It's time for some Alexercise

A comparison between reflection capabilities of activity trackers and intelligent personal assistants

by

T.I. Molenaar

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday January 9, 2019 at 11:00 AM.

Student number:	4237005	
Thesis committee:	Dr. C. Hauff,	TU Delft, supervisor
	Dr. A. E. Zaidman,	TU Delft
	Prof. Dr. M. Specht,	TU Delft

This thesis is confidential and cannot be made public until January 9, 2019.

An electronic version of this thesis is available at `https://repository.tudelft.nl/`.

TU Delft

Abstract

Recently published health reports from governments of western countries and the World Health Organization (WHO) provide insights on health benefits of being physically active. Performing physical activity of moderate intensity is, amongst others, associated with a smaller risk on symptoms of depression, cardiovascular disease and diabetes. Despite this information being widely available on the world wide web, one particular form of physical inactivity is still incorporated in the lives of many people: sedentary behaviour. An environment that is especially seductive for this type of physical inactivity is the workplace. Stress, depression, anxiety, and heart diseases are found to be work-related diseases, possibly as a result from sedentary behaviour. As a means of reflecting on physical activity, one could acquire an activity tracker such as FitBit or use one of the many available health monitoring mobile applications. However, it was found that activity trackers may not be very effective for stimulating a physically active lifestyle. In contrast, an emerging phenomenon may be able to do so: the intelligent personal assistant (IPA).

In order to be a part of the solution for aforementioned problem, we aim to stimulate reflection on physical activity during work hours in order to incorporate physical activity into daily routines. Therefore, we carried out a user study with 16 participants in two countries. For carrying out this user study, we introduced *Alexercise*: a system consisting of multiple interacting components, used for collecting, processing and presenting physical activity data. Components in *Alexercise* exchange data in quasi-real time to provide workers with current representative physical activity data. As part of *Alexercise*, we developed a custom skill for Alexa to access physical activity data and present these by voice. In addition, we developed a cross-platform mobile application for tracking and reflecting on activity using the React Native framework. We investigated whether IPAs are more effective in stimulating reflection on physical activity, compared to activity trackers. In addition, we asked participants if they felt the need to take active breaks based on the experience of the user study. At last, we intended to compare results of participants who already incorporated physical activity in their lifestyles with participants who have not.

In the development process, we encountered many challenges, such as the distribution of mobile applications and Echo Dots to participants in two countries, and facilitating real time tracking and reflecting on physical activity. Collected sensor data for Android smartphones was reported with infrequent intervals, and could therefore not be used to reliably assess physical activity. We investigated and elaborately explained how these challenges can be overcome for future work. In contrast to the challenges, participants reported that they perceived talking to Alexa as intuitive, and the majority of the participants reported that they feel the need to take active breaks based on their participation in the user study.

Preface

The most insightful aspect a master thesis has to offer, is that the learning curve never flattens for a single moment. Whereas nine months ago I considered myself relatively knowledgeable about scientific know-how, I have come to question my own ability to know how much I know. Despite the epistemological considerations, I do know that I have learned a lot about when to apply certain scientific methods, how to carry out a user study and how to develop applications for smartphones and intelligent personal assistants. Of course, this learning process could not have taken place if not facilitated by a number of people.

At first, I want to express my gratitude to Claudia Hauff, who was a beacon of rationality in the roller coaster called 'master thesis'. Whenever I managed to derail my thesis, Claudia always got me back on track, providing a solution in real time during our weekly meetings. "You have only one week to implement a complex algorithm? No problem, just tackle these problems and you'll be fine. Eight hundred new students are joining my course this year? No problem, I will just turn it into a user study to find out which learning methods work best. The Faculty of Architecture is on fire? No problem, just build a new one."

Furthermore, I would like to thank Tarmo Robal for both his remote support from Estonia, and his advice during his stay in the Netherlands. If it wasn't for Tarmo, I would have carried out my user study with only half the number of participants. After the Tarmo-era, Dimitrios stepped in, first as a participant and later as professional supervisor. I would also like to acknowledge his support in assessing each written chapter and the resulting useful feedback.

I am also grateful for the support of my colleagues Meindert and Twan at Corn Group, who provided me the necessary Apple devices in time of need, even if it requires cycling for over an hour to deliver a MacBook charger. Of course, I also want to thank my fellow computer science students and friends. Thanks to Wilko for reviewing my thesis on high-speed (literally), to Lars, Sander and Zilla for making all the TU-time bearable and sometimes even fun (especially at the place where we could sometimes 'hangout'), to Chris who has been in sync with me for my entire life and with whom I could enjoy conversations about the process of graduating (and much more), and to Maria, who was literally in the same boat as I was, and to whom I could relate all this time. Last but not least, thanks to Arine who provided feedback on my work as icing on the cake.

Writing a master thesis is also a good excuse for thanking my family and girlfriend. Thank you Laura, for your love and support at all times. I'd probably lie under a bridge somewhere looking for my motivation if it wasn't for you. I'd also like to thank my parents for the mental support during my thesis, but foremost for the first eighteen years of my life that prepared me for a life at the university of Delft. My gratitude also goes out to my Dalton brothers Jelle, Nils and Erik for the moments of joy (especially at Junk Junction).

Concluding, I am thankful that I have so many people to be thankful for. The experience of this master thesis at the TU Delft will not be forgotten.

T.I. Molenaar
Delft, January 2019

Contents

1 Introduction **1**
- 1.1 Stimulating moderate intensity physical activity . . . 2
- 1.2 Intelligent personal assistants. . . . 3
- 1.3 Research statement . . . 4
- 1.4 Approach . . . 5
- 1.5 Contributions . . . 5
- 1.6 Outline . . . 5

2 Related Work **7**
- 2.1 Detecting physical activity . . . 7
- 2.2 Stimulating physical activity . . . 9
- 2.3 Well-being assessment . . . 10
- 2.4 Comparison & user experience of IPAs . . . 12
- 2.5 Enhancing the experience of reflecting on physical activity . . . 13

3 Methodology **15**
- 3.1 Choice of IPA . . . 15
- 3.2 Tracking activity . . . 16
- 3.3 User study . . . 18
- 3.4 Evaluation process . . . 20
  - 3.4.1 Quality inspection of the data . . . 20
  - 3.4.2 Assessment of physical activity. . . . 20
  - 3.4.3 Correctness verification of measured physical activity . . . 20

4 Alexercise **23**
- 4.1 Back-end . . . 23
  - 4.1.1 Authorisation . . . 24
  - 4.1.2 Data processing . . . 24
  - 4.1.3 System Monitoring. . . . 25
- 4.2 DeskstApp . . . 25
  - 4.2.1 User identification . . . 26
  - 4.2.2 Data collection. . . . 26
  - 4.2.3 Presentation of physical activity . . . 27
- 4.3 Alexa . . . 29
  - 4.3.1 Syntax . . . 29
  - 4.3.2 My Coach . . . 30
  - 4.3.3 Architecture . . . 30
  - 4.3.4 Use case: request remaining steps . . . 31
  - 4.3.5 User identification . . . 31
  - 4.3.6 Notifications . . . 31
- 4.4 Challenges . . . 33
  - 4.4.1 Ejecting from Expo . . . 33
  - 4.4.2 Strict Apple policies . . . 33
  - 4.4.3 Distribution of My Coach . . . 33
  - 4.4.4 Installing the Echo Dots . . . 34

5 User Study **35**
- 5.1 Data quality assessment . . . 35
  - 5.1.1 Sensor equality. . . . 35
  - 5.1.2 Completeness . . . 37

5.2 Interactions with Alexa . . . 39
5.3 Comparison of performed physical activity between phases . . . 40
5.3.1 Evolution of self-set goals over time . . . 40
5.3.2 Physical activity per phase . . . 41
5.4 Evaluation of the research statement . . . 42
5.4.1 Tasks for participants . . . 42
5.4.2 Feedback from participants . . . 43
5.5 Challenges . . . 43
5.5.1 Infrequent sample rate . . . 43

6 Discussions & Conclusion **45**
6.1 Discussion . . . 45
6.2 Limitations . . . 46
6.2.1 Technical improvements . . . 46
6.3 Threats to validity . . . 46
6.3.1 Selection of participants . . . 47
6.3.2 The novelty of Alexa . . . 47
6.3.3 Real-world situation versus controlled setting . . . 47
6.3.4 Notifications . . . 47
6.4 Future work . . . 47
6.4.1 Improving data collection . . . 47
6.4.2 Evaluating the goal setting theory . . . 48
6.4.3 Sensing well-being . . . 48
6.4.4 Context-awareness and interruptibility . . . 48
6.4.5 Privacy . . . 49
6.4.6 Future applications of IPAs . . . 49
6.5 Conclusion . . . 49

A User study and System design & validation **51**
A.1 User Study . . . 51
A.2 System design & validation . . . 52

List of Figures **55**

List of Tables **57**

Bibliography **59**

1

Introduction

Two thousand years ago, the Roman poet Juvenal wrote down a modern ideal: "Mens sana in corpore sano" - a healthy mind in a healthy body. While we live in a time where knowledge about our health is accessible all over the world wide web, statistics about our health tell us there is still a lot to gain by incorporating ancient wisdom in our daily lives. In the Netherlands, the country where this study takes place, only 44% of all adults perform the recommended amount of physical activity [46]. Unfortunately, physical inactivity is related to a variety of diseases. It is estimated to be the cause of 6% of all coronary heart diseases world wide, and the total average life expectancy is estimated to increase with more than eight months for the world population, compared to the current situation if physical inactivity would be eliminated [36]. Besides, the Health Council (HC) of the Netherlands categorised studies reporting the beneficial effects of physical activity and sedentary behaviour on health [46]. The effect of physical activity on a certain disease is considered *convincing* if outcomes from multiple randomised controlled intervention studies (RCTs) and cohort studies support each other. If an effect is only measured by cohort studies, the association is considered *plausible*. Convincing evidence of gains from physical activity includes a reduced risk on depressive symptoms, cardiovascular disease and diabetes, as depicted in figure 1.1. In this context, physical activity is defined as "any bodily movement produced by skeletal muscles that results in energy expenditure" [4].

