phase, P1 and P2 were involved in the study practically full-time to the best of their ability. The intention was to collect data from as many nights of sleep as possible. Alternative to that, effectively five hours of each day were

filled with questionnaire time slots, bringing the total theoretical time spent per day on about 14 hours per participant.

3.1.3. Evaluation and Assessment

At the end of the study period, participants were interviewed using a survey based on the UEQ format. This form was one-off questionnaire, designed to capture the post-application experience based mainly on Pragmatic Quality Figure 16, thus aim at identify whether there were problems with the usable of application that should be addressed in the final phase. Moreover, already during study period, participants had the opportunity to report errors and bugs, thus allowing rapid troubleshooting and fixes.

4. DESIGN

4.1. Application Design

In the survey phase, we used a custom data collection application tailored for the study (ExSS). In this application we modeled different kinds of questions for collecting experience data to be compared with ESM survey. In order to achieve this, the application needed to fulfill some design requirements, an overview of which, can be found in Table 2.

Desired quality	Description
Scalability	Extendable to cover a larger number of participants if necessary?
Editability	Are the questions, their times and types adaptable for future use?
Ease of use	Is the ExSS easy enough to use not to become a burden for surveying participants
Understandability	Does the data reflect participants thoughts and state of mind?
Communication	Do the logs arrive on time? Is the database connection stable?
Questionnaire timing	Are the questionnaires available at the right allocated times?
Notification deployment	Will the participant receive a notification at right times?
Usage monitoring	Does the ExSS gather needed relevant information?

Table 2. Functional and qualitative requirements of the ExSS

4.1.1. Daily Questionnaires

The questions are divided into five sections on different topics. Each of the topics aims to gather information with progressively more difficult responses. For each of the topics, notification is sent to the user's mobile device, reminding them, to fill in the questionnaire. Notifications are configured to be sent on 9am, 12pm, 3pm, 6pm and 9pm. The respondent has an opportunity to fill in the questionnaires as well as they can, before the allocated time-slot closes one hour after opening. The respondent is rewarded with answer percentage displayed in the mobile application home-screen, which shows them both, percentage of questions answered during ongoing day as well as their weekly average.

Questions themselves are divided to different subcategories, in which the subject was exposed to progressively more demanding tasks. The categories are sleep quality, previous day, energy, leisure and sustenance (Table 3).

Sleep related questions will be asked when sleep is most likely to be still relatively fresh in the memory of the subject, this occurring in our survey every morning at

9am. Respondents are expected to answer sleep quality related questions, provide information about their bedtime activities, time for bedtime, and how they felt about it.

ESM topics & survey slot opening times				
9am	12pm	3pm	6pm	9pm
Sleep quality	Previous day	Energy levels	Free time	Sustenance

Table 3. How the questions are arranged within the day and what their focus topic is.

Questions related to previous day, scheduled for noon, inquire the respondent to evaluate their previous day, name one of the most memorable things from said day, and conditional question about why the respondent had not answered some part of the previous day's questionnaire.

There are two times in each day when people are more vulnerable to feel tired. 2am to 6am and 2pm to 6pm, former being stronger than the latter [7]. Energy levels are the topic of the afternoons questions. Respondent is inquired whether they have started free-time, what the respondent is currently doing and that are their plans for the rest of the day.

Last topic of the day, sustenance, is designed to gauge the respondent's use of stimulants and eating habits as these can affect greatly the following night's sleep.

This aforesaid scheduling was chosen to enable a rapid survey cycle where the app would send out notifications in average about five times a day reminding the respondent to answer the survey. In addition, in situations where the respondent had time and possibility to provide further information as part of a more analytical reflection on their experience.

4.1.2. Hardware

As the equipment used for the study was not the same between all the respondents, the raw data obtained had to be treated with necessary care, accounting for the impact of hardware itself. We carefully considered the possibility how the use of the equipment would be reflected in the results, potentially disturbing the sleep of the subject and thus affecting the results of this study. One possibility of unwanted external influence affecting this study is graphical interface of the data-collection equipment implemented in the devices used as the user might intentionally or unintentionally obtain information about their sleep data further effecting their ESM responses.

In order to conduct the final analysis, the data from the hardware is exported in CSV format and stored separately for later comparison and review. In practise, this was done by instructing the participants on how to perform said export and then afterwards have them send the .csv file to us for further analysis.

4.2. Data Collection Techniques

4.2.1. Data Format

For the ESM questionnaires the data format for saving the responses locally, that we decided to go with, was JavaScript Object Notation (JSON) – which is a lightweight text-based, interchange format. [11] It is used as a portable representation of structured data, [12] thus being relatively easy format to process if processing of locally stored responses would have been required for any reason, as well as easy to transmit to a server.

As the back-end server implemented for this project was in all simplicity just a python script capable of receiving HTTP post requests, JSON as the data format made a lot of sense as it was convenient to just have the responses as the content of HTTP requests.

Hypertext Transfer Protocol (HTTP) in itself is an application-level protocol, offering lightness and speed needed for distributed systems, [13] used commonly in collaborative and hypermedia applications. Due to its ubiquity and ease of use, it was chosen as the basis for communication method for data collection.

HTTP POST method is used for requests where the destination server accepts the message enclosed in the request as a new subordinate specified by the Request-URI in the Request-Line. [13 chapter 8.3] In this case destination server receives ESM response entity from the mobile application, extending it through an append operation.

Comma Separated Value (CSV) formatting was used as a package for transferring hardware data to further analysis. Each record is placed on a separate line, delimited by a line-break. [14] This method of data transfer was used as it was the default format in which it was possible to export data out of all the hardware used.

Generally, format of the data in the case of this thesis was not very important as long as it was capable of holding the information required and it was in at least somewhat easy to process format, but it is worth keeping in mind that using pre-existing sleep analysis devices will cause some limitations in areas such as data format due to the implementation choices made by the manufacturers.

4.2.2. Scoring and Models

Because the main focus of this study is to look for correlation between sleep quality and how much effort subjects are willing to put to the ESM questionnaires, the actual content of the questions (and thus their responses) are not as important as collecting data on user behaviour regarding responding to them. Despite of this, we wanted to at least have some of the questions provide information that can be applied to some more standardised models used in relevant research and one of these was the "circumplex model of affect" [15]. Simply put the idea of this model is to give the ability to estimate ones mood by having information on how activated the subject is and how pleasant they experience whatever they are doing at the moment. For example, if one is deactivated and finds it pleasant, their mood can be estimated to be relaxed/serene and so forth. A commonly used graphical representation of this model can be found in Figure 2.

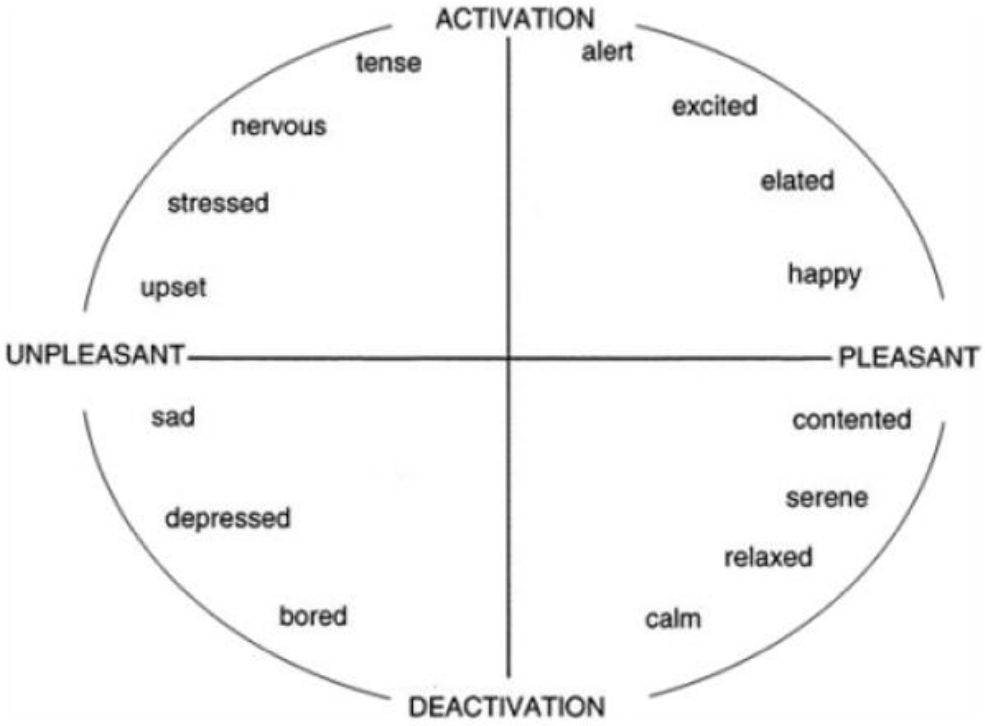

Figure 2. A graphical representation of the circumplex model of affect.

The way we are obtaining the information about the two axis (activation/pleasantness), is through questionnaire regarding free time (sent at 6pm). In it, we first ask the subject what they are doing through a multiple choice question (giving information about the activation level) and as a follow up, we ask them to rate it's pleasantness using a number input (range 1-10). Through these, it should be possible to somewhat accurately estimate the subjects mood using the model.

Another scoring/model related matter in this thesis is that in a paper from from 1988, an scoring system for accessing sleep quality is introduced [16]. While this system in itself is too wide in scope for our thesis (especially due to the inability to obtain enough information through the short ESM questionnaires), its assessment criteria can be utilized in such way that it can be used to calculate a second set of sleep scores based on raw data effectively mitigating the differences of scoring between the two devices used. In chapter 6, we will look for correlation both by using the scores provided by the devices themselves as well as scores calculated using this model.