Sedentary behaviour, in particular, is a form of physical inactivity that is widely integrated in western lifestyle [42]. While scientific evidence for detrimental effects of sedentary behaviour is not as strong as for physical inactivity, it appears to have adverse health effects [46]. In a study with 3625 participants wearing accelerometers it was found that replacing two minutes of sedentary time with an activity of low intensity seemed to have no significant effect on premature mortality [2]. However, replacing two minutes of sitting per hour with an activity of light or moderate intensity (see table 1.1) was associated with a lower hazard of death. The Dutch HC supports this by showing that being at least 75 minutes per week moderately active results in a decreased risk on heart attacks and heart failure, which further decreases as the amount of physical activity goes up to 150 and 300 minutes.

In addition, regularly breaking up sedentary time increases a person's calorie expenditure and potentially reduces the chance on obesity [68]. To specify this further: the total time of physical activity during the day should ideally be divided in multiple shorter activity breaks instead of a single long activity break, as more interruptions in sedentary behaviour is positively associated with a reduction on metabolic risk [28].

The workplace is one of the places where sedentary behaviour is fully present. From a study carried out in the U.S., it was found that participants spent 7.7 hours a day in sedentary position [42], an indication that the office is an environment where a lot is to be gained in terms of breaking up sedentary behaviour. According to the health and safety report of the EU, 14.5% of its workers suffering from work-related diseases reported they suffer from stress, depression or anxiety. In addition, 3.8% of the working European citizens with work-related health problems suffer from hearth diseases. These are diseases that are convincingly related to physical inactivity [19]. The World Health Organization (WHO) recognises health problems in the workplace, such as long working hours, that could result in physical inactivity [48]. Proposed solutions focus on facilitating a more healthy lifestyle by encouraging walking and cycling and allowing flexibility for workers in taking breaks from their work.

Health related problems in the workplace resulted in proposed solutions of the WHO and specific guidelines of the Dutch HC which inspired us to find a specific approach for encouraging physical activity in the

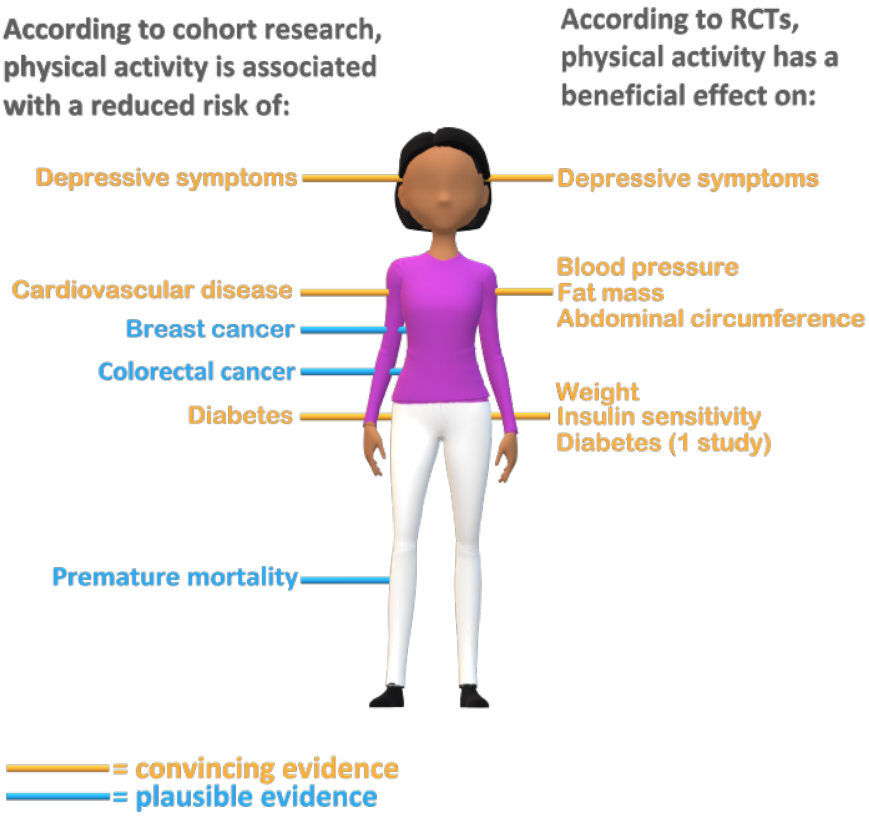


Figure 1.1: Health effects of physical (in)activity for adults [46]

workplace. To prevent ourselves from reinventing the wheel, we use techniques that have proven to be effective in prior work. The scientific novelty of our study is the use of an intelligent personal assistant (IPA) as a means of stimulation of physical activity. How we aim to stimulate workers to be physically active and what's the role of an IPA are respectively discussed in the sections 1.1 and 1.2.

1.1. Stimulating moderate intensity physical activity

Given the benefits of sufficient physical activity, this study aims to disrupt chronic sedentary behaviour by both measuring and stimulating physical activity. Physical activity should be at least of moderate intensity, as it has convincing beneficial health effects. Thus from here on, physical activity refers to moderate intensity physical activity. Due to aforementioned disturbing statistics from Eurostat about work-related health issues and the provided guidelines by the WHO, we decided that our target group is especially men and women working in the office (workers) that are often performing work while being physically inactive. The effects of physical inactivity have different outcomes for men and women. For men, there are indications that physical inactivity at work encourages a passive lifestyle for them [24]. For women, no association was observed. However, since this is a single study indicating this difference and given the health benefits of physical activity during working hours, we include both men and women in our study. Besides the fact that physical activity reduces the chance on several diseases, it can also support mental well-being by increasing vigour and reduce emotional exhaustion [58]. Because mental well-being is often of importance in the workplace besides being physically in shape, we dedicated a section to the assessment of well-being in the next chapter, where decisions for using certain techniques are explained more elaborately.

In supporting workers to be physically active, the first step would be **(1) measuring** their physical activity for them to get insight in their movement. As discussed in section 2.1, prior work often used a sensor-based approach for measuring physical activity. A follow-up action is to use this physical activity data to **(2) advice and stimulate** people to reflect on how much they should exercise. For that, prior work used gamification techniques [3, 77], quantification of physical activity [77] and other, more original approaches [10]. Zucker-

Table 1.1: Categories of physical activity intensity as defined in [2] and [46]. MET stands for metabolic equivalents, where one MET is the energy expenditure at rest

Activity	**As defined in [2]**	**As defined in [46]**	**Energy expenditure**
Sedentary	Sitting quietly, watching television, lying down	Watching television, working with a computer	1.0–1.3 METs [2] < 1.5 METs [46]
Low intensity	Sitting in class, studying, note taking, standing, making bed		
Light intensity	Casual walking, light gardening, cleaning (sink/toilet)	Playing music, washing up	1.5-3.0 METs
Moderate intensity	Brisk walking or running, lifting heavy weights	Walking, cycling	3.0-6.0 METs
Vigorous intensity		Running, playing football	> 6.0 METs

man and Gal-Oz [77] challenge the fact that using gamification techniques to support people to be physically active is more effective than quantification of physical activity data, and they vouch for facilitating "(...) reflection on meaningful aspects of physical activity by developing novel ubiquitous measures". Presenting physical activity data in a quantified way is in line with the Quantified Self movement[1] which stimulates people to get to know themselves through numbers to spark intrinsic motivations. On the contrary, the use of rewards as gamification technique encourages actions based on extrinsic motivations. According to Ploderer et al. [52], intrinsic motivation is the desired form of motivation, as "ideally, reflecting about certain aspects of one's life provides valuable insights, which in turn lead individuals to reconsider and possibly change particular attitudes or behaviours". They distinguish two types of reflection: (1) reflection-in-action and (2) reflection-on-action. The first type includes providing means to reflect on an action that is currently taking place (e.g. wearing a FitBit), and the latter is the action of reflecting on an action that is already finished.

We consider physical activity during the workday an 'action'. As long as a participant is at work, he is 'in-action'. Therefore, the focus for our user study is foremost on reflection-in-action. An example of a study applying reflection-on-action is carried out by Finkelstein et al. [22]. Eight hundred participants wore activity trackers for the duration of one year, where half of them were stimulated by external incentives for the first six months. In general, external cash incentives caused participants to be physically more active compared to the group without incentives. However, when the incentives were removed, it was found that wearing activity trackers is little effective for significantly improving physical activity levels for the long term. Section 2.2 elaborates on the details of their experiment.

1.2. Intelligent personal assistants

From Finkelstein et al. [22] we can derive that activity trackers are not yet very effective for the long-term, certainly not without external incentives. A small qualitative study in the workplace partially supports this statement, as it was confirmed by self-reports from participants that some found activity trackers useful, but all of them merely used them for a short period of time [40]. A relatively novel phenomenon might be able to support activity trackers in the presentation of physical activity data: the intelligent personal assistant.

An intelligent personal assistant (IPA), also known as digital personal assistant (PDA) or virtual assistant (VA), is "(...) a metalayer of intelligence that sits on top of other services and applications and performs actions using these services and applications to fulfil the user's intent" [62]. The 'home' of an IPA could be a smartphone, tablet, laptop, Desktop or even a headless device such as the Amazon Echo Dot[2]. A digital personal assistant is capable of conversing via techniques such as speech recognition, language understanding (LU), converting text-to-speech (TTS) and language generation (LG). Upon grasping the user's intention, an IPA should act accordingly, serving the user with the requested information via machine learning and data mining techniques. According to Sarikaya [62], IPAs are useful because they are good at stitching together multiple formerly separated tasks, such as accessing a user's calendar, reading and sending their e-mail, mak-

[1] http://quantifiedself.com/
[2] https://www.amazon.com/dp/B01DFKC2SO

ing phone calls and playing music. In this sense, a virtual assistant is also referred to as 'personal', because it (ideally) is aware of the user's interests, whereabouts and contacts. To expand the knowledge of an IPA, one could use smartphone sensors such as motion, physiological and contextual sensors, as explained in section 2.1.

An IPA can either be *proactive*, performing an action without any preceding action from the user, or *reactive*, responding to a preceding action of the user (e.g. a voice command). A proactive assistant that is aware of the user's performed physical activity could potentially advice a user on their physical activity schedule. In our study, we will make use of this feature to support participants in breaking up their sedentary time and reach their physical activity goals. For that, we will use a headless device from Amazon: the Echo Dot, home of Alexa, the personification of their IPA. Alexa has the promising feature to allow custom functionality to be added, similar to apps on a smartphone. To our knowledge, Alexa nor any IPA was ever used to provide tailored advice on physical activity to prevent chronic sedentary behaviour, while it has the capability to access and combine motion, physiological and contextual metrics to infer the user's physical activity level and possibly facilitate social support [52]. This brings us to our research statement.

1.3. Research statement

Using an Amazon Echo dot, we try to motivate people who are often in a sedentary state for subsequent long periods of time to integrate physical exercise of moderate intensity into their daily routines in order to improve their physical well-being. As discussed in section 1.1, prior work indicated that activity trackers do not have the desired effect of improving health outcomes. Therefore, we want to investigate whether the use of IPAs result in a significant increase in physical activity compared to activity trackers. Two important building blocks greatly supported our experimental setup and applied methodologies. At first, we used the work of Zuckerman and Gal-Oz [77] who inspired us to focus on the incorporation of physical activity in daily routines for the long-term using techniques promoted by the quantified self movement, which indicated to be effective stimulants. In addition, prior to our work Cambo et al. [3] carried out a study to actively stimulate workers to take breaks via distributing tasks through a custom-made mobile application. Their experimental setup supported us in shaping our experimental design. Taking into account that:

- Physical activity of moderate intensity reduces the chance on certain non-communicable diseases
- Activity trackers are not always effective in improving health outcomes
- Reflecting on activity may spark intrinsic motivation to be more physically active
- Intelligent personal assistants have not yet been applied for reflecting on physical activity, while are suitable for collecting physical activity information and providing personal information

We come to the following research questions:

RQ1 *To what extend is the taken № of steps of workers during a workday affected by proactive intelligent personal assistants compared to passive activity trackers?*

Section 2.1 indicates that activity trackers and especially accelerometers [13] are often used to detect and stimulate activity. IPAs should be compared to this current standard method of promoting physical activity to see whether it affects physical activity levels in a meaningful way

RQ2 *To what extend do intelligent personal assistants and activity trackers stimulate participants to integrate moderate physical exercise into their daily routines?*

An important aspect is to stimulate long-term integration of physical activity in daily routines at work, such that participants will not fall back to old habits. We rely on qualitative methods for answering this question.

RQ3 *What is the difference in the amount of performed physical activity for workers who already exercise outside working hours and those who do not?*

As discussed in section 1.1, passive jobs could encourage workers (men in particular) to have a more passive lifestyle. In contrast, participants with an active lifestyle may perform more physical exercise during working hours compared to other participants. On the contrary, it is possible that participants mainly perform exercise outside working hours without feeling the need for taking breaks from work. Thus, this question addresses possible relevant differences in the results.

And the following hypotheses:

H1 *Most people choose not to use Alexa frequently*
Cowan et al. [12] reported that most IPA users choose to only use it occasionally or not at all.

H2 *The use of activity trackers will not lead to a significant increase in steps*
As reported by Finkelstein et al. [22], discussed in section 2.2

1.4. Approach

In order to find evidence for answering our research questions, 16 workers were asked to participate in a user study, using Alexercise for three subsequent weeks. During the user study, we first determined the physical activity baseline of participants. Then, in the second week of the experiment half of the participants were given a mobile application reflecting their number of steps (activity tracker) and the other half were given an Amazon Echo Dot (personal assistant). In the third week, tasks between the two groups were exchanged. Movement was measured by accelerometer and GPS sensors on participants' smartphones.

In the development process of Alexercise, we stumbled on a number of challenges, especially related to the mobile application collecting physical activity data. These challenges resulted in noisy physical activity data, from which one could not draw quantitatively significant conclusions concerning an increase in physical activity. However, we observed interactions between participants and Alexa, and we elaborated on self-reports of participants regarding IPAs and their experiences with Alexercise. In addition, we described the technical limitations in details and provided recommendations for future work to overcome these limitations.

1.5. Contributions

Our study adds several contributions to the fields of HCI, in particular that of the interaction between humans and intelligent personal assistants:

- We have build a strong case for reducing physical inactivity and chronic sedentary behaviour and analysed, combined and discussed insights from prior applications with intelligent personal assistants and activity trackers.
- We created a system consisting of a mobile application, multiple web services and an IPA using modern technologies, of which all components are open-source. To our knowledge, this is the first application that aims to improve physical well-being using an intelligent personal assistant.
- We reported challenges that we faced during the design process and user study, due to the fact that we were on an exploratory mission on several areas, especially that of IPAs. These might ease software development and creating an experimental design for future work.
- We discussed the participants' perception of our system and especially how they perceived the intelligent personal assistant in the current setting, and their attitude towards it.

1.6. Outline

In the next chapter, we discuss related literature in more detail and provide background for understanding the material used in the user study. In chapter 3, we explain the rationale behind decisions we made, related to the system and experimental design. Then, in chapter 4 we elaborate on each of the system's components in detail. Chapter 5 is about the results from the user study that we carried out. Finally, we discuss the applied methods, system design, results from the user study and the limitations of each of these aspects, and set out a direction for future work.

2

Related Work

This study operates in the intersection of the areas physical and mental well-being and human-computer interaction (HCI). A number of studies that were carried out in these research areas will serve as foundation for this study and shall be discussed in this chapter. At first, for our study it is essential to be able to apply techniques on detection and stimulation of physical activity. Such techniques that have been investigated in prior work will be discussed in the sections 2.1 and 2.2. Stimulating physical activity is not necessarily a goal in itself. An underlying goal could be to improve a person their well-being, physically like reducing the chance on heart and vascular disease, and mentally like decreasing likeliness to end up in a depression. However, the relation between physical activity and prevention of mental illness is disputed by some, because it can be argued that the working of physiological and neurological mechanisms are not yet clear [67]. In order to see whether physical activity has the desired effect of improving mental well-being, a way to assess well-being and expressing it in terms of some objective metric(s) is required. Since well-being is an abstract concept that is hard to measure at once, it should be assessed on its more concrete aspects. Where the first two sections especially focus on the physical side, section 2.3 elaborates on and compares studies that concentrate on aspects of mental well-being, such as the assessment of cognitive focus, attentiveness and mental fatigue. In the last section (2.4), a particular area of HCI is discussed: intelligent personal assistants (IPAs). Studies that were carried out on the subject of IPAs especially focused on its user experience and comparison of different types and brands of personal assistants, as can be observed in the last section.