5. IMPLEMENTATION

With the design decisions presented in the previous chapter in mind, the following step was to perform the actual implementation. In this chapter we will give a more detailed overview of how everything from ExSS application to its back-end system were implemented (section 5.1). In addition, a more detailed description of the implementation of the study itself is provided in section 5.2.

5.1. Application Implementation

5.1.1. Tools

The application called ExSS was created with React Native which is an JavaScript framework, originally created by Facebook for creating native user interfaces [17]. The framework is relatively easy to use, and in the case that our test subjects might prefer using different ecosystems such as iOS, the app should not require any major modifications to be functional on both platforms, Android and iOS.

5.1.2. Questionnaire Customization

The app was made easily customizable so that it could also be used in other research utilizing the ESM method. The way this customizability was achieved is that the questions are defined in a JSON format file which the app then interprets. This approach does require rebuilding the app installer to change the questions but due to the limited time of the project, this method was ultimately used. Worth mentioning, however, is that due to the use of JSON, the app could relatively easily be expanded in the future so that the client could automatically fetch the ESM data from an outside source. The JSON format is as follows:

```
{
    "ESMs": [
        {
            "id": Index of the ESM (0, 1, 2...),
            "time":
            {
                "hours": Hour for ESM to open,
                "minutes": Minute for ESM to open,
                "closehours": Hour for ESM to close,
                "closeminutes": Minute for ESM to close
            },
            "questions": [
                {
                    "type": "Question type",
                    "question": "Question text"
                },
                {
                    "type": "Question type",
                    "question": "Question text"
                },
                ...
            ]
        },
        ...
}
```

Figure 4. ESMs.json which contains the structure and details of the questionnaires

In other words, each ESM questionnaire is an JSON format entry in an array called "ESMs", and each entry containing information on what time the ESM is supposed to be open and what questions (of which type) the ESM questionnaire should include. The supported question formats can be found in Table 5.

5.1.3. User Interface

In terms of the user interface of the app, the home screen displays information whether the user has an active ESM questionnaire to answer, as well as it displays the percentage of questions answered during the ongoing day. User is also presented with an average weekly percentage, from which the progress can be tracked. A screenshot of the home screen can be found in Figure 3

In the current implementation the weekly percentage works in such way that completely empty days are ignored. While this might lead to a situation that the user gets higher percentages as reward despite completely ignoring the surveys, it does, on the other hand, allow for things like break days in the studies without causing issues with this functionality. Also with this design, the ESM JSON file does not require information on the start/end dates since days before the beginning of the survey are

Type in JSON	Description/limitations
slider	Slider input with values 0-10
number	Text input that only allows for numbers to be inputted
text	Text input
multiplechoice	4 configurable choices. Varying the amount of checkboxes turned out to be rather difficult so this currently only supports four choices
yesno	Radio-button type of input (yes/no)
multiplechoice_SPECIAL	This is a non-configurable special case of multiple choice. In our study it was used to generate a question about previous day's responses

Table 5. Types of questions supported by the app

going to be empty from the responses point of view and thus ignored when calculating percentages.

The question views are all implemented to look as similar as possible with the only thing changing being the actual response input. These screens have the question on top, then the area for inputting response and then at the bottom they have two buttons; one for progressing to the next question and one for stopping (examples of the screens are in Figure 4). With the press of either button the response is sent to our server and on the press of stop (or upon finishing all questions), all the responses accumulated are also saved in memory as a JSON format string. Also the number of answered questions is saved separately for making the calculation of percentages easier. Since the storage functions with key-value pairs, the chosen format of saving is as follows:

```
DD_MM_YYYY_ESMID: {"question":answer,"question":answer,
                   ...,"endachieived":true/false}
DD_MM_YYYY_ESMIDnum_answered: amount of questions answered
```

In order to achieve functional app in a relatively short time span, a variety of React Native libraries, in addition to the built-in libraries, were used. List of these external libraries as well as the purposes of each one of them can be found in Table 6.

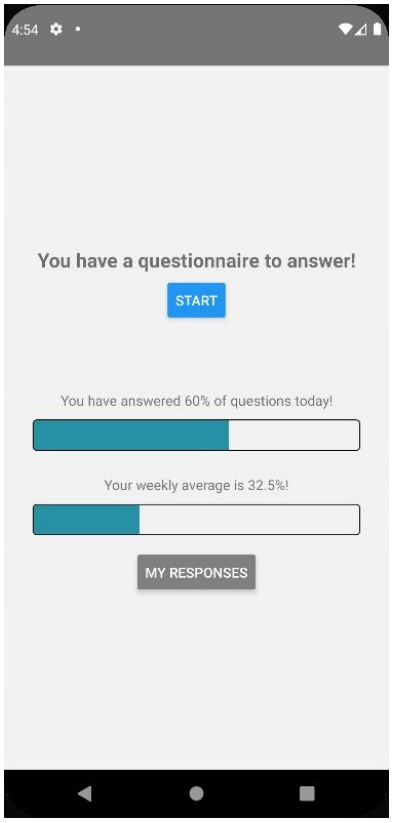

Figure 3. Home screen of the up when user has ESM to answer

Library	**Purpose**
@ptomasroos/react-native-multi-slider	Implementing questions with slider input.
react-native-paper	RadioButton component was used for implementing yes/no questions.
react-native-bouncy-checkbox	Implementing responses to multiple choice questions.
react-native-progress	Drawing percentage bars on the Home screen.
@notifee/react-native	Implementing timed notifications based on ESM open/ close times.
sync-storage	To make accessing content in storage easier. This library allows for asyncstorage to be used synchronously.
@react-native-async-storage/ async-storage	Saving of content to local storage.
@react-native-community/ netinfo	Skipping data sending attempts in case of missing network connectivity.

Table 6. Libraries used in the application

5.1.4. Creating Notifications

Notifications are generated on the first start-up, when the user consents / accepts the terms of the survey application. Notifications are generated from the ESMs.json -file

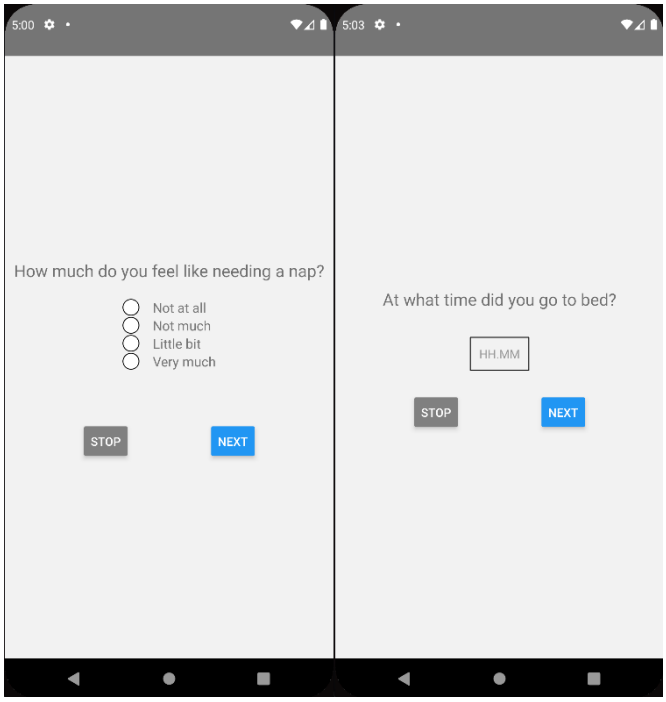

Figure 4. Examples of question views

specified in more detail in subsection 5.1.1, from where each notification is assigned its own time of deployment. This approach moves the collective control of research questionnaires to a single location and if so desired, allows quick modifications of these said parameters. Forwarding research can be carried out at times of interest to the researcher.

5.1.5. Notification Format

Visual part of the ESM notification consists of essential information such as the header, body, icon and a descending chronometer, which purpose is to create a visually represented limit to the window of time during in which a questionnaire can be answered. Thus, implying to the user that the questions are not open all day long. This steers the study in the direction of rapid question cycles, the intention being to produce discrete data at specific times of the day, rather than keeping the questions open all the time. An visualization of what the notifications looks like for the user is in Figure 5

Other, but not visual parameters include id, channel, intended events after pressing button, etc. Each notification is given a specific id to identity the correct notification in later steps. For example, when a user rejects a notification, the respected id is carried forward to the server log message, to better identify the specific notification with in which the user interacted during the study-phase.

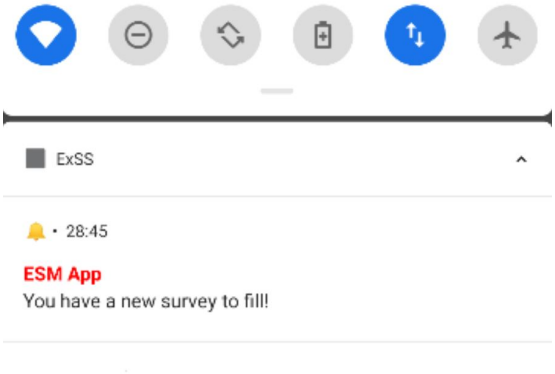

Figure 5. Notification with 28m 45s of time to answer left

5.1.6. Notification Intractability

To interact with the notification, the used library offers a couple of listening methods, two distinct ones have been utilised in this study. Since the idea of the application partly revolves around the precise triggering of notifications, monitoring their usage, as well as interpretation of the user's actions based on these events, a reliable and reproducible source of information is crucial to draw the right conclusions and thus ensure the necessary number of distinct data points to conclude whether ESM's are disruptive or not.

5.1.7. Foreground and Background Events

When listening to events, the application can either be in the foreground or in the background. Exaggerating this means that the application is either open and visible or the application is hidden / closed.

Normally, event monitoring in the foreground could be utilised to use push notifications to extend the general operability of applications by providing an additional tool for displaying important information in the notification panel. Example of these could be a visualization of a file transfer progress or, to present media control tools in a hypermedia application. Due to the small number of interaction points in our application, foreground notification handling is not particularly needed in the context of this study. Functionality of the application relies heavily on the timely delivery of notifications in the background, and on repeatable functionality. Thus, the onBackgroundEvent handling will play a significant role in all areas withing the application. Functionality of said method can be expanded to cover also the same functionalities as onForegroundEvent handler, it being built almost the same way to ensure that questionnaire notifications get through even when the application is in the foreground.

5.1.8. Background Event Handling

With onBackgroundEvent method interaction callbacks are listened, further sending them forward to server for further study.

Application listens for three callback event types: DELIVERED, DISMISSED and PRESS (Table 7). Main actions revolving around logging if or when the notification was successfully received by the users device, this is to verify whether the failure to interact with a notification is due to, for example, hardware, and thus to better clarify that the failure conclusions to press the event was solely due to the user and to separate anomalies due to hardware / software to not spoil the results of the research.

Event types for notifications		
DELIVERED	DISMISSED	PRESS
Notification deployed	*User dismissed the notification*	*User pressed notification*

Table 7. Event types for notifications

DISMISSED is an valuable event to be logged to the server, as it is one metric of whether ESM notifications are burdensome for the subject, whether the dismissal of these notifications indicate that it is not a good time to complete the survey and/or that there is no motivation this time to complete the questionnaire. More on the details as well as reasoning of these results in chapter 8

PRESS, i.e. the pressing of a notification is relevant as the user's way of entry into the application can be identified and specified to the level of notification id the application was launched from. Thus, for example the time between delivery and pressing can be calculated from the two time points sent to the server between successful delivery log and pressing log. Based on this, conclusions can later be drawn about the time taken by the user to respond to the notification, or for example, whether the response was almost immediate each time.

5.1.9. Server / Back End

The server used in our project was in all simplicity a python script which was capable of receiving HTTP requests from the app. When one of these requests was received, the content of it was saved to a CSV format file. The data in each of the requests consisted of variables: timestamp, user id, question, response, state and an extra payload. Not all of these values are required nor even relevant in every transmission. For example request containing info about opening the app would include user id, payload with the actual information and state while having rest of the variables as null.

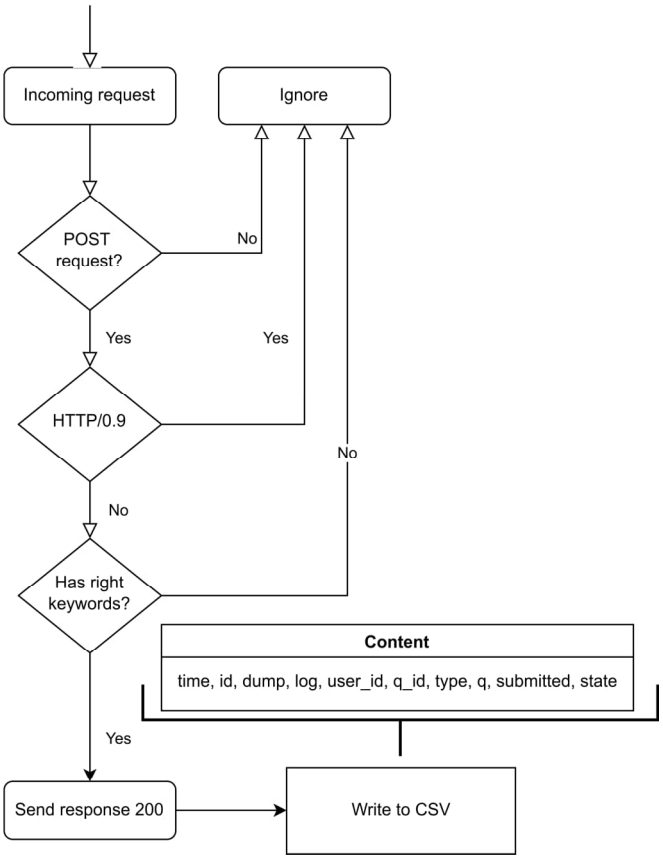

Figure 6. simplified request flow

The actual server script was based on zero dependency python web server [18] to which we added required CSV functionality as well as some filtering for incoming requests. For hosting this server for the duration of our study, we just had port opened to a Raspberry Pi 4 running it. While this solution was fine from performance point of view (considering the small scale of our experiment), some issues were encountered which will be discussed in more detail in chapter 6.

Initially the server implementation did not have any filtering of incoming requests but due to frequent attacks towards it before and during the study, some simple logic was added (illustrated in Figure 6). The filtering logic was that first and foremost HTTP/0.9 requests were discarded straight away (as they seemed to be popular in these attacks). After this, the body of the request was checked to find out if it included all keywords expected from the app (time, type, q, q_id, submitted, log, state). If all of these were not found, the request would also be discarded. Worth mentioning is that while this filtering did have some positive effects, overall the core problem was in the library used to implement the server so that a lot of attacks would crash it immediately upon starting to process them rendering the filtering useless.

5.1.10. Client Connectivity

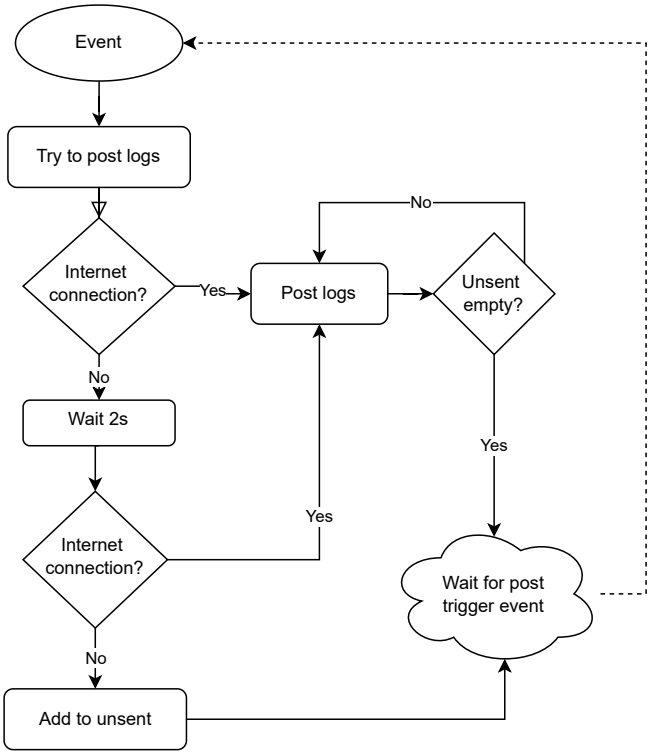

Figure 7. Client networking flow

When sending data to the server, it is important to keep in mind that there are many variables that might cause sending to fail. As already mentioned, all responses are also stored in the memory in addition to being sent to server. While this is good for redundancy, the actual analysis will get difficult if data needs to be combined from several sources and thus the server communication itself also has implementations to avoid the need of doing that.

Essentially the flow in the networking part of the app (Figure 7) is that first it is checked whether or not internet connection is available. If not, 2 seconds is waited, after which connectivity is checked again. If a connection is still not available, the content is stored to memory as unsent. The same is done in case that internet connection is indeed available but server does not respond within 2 seconds.

Whenever successful post occurs, it is after that checked whether unsent memory is empty or not. If there is something to be sent, the same log posting process is restarted for that item and the loop effectively continues until nothing remains as unsent.

While two second timeout might be a bit short (especially for the time that the app is willing to wait for a response), it was chosen so that likelihood of closing the app too soon would be reduced. The longer the time a response is awaited for is, the likelier it is that the app will be closed before successful post effectively causing data to be lost.

5.2. Study

5.2.1. Instructions for the Subjects

The subjects were provided with a sleep monitoring hardware for the duration of the study as well as before-mentioned ESM application (ExSS) for filling in the questionnaires. The application was installed on participants mobile device. All subjects participating had an android-based operating system on their primary phones, so the installation was straightforward and the application was installed from the Android Application Package (APK). This package contains the application itself, but also the associated resource-, as well as the data files.

Each user were assigned with a unique identification id, all the data generated during the study for this id could be combined, thus creating a data twin in the database stripped of any personal data that could be linked to a person. User id is beyond this step no longer linked back to any data that might lead to identification.

Some instructions for the usage of the sleep tracking hardware was also provided to the participants in order to mitigate the risk of collected data being affected by uncertainty in regard of the usage of the devices.

5.2.2. Study Phase

The two week study with human subjects was conducted between 1.4.-14.4.2022. Both participants were instructed to use sleep tracking hardware introduced in section 2.2 every night during this period and to answer as many of the questionnaires as possible. The consent of the subjects was asked in the app so that upon installing it, they would have to input their id as well as to tick a box giving consent to participating in the study. During the experiment, no further instructions nor feedback of their progress was given to the participants (other than percentages shown by the app) in order to prevent any alterations in results obtained.