2.1. Detecting physical activity

There are two ways of assessing a person their physical activity levels:

1. Subjective measurements, such as self-reports
2. Objective measurements, such as data reported by sensors

Data resulting from a study using the latter approach can differ significantly from results fetched from self-reports [54]. In the process of detecting physical activity, one often decides to go with a sensor-based approach [3, 10, 56, 60, 70]. Sensors used for measuring physical activity can be divided into the categories movement, physiological, and contextual sensors [7]. Examples of **movement sensors** are pedometers (counts steps), gyroscopes (determines orientation) and accelerometers (measures velocity in one (1-axis) or several (n-axis) directions). **Physiological sensors** measure physical characteristics of the human body, for which examples are heart rate, blood pressure and temperature sensors. **Contextual sensors** are especially concerned with the environment where a physical activity takes place. An example of such is the Global Positioning System (GPS). Movement and contextual sensors are often available in today's smartphones [53], by Android also referred to as motion, environment and position sensors[1]. According to Deloitte [14], in 2017 93% of the Dutch people between 18-75 owns a smartphone, where the group of people in the range of 25-34 years stands out with a smartphone coverage of 97%. More than 98% of the people in the Netherlands between 12 and 45 in 2017 have indicated that they have access to the internet at home with their smartphone and on average (seniors included) this is 89% [5]. The fact that smartphones are highly available, are often able

[1] `https://developer.android.com/guide/topics/sensors/sensors_overview`

to connect to the internet and the fact that they are equipped with movement and contextual sensors make smartphones suitable devices for real time physical activity tracking.

In the category of movement sensors, especially accelerometers are often used for monitoring physical activity [7], for example in Beddhu et al. [2]. An accelerometer describes the change in velocity with respect to time. In physical notation this is either in metres per second squared (m/s^2) or gravitational acceleration units (g, where $1g = 9.8m/s^2$) [6]. In general, accelerometers are small, affordable and able to measure along multiple axes [7]. In addition, they have proven to be relatively accurate: in a study carried out by Storm et al. [66], it was found that for seven activity trackers using 3-axis accelerometers, the accuracy varied from 64.6% till 99.1% accuracy in terms of steps taken. Accelerometers are used in various ways. Activity trackers might consist of a single accelerometer [77], multiple accelerometers [60] and one or more accelerometers in combination with other sensors [34]. In addition, accelerometers are suitable for counting a person their number of steps. A system that is able to count steps based on accelerometer-data is used by Khedr and El-Sheimy [34], which is simplified as follows:

1. A person's acceleration magnitude is computed using a 3-axis accelerometer, where the magnitude is the square-root of the sum of the axes of the accelerometer squared: $mag = \sqrt{x^2 + y^2 + z^2}$

2. As a person is walking, the magnitude peaks (when they are in the middle of a step) and falls (when the step is completed), resulting in a sinusoidal pattern. A step is represented by a peak and a valley, as is visualised in figure 2.1. The function fluctuates around a magnitude of ten, which means there is on average an acceleration of approximately ten, which is close to the gravitational constant [69].

3. Various filters are applied to peak/valley pairs to improve accuracy, such as comparing accelerometer-data with magnetometer and gyroscope data, filtering noise from the signal by applying adaptive filters and setting thresholds for peaks and valleys.

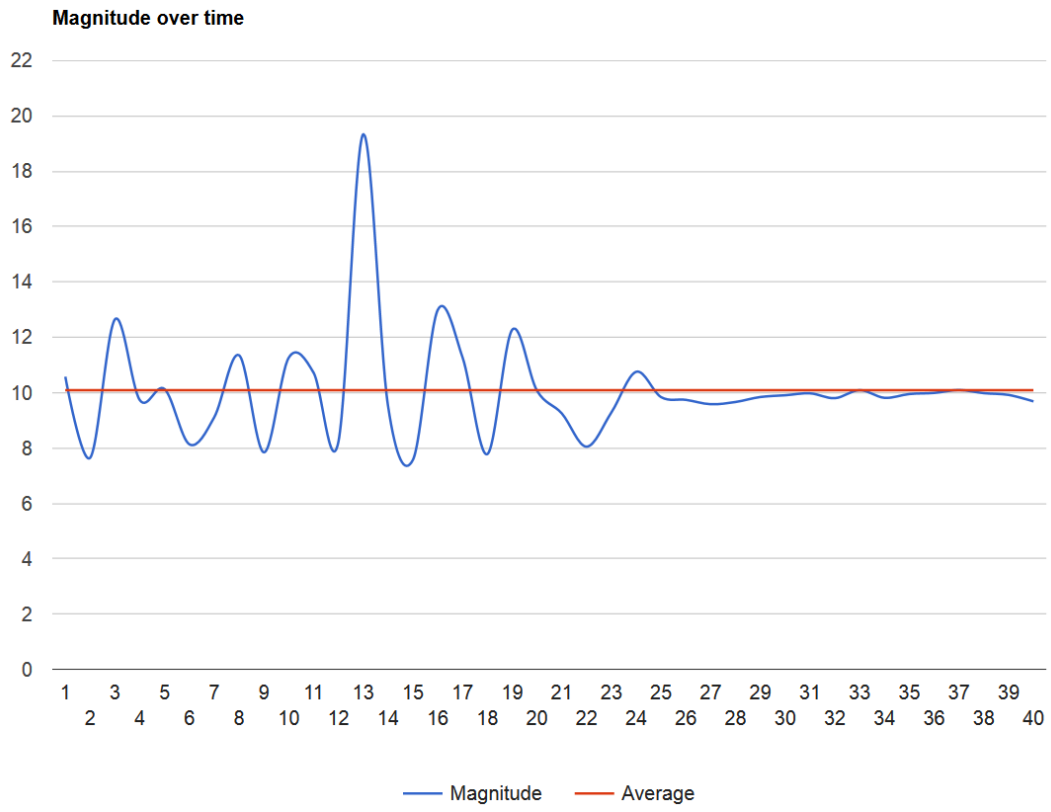


Figure 2.1: 40 subsequent recordings of magnitude show a sinusoidal pattern in the first half, which indicates the person is walking. From recording 25 and on, the person does not move, because the line is flattened and approaches the gravitational constant. Chart is based on figure 2 & 3 in Khedr and El-Sheimy [34], reproduced by DeskstApp, the application that monitors physical activity of participants as discussed in chapter 4

Counting steps is an approach for simplifying the more complex concept 'physical activity', by quantifying a person their physical activity and reducing it a single number. This quantification leads to the possibility to make a person reflect on their activity, which will be discussed in the next section. A sensor - or composition of sensors - that is determined to count steps, is also referred to as a pedometer. An (electronic) type of pedometer that is often used consists of a "horizontal, spring-suspended lever arm that deflects with vertical acceleration" [70]. Richardson et al. [59] collected data from several studies where pedometers were used as motivational tool during walking programs, which were found to result in a modest amount of weight loss. However, the downsides of accumulated step pedometers are that they are generally unable to assess intensity, frequency and duration of activities [11]. However, this can partially be overcome by providing time stamps for each counted step, such that steps can be placed on a time line to determine in which time frame a person took most steps. This concept is for example applied by Google Fit [2].

In addition to aforementioned movement, physiological and location sensors, one could decide to make use of Bluetooth Low Energy (BLE) beacons, a small battery-powered device [72], to detect a person's movement. BLE beacons are especially suitable for indoor localisation, which supplements GPS in that sense, because GPS technology is not always able to provide qualitatively good results in indoor environments [29]. BLE beacons consume, as is claimed[3] and proven [64], very little energy. There are several localisation techniques, such as spatial partitioning, where the room is divided into a number of hexagonal unit spaces and a beacon is placed at each corner and one in the centre. The hexagonal shape results in the largest possible space with the least materials required [29]. A disadvantage of this approach is that all coordinates within the hexagonal are relative coordinates. Another technique to determine a persons position is using Received Signal Strength Indication (RSSI), for which the wireless receiver determines the signal strength measured in decibel-milliwatt (dBm). Both RSSI and spatial positioning make use of unidirectional messaging, where the beacons are broadcasting to a wireless receiver. But what if the wireless receiver wants to send a message to the BLE system? This can be done by sending a message from the application on the wireless receiver (e.g. Launch Here for iPhone[4]) to the back-end via the Wi-Fi network. Lin and Lin [37] bypassed the need for a Wi-Fi network for communication from wireless receivers (e.g. smartphones) to Bluetooth beacons by setting up bidirectional communication via the advertising channels.

In the next section an application is discussed for which Bluetooth beacons are used to determine the indoor position of workers to calculate their movement during a day in the office. In addition, various ways of stimulating people to perform sufficient physical activity will be discussed.

2.2. Stimulating physical activity

A number of studies attempted to stimulate people to be more physically active. In a study carried out by Finkelstein et al. [22], the effectiveness of existing activity trackers was measured. In a randomised control trial, 800 participants were well-nigh equally divided over four groups as depicted in table 2.1. The randomised control trial took six months and then for another six months, the incentives were removed from group two and three and the long-term effects of each group were measured. During the first six months, group two clearly showed an increase in physical activity compared to other groups. However, at the end of the follow-up period without cash incentives, the significant increase compared to other groups disappeared.