5.2.3. Data Collection

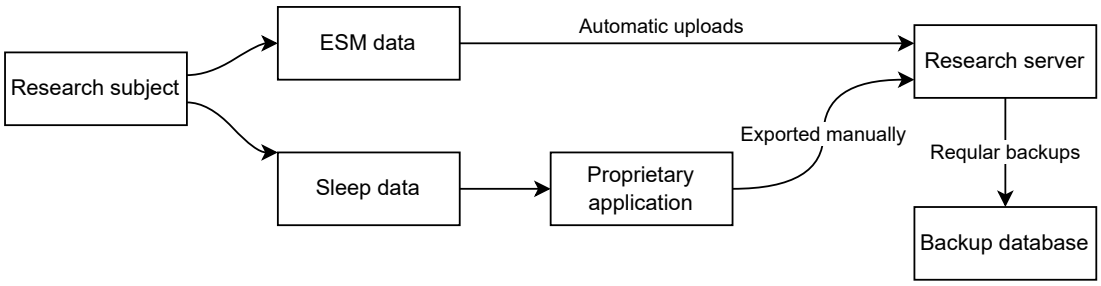

Figure 8. Flow and process of the data

During the study, due to the automatic collection of questionnaire results from the application (ExSS), manual data gathering was not necessary regarding ESM submissions. With the sleep monitoring hardware used, however, data had to be

collected manually. In both the sleep tracking devices used, the hardware manufacturer had implemented a ready-to-use export feature in their application stacks, so that allowed the export of the comma separated values be ready for use in the next phase of the study straightaway.

The overall data collection workflow is illustrated in Figure 8 which also takes account of having automated backup system in place on the server where results collected were saved. In practise this was done by a python script that regularly uploaded the accumulated CSV file to Google Drive.

6. EVALUATION

6.1. Evaluation Plan

In order to evaluate the effects of sleep quality in ESM responses, scoring of said responses must be first defined and then carried out. In practise a score was calculated based on amount of responses and quality of answers with a predefined model. As most of the question types used in this study are relatively quick to answer, assessing the quality of answers in distinct questions is hard. In our assessment, we were not so concerned with the pecking order of questions withing the same category, but as difficulty levels are defined, the questions are linked in to match scores according to their level of difficulty (Table 8) through having the questions progress from easy to hard and then having the calculated score be effected by number of questions answered.

6.1.1. Pitfalls of Engagement Pattern Recognition

To a certain extent, interactivity and engagement can be measured solely by the behaviour outside the application itself. Dismissing application notifications, leaving them open for a long time before interacting or not interacting with them at the given time, notifications thereby closing and not answering the questionnaire itself indicates a low motivation to complete questionnaires at before-mentioned times. Quantifying this is unsound practise and should not be used independently to draw final conclusions without a dedicated study of its own right. On the other hand, the level of interaction and engagement can be observed from the perspective of the rapidity of time between delivering the notification and interaction with it. The shorter this measured time being, the more motivated the perceived state of the user is at completing the task. This behaviour can further be observed on the occurrences of applications launches outside of the scope of notifications, for example starting the ESM application from the home screen. In this case it is perceived by getting application launch log without attached notification id carried with the transmission. These launch events can lead to situations where the user does not have a questionnaire to answer, them only being active at predefined time-slots, but the users engagement should still be captured this helping with breaking down the question when the user is more able to complete these given tasks. Further down, this allows researchers to corroborate a thesis with the findings from these tendencies.

6.1.2. How to Quantify Effort

In addition to before mentioned metrics of notification interactions and difficulty-tier scoring, one way of measuring quality of responses and thus effort could be based on the time spent answering. However, as the questionnaires come throughout the day, it is very well possible that user might be in such situation that they are doing something else at the same time causing the question views to be open for long time despite the actual effort being low. Because of this it is important to consider possible outliers in the sense that if the total response time is a lot longer than generally, something like

this might have occurred. On the other hand, the open text input type of questions make an exception to this as the longer the response, the more effort it will have required and thus these should be taken into consideration despite of the total time spent answering.

Question difficulty	Question type
Easy	*Slider, yesno*
Medium	*Number, multiplechoice,*
Hard	*multiplechoice_SPECIAL, text*

Table 8. An early draft of difficulty classification

6.1.3. Scoring Method

For the daily scoring of the ESM responses, the following formula was used (1):

$$\frac{Answered\ questions}{Deployed\ questions} * 100 + dailysum(time\ score) + dailysum(text\ score) \quad (1)$$

The scoring of open text input can be found in Table 9 as well as the scoring used for response time time in Table 10. The scoring was only decided on after the data collection was finished in order to have better understanding on things like how long answering had usually taken. Considering this information, the limits for different scores were decided on so that distribution between them would be balanced meaning that for example singular, longer than normal, response times would not be given too much weight. In addition, in order to apply more weight to score representing time spent answering, range of it was double to the scoring of open text responses.

In practise, for answering-time scores, 90 seconds was decided to be the upper limit on increasing the score. While the recorded values went up to around 250 seconds, there was only total of four times where the value was higher than 100 seconds and thus these could be considered outliers and scoring them higher than the rest of the values is not sensible. The scoring of the open text answers is quite the opposite, however. All lengths, were six or less with the exception of one being eleven. As the response with even this higher value had a proper response content, it was chosen to allow it to have a higher score compared to the others. So unlike in response times, "grouping" of the higher values was not done but rather the one "outlier" response was allowed to get one point more in comparison to the others (by being the only value falling into the category 7+ in Table 9).

Response length (words)	Score
0	0
1-3	1
4-6	2
7+	3

Table 9. Scoring of open text responses

Response time (s)	Score
<15	0
15<t<30	1
30<t<45	2
45<t<60	3
60<t<75	4
75<t<90	5
>90	6

Table 10. Scoring of time spent answering

6.2. Overview of the Data

During the two-week data collection period, a total of 78 ESM questionnaires were at least partly answered by the participants. Both of them received a total of 70 questionnaires to answer and the other participant responded to 84,2% of the questionnaires while the other to 27,1%. This low response rate for the latter one is, however, partly explained by technical issues encountered by P1 at the beginning of the study period where their mobile device had blocked internet connection from our app (and despite our implementation of backup system for missing connection in the app, the responses were ultimately lost). Sleep data was mostly successfully collected with the only exception being data missing from last two nights on P2. For one night the missing data was explained by SleepScore Max randomly freezing and for the other it was caused by their phone rebooting during the night.

Further analysis on the data based on the metrics introduced in section 6.1 can be found in section 6.3, but in this subsection a brief overview of relevant data is given.

6.2.1. Values for ESM Scores

Starting with time spent answering, the longest time recorded was 250,5s while the shortest was only 13,1s. The average was 44,0s and median 33,7s. For the most part (93,6% of the responses), time spent answering was under 60s meaning that the actual effort needed to filling a full questionnaire was not very high. Distribution of the times is shown in Figure 9.

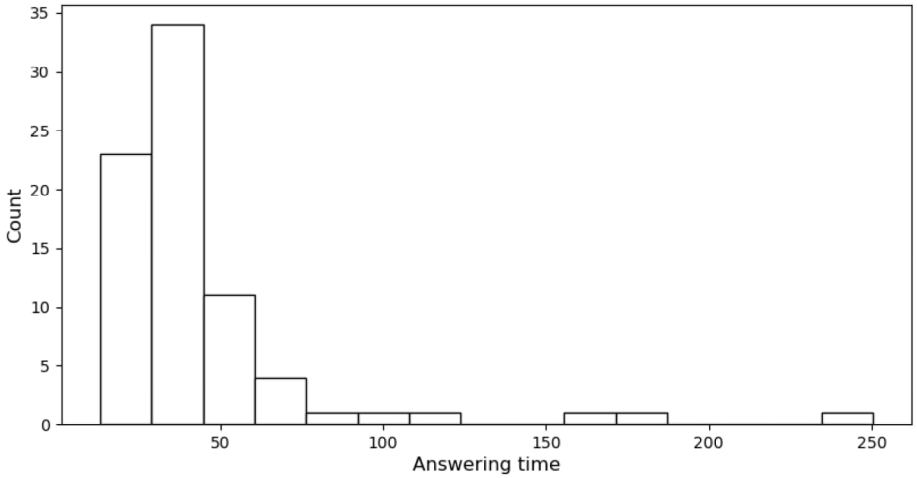

Figure 9. Histogram of answer times

In terms of user activeness, we will mainly focus on time between notification and user interacting with the app. In this regard, the shortest time was 1m 1s and the longest 57m 54s. Average was 16m 36s, median was 11m 19s and the variance was 15m 51s. Distribution of these can be found in Figure 10. From this it is clear that it was more common for the participants to respond to the ESMs close to the opening time suggesting somewhat high motivation to participation.

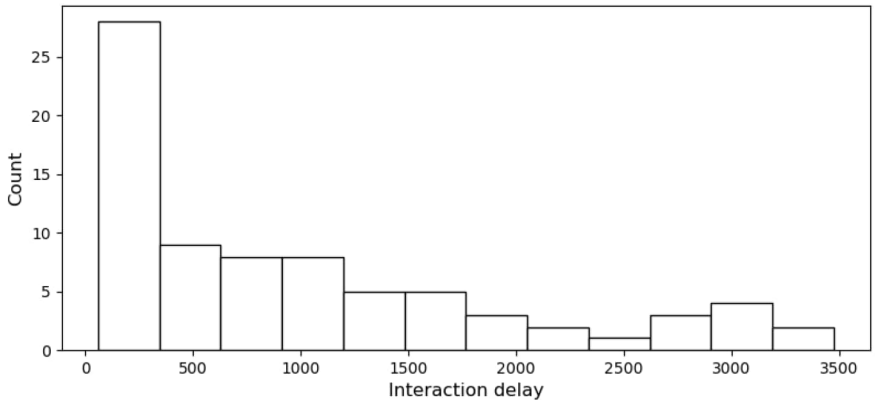

Figure 10. Histogram of interaction delays

The lengths of open text answers very much follow the same pattern with the previous ones in a way that lower values dominate. The variance etc. are significantly lower, however, due to the fact that the value range for open text answer length only varies between 1 and 11. Distribution of these is visualized in Figure 11 and the main statistics are: average: 3,04, median: 3, and variance: 1,73.

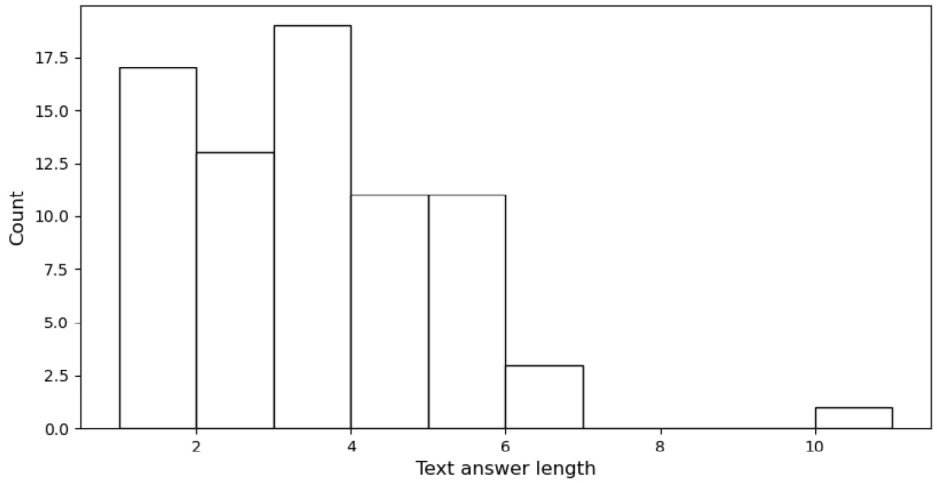

Figure 11. Distribution of answer lengths

Detailed score values and calculated ESM scores can be found in Table 11 (P1) and Table 12 (P2).

Day	Answer %	Sum of time scores	Sum of text scores	ESM score
1	0,00 %	0	0	0
2	0,00 %	0	0	0
3	0,00 %	0	0	0
4	0,00 %	0	0	0
5	41,18 %	4	2	47,18
6	64,71 %	4	2	70,71
7	58,82 %	4	3	65,82
8	35,29 %	9	2	46,29
9	23,53 %	2	1	26,53
10	100,00 %	6	5	111
11	35,29 %	3	2	40,29
12	23,53 %	1	0	24,53
13	0,00 %	0	0	0
14	0,00 %	0	0	0

Table 11. ESM data of p1

6.2.2. Sleep Metrics

As both of the sleep tracking devices had a built-in scoring system, no score value had to be calculated based on their data in the first stage of analysis. The relevant data collected by these devices is available in Table 13 and Table 14

Day	Answer %	Sum of time scores	Sum of text scores	ESM score
1	82,35 %	18	6	106,35
2	76,47 %	10	5	91,47
3	82,35 %	12	6	100,35
4	58,82 %	8	5	71,82
5	100,00 %	15	11	126
6	82,35 %	6	5	93,35
7	100,00 %	6	6	112
8	94,12 %	13	7	114,12
9	64,71 %	7	6	77,71
10	76,47 %	6	5	87,47
11	100,00 %	11	6	117
12	100,00 %	14	8	122
13	100,00 %	6	6	112
14	58,82 %	4	3	65,82

Table 12. ESM data of p2

Night	Sleep duration	Sleep efficiency	Number of awakenings
1	6:52	87	4
2	7:26	94	4
3	7:08	94	4
4	7:48	94	4
5	7:13	95	5
6	6:06	93	3
7	5:36	88	2
8	5:39	94	1
9	8:55	98	2
10	6:45	97	1
11	5:28	96	1
12	6:57	92	2
13	8:29	95	3
14	7:20	97	2

Table 13. Sleep duration, efficiency and number of awakenings recorded by Dreem2 (P1)

Night	Sleep duration	SleepScore	Number of awakenings
1	7:50	95	2
2	6:18	84	3
3	7:16	93	1
4	7:18	89	3
5	6:19	86	4
6	6:56	94	2
7	7:10	94	2
8	5:58	82	4
9	9:10	94	4
10	7:28	95	4
11	6:56	93	4
12	7:20	94	3
13	-	-	-
14	-	-	-

Table 14. Sleep duration, Score and number of awakenings recorded by SleepScore Max (P2)

6.3. Analysis of the Data

6.3.1. Brief Description of Key Parameters

As the terminology and measured variables between the two devices used are not equal, an overview of some key parameters is in place to clarify following analysis.

Start time corresponds the beginning of the night. This value represents when the user clicks "Start your night" in the "Sleep" section of the App. Stop time corresponds to end of the recorded night, for example the removal of the device or interruption exceeding one hour. Sleep onset duration corresponds to the time between user begin to fall asleep and the time of first phase of sleep recorded, based on devices electroencephalogram. Number of awakenings correspond only the number of nocturnal awakenings. SleepScore Max defines "Bedtime" as the time when the user has activated the sleep monitoring and thus is gone to bed.

Both SleepScore Max and Dreem2 display the estimated time taken for the user to fall asleep with tags; Time to Fall Asleep and Sleep Onset Duration, respectively.

6.3.2. Correlation of Scores

Starting with the main question of the study; do the calculated ESM scores have any correlation between them and scores given by our sleep tracking devices? There are several ways of looking for correlation but in this study, we are using three main ways introduced in [19]; scatter plot, Karl Pearson's correlation and Spearman's correlation. As mentioned, due to technical difficulties, there are such days in the dataset where either the sleep score or the ESM score is zero not representing the reality and thus these kind of days were discarded during analysis.

Starting off simple with a scatter plot in Figure 12, we can see that no clear correlation exist: high sleepscores can result in poor ESM performance while lower sleepscores might result in higher ESM scores. Fitting a trend line to the plot reveals slight negative correlation and the results of calculating both Karl Pearson's correlation and Spearman's correlation, reveal similar results. Spearman's correlation is -0,256 (with p-value of 0,274) while Pearson's correlation is -0,309 (p-value 0,185). Based on these raw correlation measures between the two scores, it slightly seems as if the correlation is indeed negative meaning that the lower the quality of sleep, the better the ESM performance. It is, however, very important to consider the fact that in our study, we only had two participants and the study period only lasted for two weeks resulting in only total of 20 valid value pairs meaning that value pairs of low sleepscore and high ESM scores might actually be outliers which due to small data set, are pushing the correlation down. Another important point is that the negative correlation measures are relatively small meaning that this could, especially considering the small data set, also be interpreted as having no correlation in practise. In short, due to small data set, it is impossible to unambiguously determine whether correlation exists or not.

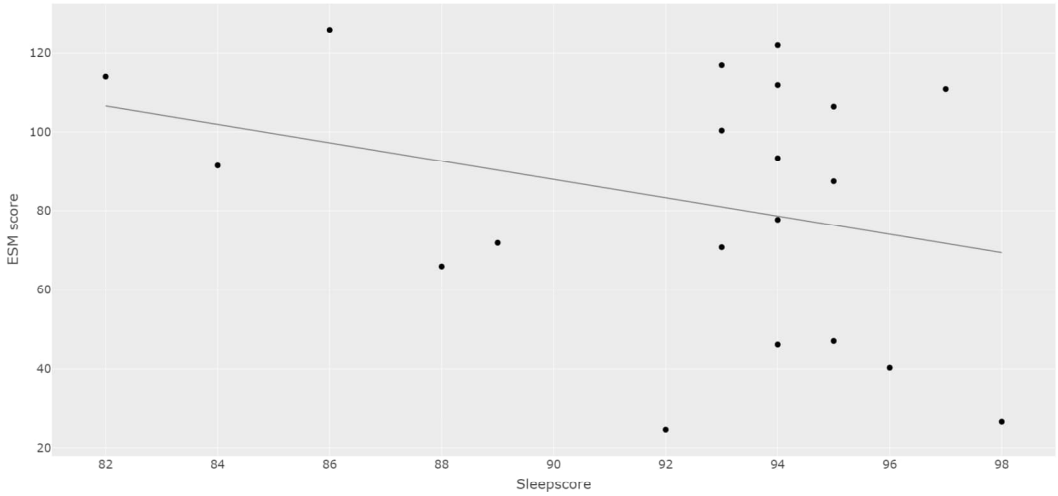

Figure 12. Scatterplot of sleepscore vs. ESM score

6.3.3. Habitual Sleep Efficiency and Pittsburgh Sleep Quality Index

Because sleep quality is difficult to quantify and determine precisely, we wanted to use a wide variety of different validated tools to score the resulting records. This extra step is also conducted because the data obtained came from two different instruments and these are not directly comparable, starting from the fact that the measurement methods and scoring scales are different. Pittsburgh Sleep Quality Index (PSQI) was developed with a reliable, valid and standardized measure of sleep quality in mind. [16] Scoring of the PSQI can be done in components which in put together define the total score for the index. As we were mainly interested in determining the sleep efficiency of each

night we used the Component 4: Habitual Sleep Efficiency which is calculated with the formula (2):

$$\frac{Number\ of\ hours\ slept}{Number\ of\ hours\ spent\ in\ bed} * 100 \tag{2}$$

Data to this calculation is fed from two different sources in parallel to compare the differences of user inputted and device gathered sleep estimations. This is also one metric to determine the reliability of the efficiency / score reported by the device.