Table 2.1: Four groups of participants as in [22]

Group	**Participants were ...**
1. *Activity tracker*	Wearing activity trackers without any extra incentives
2. *Activity tracker + cash incentive*	Wearing activity trackers and stimulating people with cash incentives of S$ 15 (Singapore Dollar) for 50000-70000 steps, or S$ 30 for > 70000 steps per week
3. *Activity tracker + charity incentive*	Wearing activity trackers and stimulating people with donations to charity for the same amounts as for the cash incentive group
4. *Control*	Not wearing any activity trackers nor stimulated in any way

[2] https://www.google.com/fit/
[3] https://www.bluetooth.com/bluetooth-technology/radio-versions
[4] http://launchhere.awwapps.com/

In addition, the FitBit wear dropped quite drastically for group one, two and three after the sixth month. Concluding, there is some evidence that wearing activity trackers improves step activity, however group one did not show an increase in steps and no improvement in health outcomes were detected for neither of the groups. The latter, however, could be explained by the definition of improvement in health outcomes. Finkelstein et al. [22] used ActiGraph[5] accelerometers to measure the intensity of the activity, which can be classified as light, moderate or vigorous[6]. They focused on bouts of ten minutes with at least moderate-to-vigorous activity (MVPA), because a bout less than ten minutes was not considered sufficient for improving health outcomes, according to the United States Centers for Disease and Control Prevention (2008)[7]. However, in their new edition (2018) of physical activity guidelines, the ten minute condition is eliminated, as any length of bouts was found to contribute to health benefits [45].

In Cambo et al. [3], a study was carried out in order to measure the effects of promoting mobility in the workplace. The workplace was equipped with a Bluetooth beacon infrastructure. Workers (N=6) received a smartphone and smart watch, which, together with the Bluetooth beacons, captured physiological and location data. In addition, BreakSense, a mobile application, was installed on the phones that were handed out to the participants. The worker was then given a short challenge through notifications on their phone, as soon as they would leave their work spot. The challenge included collecting stars somewhere in the building, in order to achieve the necessary physical mobility.

Another application that aimed to improve people's physical activity levels is UbiFit Garden: a fitness device, interactive application and glance-able display [10]. For three weeks, it tracked the movement of participants (N=12) by means of a 3-axis accelerometer and barometer. A list of daily activities and progress towards the daily goal is displayed to the user by the interactive application. The display visualises a garden that blooms as the user performs more activity. This glance-able display was well-perceived by the participants.

In both of these studies, gamification elements were used to stimulate the user to be more physically active. The underlying assumption could be that gamification is believed to be effective. However, a question that needs to be answered beforehand is: does gamification lead to more (active) breaks? Zuckerman and Gal-Oz [77] investigated the effect of virtual rewards and social comparison on break-taking and daily activity. They built an accelerometer-based mobile application that was used for promoting physical activity, called StepByStep. The authors intended to integrate moderate physical activity into participants' daily routine in a two week period, using two different versions of the app:

v1. A quantified version, reflecting the user's activity trough numbers
v2. A quantified, virtual reward and social comparison version, where virtual rewards consisted of virtual points and social comparison consisted of a leader board, visualising the points of each of the participants.

In v1, StepByStep monitored the physical activity of participants (N=30) for three days to set a baseline for each participant. After three days, a goal was set for each participant, reflecting a 10% increase of the baseline. Participants were presented with their own physical activity statistics, such that they were aware of their progress. In v2, gamification techniques were used to stimulate participants (N=59), such as giving points as reward, or showing a leader board reflecting a comparison of physical activity metrics of participants. Results for v1 indicated that a quantified version of the application increases activity levels compared to the baseline levels. In the second study it was found that gamification and social comparison are as effective as a system that quantifies a person's daily activity.

2.3. Well-being assessment

Systems that promote physical activity may rely on the idea that a persons well-being is improved if they perform sufficient exercise. In order to see whether physical activity has this desired effect, well-being has to be assessed. Well-being tells us something about the 'state' of a particular person, expressed in a gradation of 'goodness'. As this is an abstract concept, Veenhoven [71] distinguished two types of well-being: (1) chances on a good life and (2) the actual outcomes (or results) of life. Both chances and outcomes can tell us something about (3) external and (4) internal states of being, which corresponds respectively to well-being in an environment and well-being as an individual. These distinctions are visualised in table 2.2.

[5] https://www.actigraphcorp.com/
[6] https://actigraph.desk.com/customer/portal/articles/2515802
[7] https://actigraph.desk.com/customer/portal/articles/2515834

Table 2.2: Varieties of well-being [71]

	(3) Outer qualities	*(4) Inner qualities*
(1) Life-chances	Living in a good environment	Being able to cope with life
(2) Life-results	Being of worth for the world	Enjoying life

Variant (1,3) is about good living conditions in an environment, which can also be referred to as 'welfare', typically a type of well-being a politician is concerned with. (1,4) is about the ability of a person to deal with problems and difficulties in life, something that concerns a psychologist. Other names for this variant are 'health' or 'fitness'. Aspect (2,3) is about doing good for your environment, which could typically be advocated by (religious) moral advisers. Enjoying life (2,4) is the user-perception side of well-being, or in other words: whether they are happy. Studies presented in this section will especially focus on (1,4): the assessment of a person's health or fitness.

Rabbi et al. [56] attempted to measure mental and cognitive well-being by capturing audio and motion data of elderly. They hypothesised that "Continuous recording of daily audio patterns, specifically relating to the amount of human speech, would be linked to social and mental well-being". Participants (N = 8) wore a mobile sensing device which contained several sensors, such as a 3-axis accelerometer and a microphone. They found a strong negative correlation (R = -0.73) between the amount of sensed speech and depression using the Center for Epidemiologic Studies Depression (CES-D) scale, and a strong correlation (R = 0.82) between the amount of sensed speech and mental health based on the mental health score resulting from the Short Form 36 (SF-36).

Norris and Ph [44] focused on another mental health aspect: attentiveness. Participants (N = 37) were given either a meditative audio tape or an audio control tape. After listening to the 10-minute tape, they had to complete the Flanker task. They found that the meditation group had increased attention levels for incongruent trials compared to the control group, except for individuals higher in neuroticism.

In Zeidan et al. [74] it was found that a short meditation training would improve people's performance on cognitive tasks. Half of the 63 participants were given four meditative sessions in four days, whereas the control group were to listen to an audio book ("the Hobbit" by Tolkien). The group who attended the meditative sessions had increased cognitive functions compared to the Hobbit-group.

A study that is more focused on the long-term effects of meditation was carried out by Fabio and Towey [20]. Thirty-six participants were divided over two groups: one half consisted of long-term meditation practitioners performing meditation for over six years, and the other half consisted of people who never practised meditation. Results indicate that there is a relation between long-term meditation and some high-order cognitive processes. Hao and Chan [27] also focused on long-term effective meditation. They captured & analysed breath cycles and estimated real-time how breathing progressed using MindfulWatch, a smart watch based system. MindfulWatch can be used to improve the effectiveness of breathing exercises, and potentially support in reducing stress levels.

Besides meditation, there are other ways to improve and assess attentiveness. Participants in this area are often (high school) students [16, 17, 76]. Relying on a teacher their experience for determining students' attentiveness might not suffice or not be reliable in some cases. Therefore, one could make use of sensors, such as accelerometers and gyroscopes to assess attentiveness. An application of sensors for measuring attentiveness is found in handwriting detection and comparison [76]. A system that captures the student their attention is an instance of an Ambient Intelligent (AmI) system [17]. Such a system should be context-aware, adaptive and hidden in the background. A component of an AmI system that is regularly used for data collection is a monitor for mouse clicks and keyboard strokes [16, 17, 33, 49]. An advantage of this component is the invisibility and flexibility of data collection, and the fact that results can be monitored in real time. For example, in Durães et al. [16] attentiveness was measured using and AmI system. High schools students (N = 13) were given two kinds of lessons in which they had to perform tasks: a 'normal' lesson and an 'assessment' lesson. Mouse activity was monitored during these sessions. Applying the Kruskall-Wallis test, a weak correlation between attentiveness and task scores was found (R = 0.41), which could be explained by the difficulty of the tasks and the intelligence of the students. Closely related is a study by Pimenta et al. [49] who monitored mouse and keyboard activity to detect mental fatigue. Detecting mental fatigue is complex, as it is subjective and depends on many factors such as age, gender, work and food consumption. Since self-reported results (e.g. by means of a survey) might be inaccurate, behavioural results were collected. In particular, 15 mouse

and keyboard features were monitored for 20 participants. These data were then transformed such that it could be used for classification. The classification results were then visualised on a time line. Six of the aforementioned metrics showed significant differences for people who were energetic compared to people who were fatigued. In a follow-up study, the system architecture is described in more detail, and a neural network was used to classify fatigue in real-time [51]. Additionally, another study of Pimenta et al. [50] focused on the mental workload in relation to fatigue by classifying mouse & keyboard interactions combined with questionnaire data, which shows that in general the mouse & keyboard fatigue indicators aligned with how people experienced their fatigue. Besides using mouse & keyboard metrics to detect mental fatigue, results from Khan et al. [33] indicate that even a person's personality can be measured by monitoring their mouse and keyboard use.

Other sensors than mouse and keyboard have been used. An example of such is a system designed by Hall et al. [26] to monitor vital signs. They developed a system to measure a person their heart rate wireless with an accuracy of 90% in an office cubicle environment. In addition to this versatile system, a more specific approach was chosen by Wijsman et al. [73], as their focus was especially on detecting mental stress in environments similar to the office. They systematically measured physiological properties such as heart activity using an electrocardiogram (ECG), respiration and skin conductance. People their conditions were classified as 'stress' or 'rest' with an accuracy of 74.5%.

For most of the aforementioned field studies in the areas of detecting and stimulating physical and mental well-being, we can observe that the **number of participants** is relatively small compared to survey-based studies. For our user (field) study, we experienced difficulties regarding three aspects:

1. A limited amount of resources was available for distribution to participants
2. Participants that meet the conditions of the user study are scarce
3. A user study requires intensive guidance

Survey-based studies are to a lesser extend limited by these aspects, for (1) as their resources are inexhaustible and for (2) as surveys can often be distributed to a relatively large group of people, seizing a relatively small amount of their time. Aspect (3) concerns answering questions participants have about the user study, solving technical, logistic and planning problems ad hoc, and observing whether both participants and supervisors do not influence the outcome of the user study in an artificial way. Increasing the number of participants for a field study requires sufficient resources, or monetary means as in Finkelstein et al. [22] to obtain sufficient resources and reward participants for their participation.

2.4. Comparison & user experience of IPAs

Intelligent personal assistants could potentially be used to strengthen the social bonds of elderly. Reis et al. [57] compared Microsoft Cortana, Amazon Alexa, Google Assistant and Apple Siri on five features:

1. Basic greeting
2. Email management
3. Social network management
4. Social and family events management
5. Social games

These features should lead to user identification, state of mind assessment, obtaining information on the current context, personal information acquisition and providing a set of activity proposals. Amazon Alexa contains most of aforementioned features.

In Purington et al. [55] it was observed how people perceive Amazon Alexa, how it can be predicted whether people are likely to personify Alexa and how this can be influenced by factors such as integration with other services. This was done by collecting reviews of the Echo device from `www.amazon.com`. The content of these reviews was analysed and factors such as degree of personification (e.g. people referring to the device as a person), degree of sociability (perceived by the user), integration (with other services), technical qualities and household characteristics (who is/are using Alexa) were measured by means of this content. From the results, it follows that technical qualities and degree of personification are most correlated with the overall user satisfaction. Households with children who use the Amazon Echo are more likely to personify the device than single users.

The degree of sociability consists of five categories, of which users could select more than one:

1. Information source (news / weather)
2. Provider of entertainment (music / jokes)
3. Assistant (schedules / timers)
4. Companion (conversation partner)
5. Friend

79% of the users described Alexa as a provider of entertainment. One third of the users think of Alexa as an assistant, 5.5% of the users see Alexa as a companion, and 7.2% as a friend.

Personification of personal assistants also play a role in Cowan et al. [12] who examined the experience of infrequent IPA users by means of a survey. They found that infrequent users were frustrated when the IPA asked them to interact with a (smartphone) screen rather than interacting with an IPA through speech. In addition, the lack of integration with third party services was experienced as frustrating. The use of IPAs also depended on the context of a user: people who did not make use of it frequently were uncomfortable talking with Siri in public spaces, an effect that is indicated in previous literature [18]. Also, trust in privacy, the IPAs accent, and the lack of conversation functionality were perceived as weaknesses for infrequent users.

In addition to aforementioned survey-based approach, the user experience of IPAs was extracted in a series of interviews with 14 different users by Luger and Sellen [39]. A convenience experienced by users was that voice commands allowed multitasking, because it can be done hands-free. In addition, it was observed that users conversed differently with IPAs compared to people, for example by talking slow and clear and by leaving out unnecessary words. One of the users tends to talk more informally in private than in public. It was also found that satisfaction of IPAs was strongly correlated to the willingness of investigating the possibilities of an IPA. Also, people who used it more frequently were more successful in using it. First engagements with an IPA started with playful interactions for almost every user. In general, it can be concluded that there is a gap between user expectation and system operation that could be bridged by providing ways to reveal the system their intelligence and skills to the user.

Saad et al. [61] hypothesised that adding a visual component to a virtual personal assistant (VPA) would improve the quality of (user) experience (QoE). QoE is measured on eight aspects, amongst others user satisfaction and user expectations. They developed four systems: a VPA (s1), a VPA + an avatar on a 2D display (s2), a VPA + an avatar on a 3D display (s3) and and a VPA + an avatar living in a virtual reality, visualised with an Oculus Rift (s4). Participants had to perform simple tasks such as sending an e-mail and adding events to a calendar. s2 had the least variance in QoE, whereas s4 was perceived as having the most potential to increase QoE when communicating with a VPA.

Studies discussed in this section especially examined user experiences regarding IPAs. The majority of the people describe Alexa as a tool for providing entertainment such as music, and as an assistant e.g. for scheduling their calendar. It was also found that people who infrequently interact with IPAs rather converse by voice than via their smartphone, and that talking in public spaces is perceived as uncomfortable. In addition, people conversing with an IPA behave in a different way than people conversing with other people, indicating an IPA may be personal but is not yet 'a person'. Combining virtual reality and IPAs is promising for enhancing the user experience.

2.5. Enhancing the experience of reflecting on physical activity

Prior work investigated how to detect physical activity using mostly objective metrics from sensors, such as movement, physiological and contextual sensors. Many of these sensors are embedded in modern smartphones, and since nowadays most people possess a smartphone with internet access, transmission of sensor data to other devices and services is possible. Studies carried out by Cambo et al. [3] and Zuckerman and Gal-Oz [77] used smartphones for detecting and stimulating physical activity, where the latter presented quantified physical activity data to participants in order to reflect on it. However, results from Finkelstein et al. [22] indicate that activity trackers are ineffective for stimulating a physically active lifestyle. Section 2.4 elaborates on the perception of IPAs, mostly obtained through interviews and surveys. Nonetheless, there is a limited number of studies actually using the IPA for a field study. There are good reasons to believe that IPAs are able to enhance a person their experience on physical activity reflection, as they promise to be aware of a person their whereabouts and because they can be proactive [62]. Taking these considerations into account, we designed Alexercise, a system for our user study on which chapter (4) will elaborate. The next chapter will will elaborate on how Alexercise is established, and how the user study was designed.

3

Methodology

As mentioned in the introduction, the intention of this user study is to track and stimulate physical activity of participants during office hours, and to verify whether an intelligent personal assistant (IPA) has a more positive effect on the amount of exercise compared to activity trackers (e.g. smart watches or smartphone fitness apps). In section 3.1 it is explained why Alexa is the IPA of choice. A participant their insight in physical activity levels is necessary for them to be able to reflect on their amount of exercise. Methods for perceiving and collecting these insights are described in section 3.2. With the ingredients 'personal assistant' and 'activity tracker', a user study was tailored. Information about the group of participants that participates in this user study and the exact experimental setup are described in detail in section 3.3. The last section (3.4) especially focuses on how results from the field study are being evaluated.

3.1. Choice of IPA

Nowadays, one is able to choose from a variety of IPAs. Companies such as Amazon (Alexa), Google (Assistant) Apple (Siri), Microsoft (Cortana) and Baidu (DuerOS) respectively configured their assistants on headless devices such as Amazon's Echo[1], Google's Google Home[2], Apple's homepod[3], Microsoft's Harman Kardon speaker[4] and Baidu's DOSS Smart Speaker[5]. Besides these companies, smaller organisations also enter the market of IPAs with their product. An example of such an intelligent personal assistant is Mycroft: an open source voice assistant running on the headless device Mark II[6]. In addition to headless devices coming with a personal assistant, Mycroft released Picroft, a software package containing the implementation of the Mycroft personal assistant, which can be installed on a Raspberry PI[7].