Overview of how the scores calculated using this formula compare to the scores given by the devices themselves is in Table 15 as well as visualisation based on this data in Figure 13. From this it can be concluded that the calculation of the efficiency score of the Dreem2 device corresponds very closely to the calculation of the PSQI fourth component, defined in the Pittsburgh study. while only minor relation between it and scores from SleepScore Max can be seen. Thus, SleepScore Max must calculate its scores based on different values than just deriving it from amount sleep and total record time. Manufacturer reports that SleepScore is based on six metrics, which are compared to the person's personal attributes. The parameters are: Total sleep duration, time to fall asleep, light sleep, deep sleep, rapid-eye-movement (REM) and nocturnal awakenings.

	Dreem2 (p1)		*SleepScore Max (p2)*	
Night	**Habitual Sleep Efficiency**	**Sleep Efficiency (Device)**	**Habitual Sleep Efficiency**	**SleepScore (Device)**
1	86,83	87	94,19	95
2	94,03	94	94,97	84
3	94,36	94	97,10	93
4	94,38	94	91,63	89
5	95,05	95	89,81	86
6	92,59	93	95,85	94
7	87,40	88	96,41	94
8	94,35	94	89,28	82
9	98,21	98	94,18	94
10	96,58	97	92,37	95
11	96,38	96	90,83	93
12	91,75	92	89,98	94
13	94,68	95	-	-
14	96,81	97	-	-

Table 15. Habitual sleep efficiencies based on sleep times extracted from the device

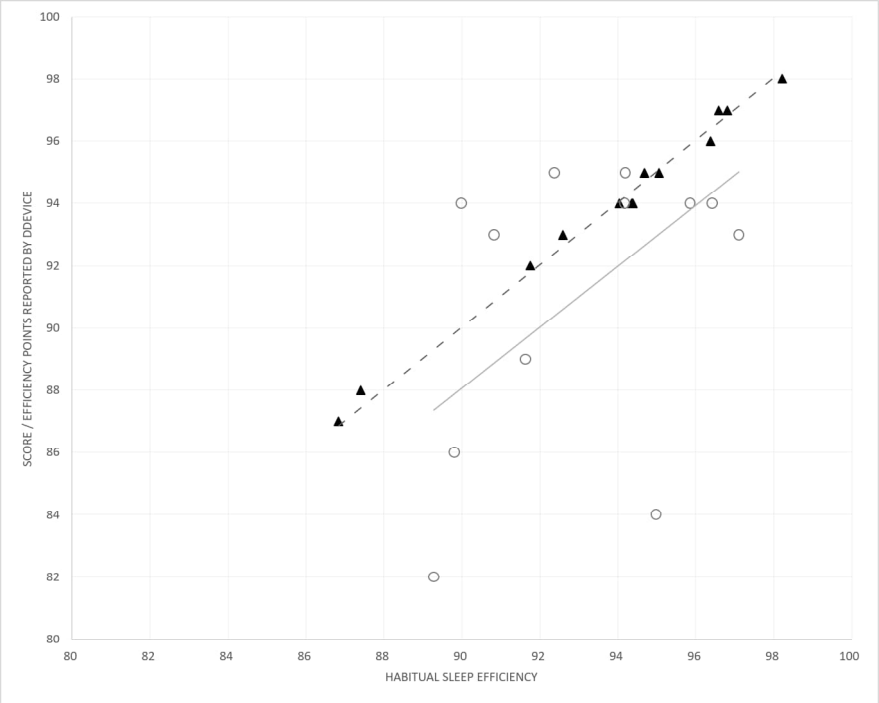

Figure 13. Scatterplot of calculated Habitual Sleep Efficiency vs. Sleep Score. Dreem2: Triangles, SleepScore Max: Rings

Calculating correlations between Habitual Sleep Efficiency scores and ESM scores, the outcome is similar to results of correlation between scores from sleep tracker devices meaning that slight negative correlation is revealed. For these, Pearson's correlation is -0,264 with a p-value of 0,260 while the Spearman's correlation is -0,263 with a p-value of 0,262. Scatter visualization can be found in Figure 14

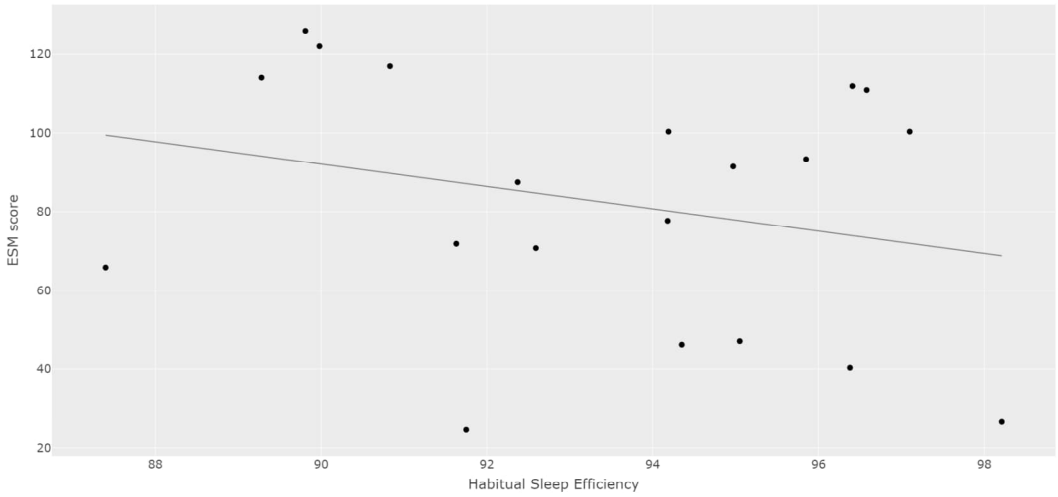

Figure 14. Scatterplot of Habitual Sleep Efficiency vs. ESM scores

6.3.4. ESM Input of Sleep Times

The first questionnaire of the day in the ESM application consisted of questions on sleep diary keeping to gather information of the test subjects recollections on the previous night. These 9am ESM questions inquire the user to estimate the time of going to bed and time of waking up. Example of ESM question asking user what time did they go to bed can be seen in more detail in Figure 4

Sleep diary collected in the study through ESM questionnaires shows, for example, that there are significant differences between the observations made with the hardware and the user reports regarding bedtime and wake-up time. The subjects had reported their bedtimes with a difference of up to more than an hour to what the device had reported as the time of bedtime. Thus, it can be concluded that the data collected from this morning ESM survey is not particularly reliable, especially minding the size of the data set in question. Before mentioned inaccuracies because of small data set combined with gaps means that the main role of this sleep diary ended up being a 'sanity-check,' and to confirm that the device used did perform as expected.

6.3.5. Circumplex of Mood

As mentioned in subsection 4.2.2, one model that we incorporated to the questionnaires themselves was the circumplex of mood (also known as circuplex model of affect). This inclusion was done in the questionnaire opening at 18:00 by first asking the user what they were doing (choices: relaxing, eating, working/studying, exercising) and then asking them to rate it in terms of pleasantness giving rough estimation for values needed for the two axis of the model. As the circuplex of mood is essentially a circle, these values were scaled to be in range of [-1,1] for easy plotting on the model.

While the initial plan was to be able to apply this information on a day level, it was realised that comparing how someone feels between 6pm and 7pm to ESM performance of the entire day is not sensible. Despite this, the data can be used on smaller scale to see how mood affected the one questionnaire in question. To do this, the questionnaire in itself needed to be scored. In practise this was done by considering the time between interacting with the app after notification and time used for answering. For answering time, the same scoring was used as in subsection 6.3.2 while the scoring for interaction can be found in Table 16. The final score was then calculated as the sum of these (3) and visualization of how different scores landed on the model can be found in Figure 15.

$$AnswerTimeScore + InteractionSpeedScore \tag{3}$$

Time between notification and interaction (s)	Score
<600	6
600<t<1200	5
1200<t<1800	4
1800<t<2400	3
2400<t<3000	2
3000<t<3600	1