At the moment of writing, there is little scientific research available that compares different IPAs on functionality. One of the few examples of a study that aims to do this is Reis et al. [57], as discussed in section 2.4. They concluded that Amazon Alexa is a device with a broad set of features compared to other IPAs, such as sending an e-mail, checking the weather and online shopping. However, this result could be very different when IPAs are assessed on different properties. Before this study was initiated we already possessed an Echo Dot device. Since scientific literature does not point out significant differences between IPAs, we decided to first discover options with devices that were already in our possession. Thus, for this study, we explored the possibilities of Alexa and Mycroft in practice. Mycroft is the most flexible option in the sense that it is open source and could therefore be configured to our needs. In addition, it could be implemented on different types of hardware, such as smart speakers, personal computers, smartphones and even on Raspberry Pi's. We carried out an exploratory experiment with Picroft using the Raspberry Pi 3 model B[8] with an external microphone, power supply cable, a micro SD containing the Picroft OS and external speakers as in figure 3.1b. For

[1] https://www.amazon.com/dp/B06XCM9LJ4
[2] https://store.google.com/product/google_home
[3] https://www.apple.com/homepod/
[4] https://www.microsoft.com/en-us/p/harman-kardon-invoke-with-cortana-by-microsoft/8rl7xlnwn95v
[5] https://dueros.baidu.com/en/
[6] https://mycroft.ai/product/introducing-the-mycroft-mark-ii-pre-order/
[7] https://mycroft.ai/documentation/picroft/
[8] https://www.raspberrypi.org/products/raspberry-pi-3-model-b/

comparison, we also evaluated the response time of an Echo Dot, see figure 3.1a. We asked questions of a different nature, such as factual and mathematical questions, a question where external services were involved, and a question that requires the capability of combining facts into new insights. In addition, we verified whether questions with the same intent but a different syntax resulted in identical answers. Overall, we observed that the response time of Mycroft was relatively slow compared to Alexa, as shown in table 3.1. The voice could be customised, however most of the available voices sounded relatively unnatural. Also, Mycroft seemed to not always understand the question if asked in a different way, for example it understood "What's tomorrow's weather in Delft", but Mycroft didn't know how to answer the question "What's the weather in Delft tomorrow?". In contrast, Alexa was able to understand both sentences and interpret them in the same way. Considering the downsides of Mycroft at the moment of writing, we decided to explore the possibilities of the Amazon Echo Dot. The Echo Dot is fairly adaptable: using the Alexa Skills Kit[9] one could add custom (programmable) functionality to the already wide range of features of Alexa. In addition, we could observe that it is relatively fast (see 3.1). Furthermore, the Echo Dot has several built-in attractive features such as "Alexa, tell me a joke", which possibly enhances the user experience. To ease development of a custom skill, a number of sample projects are available to quickly understand the practice of creating a custom skill[10]. We installed several third-party skills to explore the possibilities and limitations of custom skills. Examples of such are "Question of the day"[11] and "BBC News"[12]. Besides installing third-party skills of both informative and attractive nature, we created an Amazon Developer account and built our custom skill "GreetingFactory" using the Command-line interface of the Alexa skills kit (ASK-CLI)[13] to test basic reactive functionality (e.g. "Alexa, ask GreetingFactory to say hi"). The design and implementation of the custom skill for the Echo Dot will be discussed in more detail in chapter 4. We found that Alexa understood us well, responded relatively quick to our questions and for custom skills it is relatively easy to connect to third party services and APIs. Concluding, based on the good impressions from this exploratory experiment we decided to use the Echo Dot as intelligent personal assistant device for the current study.

Table 3.1: Benchmark tests of Mycroft & Alexa, using a Wi-Fi network of 52.0 Mbps down and 27.3 Mbps up, posing each question 5 times per device taking turns. Numbers in the first columns are approximations of average response times (RT), and the values in the second column indicate whether the IPAs provided a correct answer (CA).

	Mycroft		**Alexa**	
Question	*RT*	*CA*	*RT*	*CA*
Tell me a joke	4.5s	Yes	1.5s	Yes
What is an elephant	10.0s	Yes	1.5s	Yes
What is 154*72	4.5s	Yes	1.5s	Yes
What's tomorrow's weather in Delft?	4.5s	Yes	1.5s	Yes
What's the weather in Delft tomorrow?	4.5s	No	1.5s	Yes
Who's the current prime minister of Tuvalu?	6.3s	Yes	1.8s	Yes
What would happen if there were no moon?	7.5s	No	2.0s	No

3.2. Tracking activity

Collecting data about physical activity is required for Alexa to provide user-tailored advises. Measuring a person their physical activity can be done in several ways: (1) by means of activity trackers such as sport watches, e.g. Garmin[14] watches and (2) smartphones, using motion and contextual sensors such as the accelerometer and GPS sensor. For Alexa, it is important that physical activity data is regularly updated, such that the participant has access to their most recent statistics when informing Alexa about their physical activity. All communication with a custom skill happens trough the internet, i.e. Alexa sends requests over Wi-Fi to the

[9] https://developer.amazon.com/alexa-skills-kit
[10] https://github.com/alexa/
[11] https://www.amazon.com/gp/product/B01N6QUAXX
[12] https://www.amazon.com/TuneIn-BBC/dp/B01JHLI06S/
[13] https://developer.amazon.com/docs/smapi/quick-start-alexa-skills-kit-command-line-interface.html
[14] https://www.garmin.com

(a) Benchmark setup of the Echo Dot

(b) Benchmark setup of Mycroft running on a raspberry pi (Picroft), external microphone, and speakers

custom skill and the custom skill is able to make requests to external back-end services, like a service built on AWS Lambda[15]. Therefore, in terms of compatibility, it would be best if an activity tracker transmits (real-time) information via the internet, for example to a service that stores the information such that the Echo Dot could fetch that information whenever they require it. Such an activity tracker would best be wireless, as it prevents physical movement restrictions on the user. For that, a device using a network connection or Wi-Fi is required to provide regular updates. A smart watch that transmits data over Wi-Fi or to a smartphone via Bluetooth such as the TicWatch Pro[16] (or another watch using Google's WearOS[17]) meets these requirements. However, a limitation of smart watches is that either Wi-Fi should be available, or a smartphone should be nearby to keep up a connection to the internet via Bluetooth. For this study, it is likely people will go outside the office building where there is no Wi-Fi, meaning people have to take their smartphone with them at all times and keep Bluetooth enabled in addition to wearing a smart watch at all times. A simplification of such a system is a smartphone-only solution, where it is used as tracking device using the available movement and contextual sensors. A smartphone meets aforementioned requirements, as they could connect to the internet in indoor (Wi-Fi) and outdoor (mobile network) environments, have a number of movement and contextual sensors, and many people possess one as mentioned in section 2.1. For Android and Apple, there are around 13 common sensors included in smartphones [53]. In addition, there are many apps available that are able to convert sensor data into activity metrics such as number of steps. Sense-it is a mobile application that is able to detect all motion, position, environmental and body sensors on a smartphone [63]. Data can be viewed as a graph (Explore), downloaded as CSV file (Record) and send to a platform called nQuire-it (Share), used for obtaining scientific insights from the acquired data for a specific mission. An example of such a mission is the recording of sunlight via the ambient light sensor, for which people all over the world can join such that levels of sunlight can be compared over time for different locations. However, for the current study we need access to recent physical activity data and neither of the aforementioned options provided that option. An alternative to Sense-it is Google Fit[18], a mobile application for Android and web page to keep track of activities. Google Fit allows developers to access physical activity data via their API[19]. For access to the API, a Google account is required. Data such as number of steps, calories burned and location are request-able[20]. However, a downside of both Google Fit and Sense-it is that they are Android-based. At the time of writing, many potential participants were in the possession of an iPhone. Therefore, we had the following options:

1. Make use of an existing cross-platform smartphone application such as FitBit[21]
2. Build a cross-platform app for Android and iPhone smartphones

[15] https://aws.amazon.com/lambda/
[16] https://www.mobvoi.com/benelux/pages/ticwatchpro
[17] https://wearos.google.com/
[18] https://www.google.com/fit/
[19] https://developers.google.com/fit/
[20] https://developers.google.com/fit/rest/v1/data-types
[21] https://www.fitbit.com/app

The second option was preferable to the first one for several reasons. At first, available cross-platform solutions are not always clear about the interval of physical activity updates to the server. Second, the physical activity properties are already determined in case of existing applications. Building an application ourselves would give the opportunity to tailor the system to meet exactly our requirements without unnecessary functionality. However, a downside is that all sensor data processing mechanics should be configured manually. Considering pros and cons, we decided to go with a tailor-made cross-platform mobile application. This mobile application detects movement by means of exploiting the accelerometer and provides context and location data via GPS. As pointed out in section 2.1, describing physical activity in terms of number of steps is a method that is often used and easy to reflect on, because it is a quantification that is easy to grasp. Further design details and the design process of this mobile application are discussed in section 4.2. The next section elaborates on the user study for which this application will be used.

3.3. User study

The experiment was carried out in four stages over a period of two and a half months, for which 16 people participated in total. Rounds one and two took place in Delft, The Netherlands and the latter rounds in Tallinn, Estonia. For the first and second round, entrance requirements for participants were (1) they should work in the office and (2) either possess an iPhone or Android phone. Because remote distribution of iPhone apps is hard to establish, the constraints became stricter for the last two shifts, i.e. the participant required an Android phone to participate. Further details of the experimental setup can be observed in table 3.2.

Table 3.2: Details about experimental setup per round

	Start	**End**	**№ of participants**	**№ of iPhones**	**№ of Android phones**	**Locations**
1st round	16/07/18	03/08/18	5 (♀= 1, ♂= 4)	4	1	Delft
2nd round	06/08/18	24/08/18	1 (♂= 1)	0	1	Delft
3rd round	03/09/18	21/09/18	4 (♀= 1, ♂= 3)		4	Tallinn
4th round	17/09/18	05/10/18	6 (♀= 1, ♂= 5)		6	Tallinn (4), Delft (2)

Participants received the following tools for either tracking of and reflection on physical activity:

1. DeskstApp, an application for Android phones and iPhones to record and visualise the participant their number of steps, depicted in figure 4.4a. Further details of this application are discussed in section 4.2.
2. An Amazon Echo Dot (Alexa) to voice-support the participant in being physically active.