Table 16. Scoring for notification interaction

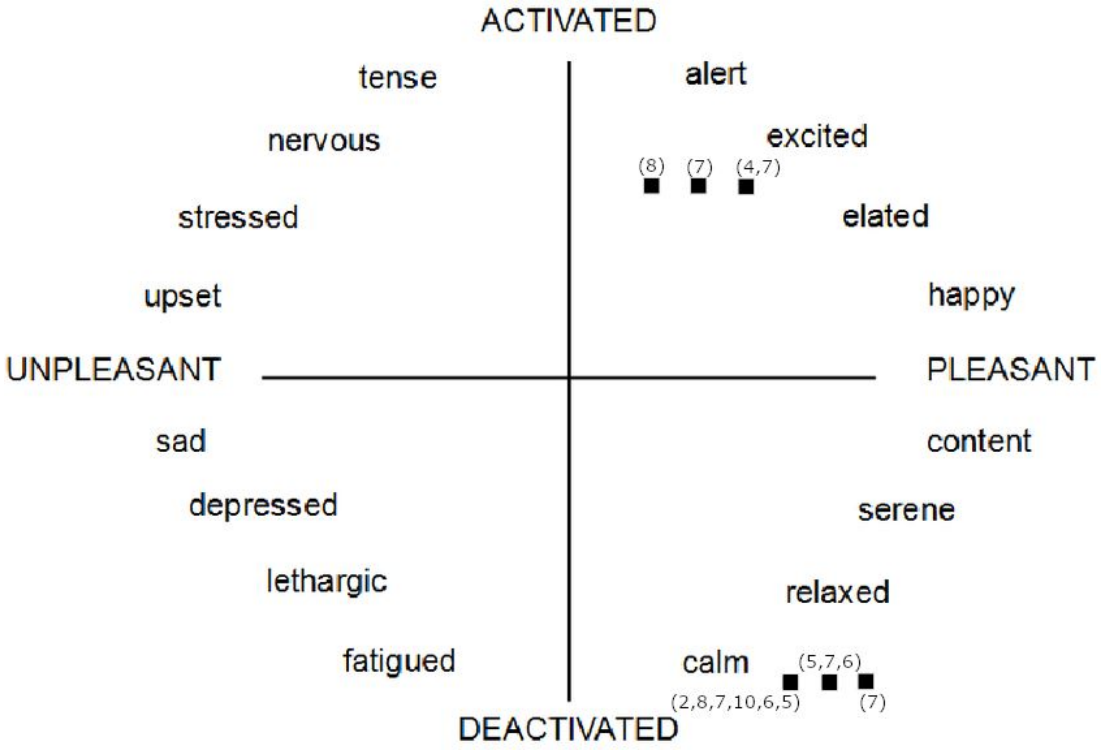

Figure 15. Scores on circuplex of mood

As it can be observed from the visualization, the collected data does not give any major proof on mood having an effect on effort put to ESM questionnaires. The problem here is that the data points essentially produce two groups, pleasant/deactivated and pleasant/fairly activated, with both having scores ranging from low to high. To emphasise this, the minimum and maximum calculated scores (2 and 10) are actually both at the same location on this model. This bad distribution is partly explained due to only having 15 valid responses on this particular questionnaire and in addition, the y-axis only had four options to choose from greatly lowering the resolution in that direction. With more data and better resolution especially on the y-axis, the results would likely be more useful due to significantly increased accuracy.

Another way to emphasise the inaccuracy of this method of analysis in the case of this thesis is that calculating the averages of scores within the two identified groups result in 6,5 for the higher activation and 6,3 for the deactivated group. This further shows that from the data available, there is no way, using the chosen scoring parameters, to draw conclusions on whether or not participants mood might have an effect in ESM questionnaire performance.

6.4. Application Evaluation

During the study, respondents faced no major problems in using the application, however, minor adversities were experienced with e.g., in automatic updates interfering with the deployment of already placed notification events (P1, P2), in this case, operating system update caused notification deployment queue to clear for the following day. Second unanticipated operation occurred with slider question, where the data was not properly cleaned before next launch (P1). These occurrences were recorded and subsequently tackled with on the fly.

On two separate occasions, sleep tracking hardware failed to record the whole night uninterrupted. Both of these were caused by movement during sleep on P1, causing Dreem2 headband to off in the middle of the measurement period. These occurrences lasted approximately 15 minutes and 20 minutes, respectively. For P2, the entirety of last two nights is missing due to the SleepScore Max freezing one night and then their phone restarting during the other one causing tracking to be cut.

Metrics used to evaluate the software focuses on UEQ evaluation, which consists of a short survey conducted after the study phase. This method aims to gather data of the user's perceived experience of using the application and we were interested of whether the application itself influenced the responses of the testers and whether the interface of the application was sufficient to facilitate the creation of reliable data. User Experience Questionnaire (UEQ) was chosen as an evaluation method for in which it should be able to identify nuances that represent how challenging or obstructive the application was to use.

6.4.1. UEQ Assessment

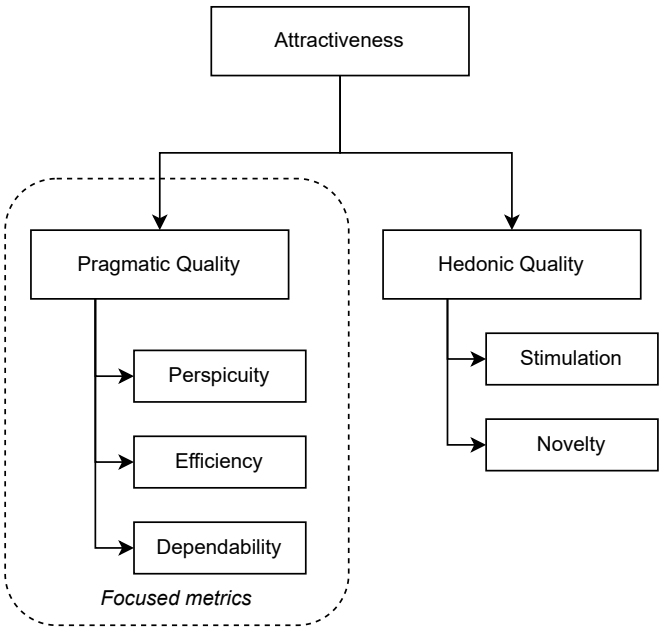

Figure 16. Assumed scale structure of the User Experience Questionnaire (UEQ) [20] with evaluation focus highlighted

UEQ data analysis tool version 10 contains data of a 468 previous evaluations, with a total of 21175 participants combined. Sample sizes ranging from 3 respondents to 1390 respondents. [20] Comparison of results obtained from our application in relation with the benchmarks allows conclusions to be made about the quality of our application compared to other products in the benchmark set. As the benchmark supports extremely small sample sizes, as is the case with this study, it was seen as fitting option, but like other sections, it can be scaled up to fit a larger sampling size if deemed necessary for further research. UEQ divides the measurement types into sic scales. Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation as well as Novelty. [20] As the sample size used is extremely small, in fact way too small to be used as a stand alone mean for drawing conclusions, we wanted to further investigate the applications Pragmatic Quality, containing applications perspicuity, efficiency and dependability. We saw this helping to interpret the data obtained and to further identify potential interface obstructionism. For example the respondent not feeling to be able to answer truthfully, or the most descriptive way.

6.4.2. UEQ Benchmark Results

Based on the benchmark and survey, it can be deducted that the application does meet the basic needed functionalities and the respondents had no major stumbling blocks in the way of submitting their results. However, it should be noted that as this evaluation was carried out only at the end of the study, it does not take into account the initial struggle of use of the application, thus subjects may have adapted to initially obstructive behaviour.

Confidence intervals (p=0.05) per scale						
Scale	Mean	Std. Dev.	N	Confidence	Confidence interval	
Attractiveness	1,167	0,236	2	0,327	0,840	1,493
Perspicuity	**2,375**	**0,177**	**2**	**0,245**	**2,130**	**2,620**
Efficiency	**1,500**	**0,354**	**2**	**0,490**	**1,010**	**1,990**
Dependability	**1,500**	**0,354**	**2**	**0,490**	**1,010**	**1,990**
Stimulation	1,375	0,530	2	0,735	0,640	2,110
Novelty	-0,125	1,237	2	1,715	-1,840	1,590

Table 17. Measure for the precision of scales

Benchmark results fall within the accepted specifications when looking at pragmatic qualities of the application and thus it can be said that the application has achieved its objectives in the regard of Pragmatic Quality. In practice, the values from -0.8 to 0.8 represent a neutral evaluation of the corresponding scale. Values below that range represent negative evaluation and above that positive evaluation. Because of avoidance of extreme answer categories, its extremely unlikely to observe values below -2 or above 2 [21] The scores are as follows:

Scale	Mean	Comparison to benchmark	Interpretation
Attractiveness	1,17	Below average	50% of results better, 25% of results worse
Perspicuity	2,38	**Excellent**	In the range of the 10% best results
Efficiency	1,50	**Above Average**	25% of results better, 50% of results worse
Dependability	1,50	**Good**	10% of results better, 75% of results worse
Stimulation	1,38	Good	10% of results better, 75% of results worse
Novelty	-0,13	Bad	In the range of the 25% worst results

Table 18. ESM application in relation to benchmark data set

As the variation of the selected scales remain proportionally small, and as the predefined purpose of the metric was to examine the consistency of respondents experience with the application, using a well tested benchmark tool. It can be concluded that the subjects were in unison in their evaluation of the application when looking at purely these selected three operational scales. It is to be expected that as the size of the sample increases the scores themselves will undergo a drastic change.

6.4.3. Evaluation of Other Technical Decisions

While the application itself was determined to be reasonably good, the same cannot be said about the implementation of the back end. The main issue was that running a relatively unprotected HTTP server in home network with an open port lead to a situation that the server was attacked causing it to become unresponsive for further incoming requests. The first one occurred before the actual study phase and with a little bit of research regarding the content in the attack request, we determined that the attack in question had to do with Remote Desktop Protocol (RDP)[22]. In short the attack was likely one in which open ports vulnerable to RDP exploitation are attempted to be found. The downtime caused by this attack did not end up having any effects for the actual study as it occurred during the testing phase before starting the study. Having no idea how frequent similar attacks would get over time, the implementation was not changed at this stage despite the fact that it still could have been done with relative ease.

In addition to this pre-study attack, many other types attack were encountered practically on daily basis requiring constant monitoring of the server to restart it whenever it was crashed. Thankfully, the backup system built into the app was functional and thus no responses were lost due to these attacks. These did, however, show that our solution for the back-end was far from optimal and so if our app would ever be used in another study, alternative solutions for back-end should be considered.

7. DISCUSSION

7.1. Reflection - Goals

The main goal of the study was from the beginning to find out if there is correlation between sleep quality and participants performance on a study conducted using experience sampling method. Through the small scale experiment, data needed to determine this was collected, but due to the very small amount of participants as well as short study period, the results obtained did not give any conclusive results. While technically the results obtained would suggest reverse correlation, for the previously mentioned reasons, no proper conclusions can be drawn. In any case, the main goal of this study was achieved as well as possible with the limited tools and time available for completing this thesis.

In addition to the main goal, looking for correlation between the users mood and ESM performance was an fascinating scientific adventure to explore. As this was not the main focus of this study and thus the ability to gather information on this was only embedded into one of the five daily questionnaires, the results in this regard are even more unreliable. On the other hand, like with the main focusing points, the tools for studying this should be usable in future work after minor modifications. With a larger scale experiment, the results could possibly have been more significant.

In the early stages of the project, we also had this larger goal in mind about being able to determine if sleep measurements could be used to improve the overall ESM process. Due to the fact that even the results regarding whether there is correlation between sleep quality and ESM performance in the first place are so questionable, no conclusions regarding this were really possible to draw. Again, if we would have had the ability to conduct a larger scale experiment, the results could have been rather different and in the optimal case, even this question could also have been answered.

While the original goals were not perfectly met, there were some interesting extra findings that can be taken away from the project. One of these is that it looks like peoples memory based reporting of sleep is not very accurate. As mentioned in subsection 6.3.4, there were major differences between sleep data gathered through devices and what users reported through questionnaires. While it is important to recapitulate that our study was extremely small scale and that might be a major factor behind this finding, this might suggest that if further studying of this topic is done, using memory-based system of collecting sleep data might not be ideal.

7.2. Reflection - State of the Art

As there are only a few studies done on the topic of ESM performance/effort in relation to sleep, directly assessing how well our results are in line with other studies is not feasible. As mentioned in subsection 2.1.1 however, it is common knowledge that overall sleep quality does have effect on how how people perform on daily tasks and so comparing our results to this idea, it is very much the opposite. As mentioned, based on our results, it would seem that poor sleep quality yields better performance and so in comparison to previous studies on the topic, this is very contradictory. It needs to be remembered, though, that our small sample size and short study period likely are the

reasons behind this and with a better experiment, the results could be expected to be more in line with those results.

In the chapter 2 we also briefly discussed about how sleep quality is experienced differently by different individuals and thus raw numbers coming from different tracking devices are not necessarily the whole truth. However, in our experiment we did notice that assessing how the individual experience their sleep can be rather difficult. This is because things like start/end times were reported significantly differently between the people and the devices and so when looking for correlation, we had to almost fully ignored the personal experience side of things.

Required quality	Realisation
Scalability	ExSS retrieves the data from centralised location, which with desired edits can be linked to an external source
Editability	Attributes of the questionnaires came from centralised location, changes made can affect both temporal and content aspects.
Ease of use	According to UEQ, participants did not encounter obstacles with the User Interface (UI) or other aspects of using the application.
Understandability	There were no issues in the interpretation of questions.
Communication	Logs arrived on time, although problems were faced with denial of service in the back end
Questionnaire timing	Questionnaires deployed in time
Notification deployment	System updates caused momentary interruption.
Usage monitoring	Information on application usage, opening, answering questions, responding to notifications and other usage statistics were collected successfully, although enough data was not obtained to use the circumplex of mood at its full extent.

Table 19. Fulfilment of the ExSS requirements

7.3. Future Work

Resulted presentation of ESM questionnaire usage for studying sleep quality successfully verified the functionality of the ExSS application under field-conditions (Table 19) and, with a perspective of future research, ensured that the scalability of ESM application is sufficient to facilitate even largest of groups. The main thing regarding future work would actually be that correlation between sleep quality and ESM performance should be studied with larger group of participants and with a longer study period. Then, if positive correlation of the two would be possible to find, it could further be studied if the ESM process can be improved by, for example, adjusting the triggering times of questionnaires based on sleep quality. In practise this likely requires an experiment where different rules for timings would be used on different groups of

people based on sleep data and then try finding if, triggering questionnaires earlier after bad sleep quality would have an positive effect.

These dynamic ESM deployments could achieve greater number of data-points if the questions were deployed at times when the subject is most likely to answer. As demonstrated, technology allows conditional and smart deployments of questionnaires themselves. Importance of this variable deployment should be emphasised and mobile ESM applications should not be afraid to vary the question types to better suit the research topic while also maximising the number of responses.

Questionnaires must be planned well in advance and they should not focus too broad topics, but to rather on distinct, well predefined and scorable questions of main points of the research. For example one question per topic per day cannot guarantee the necessary reliability and in particular, satisfactory amount of data to ensure that the results obtained are adequate to draw meaningful conclusions.

In terms of conducting the ESM questionnaires, the method for storing responses should be reconsidered meaning that if our application would be used in any other study, the current server implementation should not be used. As mentioned in subsection 6.4.3, it was extremely sensitive for external attacks and thus in a larger scale study, where effects of downtime would be even worse than in our case, it would likely just cause a lot of unwanted extra work. As the app implemented uses regular HTTP requests for communication, however, the changing of the back end system should be rather easy.

8. CONCLUSIONS

While sleep is a crucial part of life and performance of different tasks can be significantly effected by a lack of it, no correlation between it and users effort put into ESM questionnaires were found in the study conducted in this thesis. To be more precise, a slight negative correlation between the two was found. While the negative correlation could more easily be interpreted as having no correlation, considering the shortcomings of the study, even that does not align with expected results. All this is likely explained by the fact that the experiment run was extremely small in scale and thus the obtained results do not hold as much value as results collected from a larger scale study.

This study did not reveal any obvious shortcoming in the hardware and the data it provides, but as a basis for future research, it is worth considering whether to use raw data instead of the data exported from the device's dedicated application and if so, to develop own measurement metrics to obtain specific type of information. When using different devices, it is important to ensure that the parameters used are the same or similar and thus be sure that they can be used together as such.

As in any research, the correctness of data handling is important. The challenge that rises is that the subject should never have access to information about their sleep data prematurely, and the diary questions they complete should reflect their mood and opinions before the participants have had a chance to look on their sleep scores from the night before. This is very challenging to ensure working with a proprietary application and in environments where the application is run in test subjects personal device, thus them always having the possibility of viewing their sleep data against the intentions of the authors of the study. We recommend that data should be collected in a different route where possible, similarly with ESM questionnaires.

While the overall results are unwanted in terms of quality, the greatest single outcome of the project is the purpose built ESM application which could be used in further studies regarding this topic. Perhaps in any other study utilizing the ESM method as the questions in the app are easily configurable. The main limitation would be that if other question formats than what were used in this thesis were to be needed, some further implementations would need to be fabricated on the app. Even so, implementing these changes would likely be a lot simpler than building a new application from ground-up.

All in all, the main contribution of this study was to provide a structural basis for future research. Lessons learned can be summarised in three points: ESM questions should be as simple as possible, they should be able to be optimised dynamically over time and one topic question per day is not enough to obtain enough information regarding multiple subject matters.

9. REFERENCES

[1] Hektner J.M., Schmidt J.A. & Csikszentmihalyi M. (2007) Experience sampling method: Measuring the quality of everyday life. Sage.

[2] Van Berkel N., Ferreira D. & Kostakos V. (2017) The experience sampling method on mobile devices. ACM Computing Surveys (CSUR) 50, pp. 1–40.

[3] Cheng S.H., Shih C.C., Lee I.H., Hou Y.W., Chen K.C., Chen K.T., Yang Y.K. & Yang Y.C. (2012) A study on the sleep quality of incoming university students. Psychiatry research 197, pp. 270–274.

[4] Niemi J., Risto R. & Salo S. (2021) Digital sleep: Expert evaluation of commercially available digital sleep trackers .

[5] Huber R., Felice Ghilardi M., Massimini M. & Tononi G. (2004) Local sleep and learning. Nature 430, pp. 78–81.

[6] Carskadon M.A. & Rechtschaffen A. (2011) Monitoring and staging human sleep. Principles and practice of sleep medicine 5, pp. 16–26.

[7] Chokroverty S. et al. (2010) Overview of sleep & sleep disorders. Indian J Med Res 131, pp. 126–140.

[8] Rauchs G., Desgranges B., Foret J. & Eustache F. (2005) The relationships between memory systems and sleep stages. Journal of sleep research 14, pp. 123–140.

[9] Hor H. & Tafti M. (2009) How much sleep do we need? Science 325, pp. 825–826.

[10] Hale L. (2005) Who has time to sleep? Journal of Public Health 27, pp. 205–211.

[11] Bray T. et al. (2014) The javascript object notation (json) data interchange format .

[12] ECMA (2017) Standard ecma-404 the json data interchange syntax .

[13] Berners-Lee T., Fielding R. & Frystyk H. (1996), Hypertext transfer protocol–http/1.0.

[14] Shafranovich Y. (2005) Common format and mime type for comma-separated values (csv) files .

[15] Posner J., Russell J.A. & Peterson B.S. (2005) The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology 17, pp. 715–734.

[16] 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.

[17] Gackenheimer C. & Paul A. (2015) Introduction to React, vol. 52. Springer.

[18] Python webserver. URL: https://github.com/nickjj/webserver. Accessed 25.3.2022.

[19] Gogtay N.J. & Thatte U.M. (2017) Principles of correlation analysis. Journal of the Association of Physicians of India 65, pp. 78–81.

[20] Schrepp M., Hinderks A. & Thomaschewski J. (2014) Applying the user experience questionnaire (ueq) in different evaluation scenarios. In: International Conference of Design, User Experience, and Usability, Springer, pp. 383–392.

[21] Schrepp M. (2015) User experience questionnaire handbook. All you need to know to apply the UEQ successfully in your project .

[22] RDP attack. URL: https://www.cisa.gov/uscert/ncas/current-activity/2018/09/28/IC3-Issues-Alert-RDP-Exploitation. Accessed 28.3.2022.