DeskstApp both records and visualises of the participant their number of steps. Alexa converses with a participant in a reactive manner, supporting them to increase their physical activity levels. To make connecting with the Echo Dot more appealing, some attractive features were be implemented. Surprising functions are found to be one of the first encounters with an IPA for people who are not yet familiar with IPAs [39], with the underlying assumption that most of the participants did not use an IPA before as it is a relatively novel phenomenon. This assumption is later confirmed by the participants through a survey, discussed in section 3.4.3.

Every round in the experiment consists of three phases, where each phase has the duration of one week:

1. **Passive tracking phase**: there is no reflection on the performed physical activity
2. **App reflection phase**: physical activity is visualised in DeskstApp
3. **Alexa reflection phase**: physical activity can be requested from Alexa

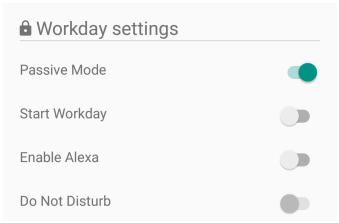
In the passive reflection phase, we determined the physical activity baseline of participants. In the subsequent phases, the performed activity for each current day was revealed to them, and they were able to reflect on their number of steps. In addition, we added an extra level of motivation to DeskstApp. According to the goal setting theory described by Locke and Latham [38], a specific goal of above average difficulty is more effective than merely urging people 'to do their best'. Also, self-efficacy commits a participant to his goal. Additionally, prior work indicated that participants wanted to dynamically update their daily goal [77]. Therefore, we allowed participants to set a step goal to their liking, starting at a presumably above-average difficulty of 8000 steps. They were urged to set a realistic goal on a daily basis and they were allowed to change their goal

at any time. Each experimental round was carried out in a time frame of three successive weeks. Participants were divided into two groups, performing the same tasks in the first week. In the second week, the first group reflected on their physical activity using DeskstApp and the second group using the Echo Dot. The week after, the groups switched roles. Initially, the group sizes of both groups should have been identical. However, in practice this was not the case. In round four, the division of participants should have been 2:4 instead of 4:2. However, in order for the Echo Dot to work a stable Wi-Fi network is required. In the second week of the 4^{th} phase there was a maintenance in the office building of two of the participants, which caused the Wi-Fi network to be down, meaning they had to be shifted from the Alexa reflection phase to the app reflection phase.

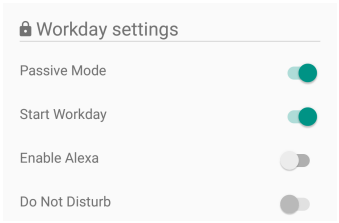
For the passive tracking phase, the detected taken number of steps was hidden for the user until the beginning of the next phase. This was done by toggling the 'passive mode' switch in the app, which would disable any means of reflection on physical activity for the participant. Implementation of this switch ensures a smooth transition between the phases, with little interference from our side. For three subsequent weeks, every day the user had to 'start' his workday manually, meaning that they should toggle a switch in the app. When the day was over, they should again toggle the switch to stop the workday. An example of such can be seen in figure 3.2a and 3.2b. This way, there is a clear time frame for which physical activity was measured. For participants in the app reflection phase, an instruction sheet was handed out, informing them of DeskstApp their features. They had to disable the passive mode, such that a visualisation for the number of steps became visible on another screen, like the left hand side of figure 4.4a in section 4.2. In the Alexa reflection phase, participants would receive an Echo Dot device and instructions on how to use it, such as example phrases for the custom skill. participants should enable the 'enable Alexa' switch in the app. This would disable physical activity visualisation in the app and enable a message that informed users that Alexa would provide information about their physical activity. This message is visualised on the right hand side of figure 4.4a. As the user would be in the Alexa phase, they could receive notifications about their activity. Notifications could in some cases be timed at an awkward moment, e.g. in a meeting. Therefore, participants had the option to enable the do-not-disturb switch, in order to disable notifications. They would still be able to communicate with Alexa while do-not-disturb was enabled.

Figure 3.2: Different states of the app

Passive tracking phase

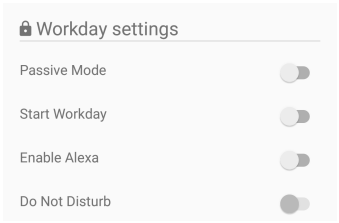


(a) Passive tracking - workday off

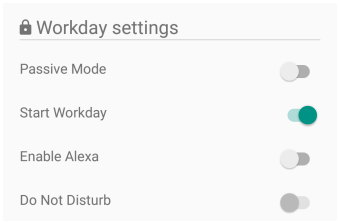


(b) Passive tracking - workday on

App reflection phase

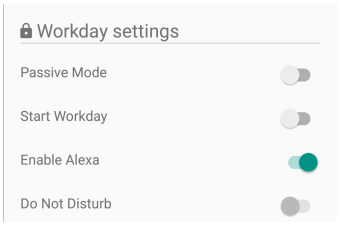


(c) App - workday off

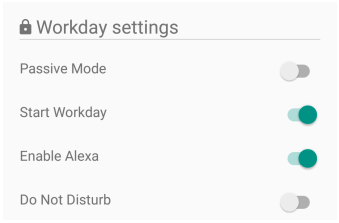


(d) App - workday on

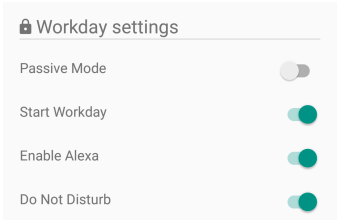
Alexa reflection phase



(e) Alexa - workday off



(f) Alexa - workday on



(g) Alexa - workday on with DND

3.4. Evaluation process

Metrics for the evaluation of physical activity consisted of the number of steps taken by participants, GPS data and interactions with Alexa, which are analysed on both individual and aggregated level. Results from the pedometer are leading for determining physical activity, and GPS data is especially used to verify the correctness of pedometer data. Interactions with the Echo Dot should verify whether a possible increase in physical activity levels would indeed be caused by Alexa. Mainly, results should point out to what extend participants increased or decreased their physical activity levels when using Alexa compared to using the app, as follows from the first research question discussed in section 1.3. Before such conclusions could be drawn, as a first step in the process of evaluating the results, data should be verified on its correctness. How this was done will be explained in section 3.4.1. In section 3.4.2, we discuss how physical activity was determined based on results from the pedometer. The last section focuses on how GPS data was used to verify physical activity and what role interactions with Alexa and self-reports of participants play in the process of evaluation.

3.4.1. Quality inspection of the data

Quality and reliability of the data could be skewed by a number of factors. Collected data should be assessed on the following criteria:

Sensor equality - for this study, different types of phones are used. Sensors from each phone might behave different, possibly producing deviant results. It should be verified whether the step count detection algorithm works identical for every participant in the study. In addition, it should be confirmed that the phone of each participant does not only sends the right data, but also sends the same amount of data. As described in section 4.2, approximately every ten seconds a batch of magnitude-data is sent to the server, based on values produced by the accelerometer. The interval for which the accelerometer collects data might be different for each type of phone.

Disciplinary behaviour - we kindly request participants to perform certain tasks, such as starting and stopping the workday in the app and taking their phone with them at all times. However, it might occur that participants forget to do either of these tasks, thus we need to take these limitations into account when determining whether there is an overall increase in physical activity levels.

If these criteria are met, it ensures data uniformity amongst participants. Further details on how data are collected are specified in section 4.2.

3.4.2. Assessment of physical activity

The leading metric for assessing physical activity is the number of steps. This number is calculated by using data collected by DeskstApp for the step detection algorithm. How this number is determined exactly is reported in section 4.1.2. We investigated in a post-experimental analysis whether the level of physical activity expressed in number of steps increased for individual users. For each individual participant we looked at the average daily number of steps per week. The daily average is chosen over the weekly total number of steps, because participants were not necessarily asked to work five days a week, as we wanted to approach a real-life situation as much as possible. We compared the passive tracking phase, app reflection phase and Alexa reflection phase using the average daily physical activity per week to research whether there is an increase in physical activity levels. Besides observing a change on individual physical activity levels, it can be determined whether there is a collective increase or decrease in physical activity per phase.

3.4.3. Correctness verification of measured physical activity

In addition to the number of steps, several other aspects are measured, such as a participant their location, their interaction with the buttons in DeskstApp, interactions with Alexa and notifications sent from Alexa to the participant. GPS data is used to verify movement detected by the step detection algorithm and whether the amount of detected physical activity seems correct. In addition, it is used to calculate the walked distance on a single day, by adding up the distance between each of the measured points. It should be taken into account that GPS data were often collected when the participant was located indoors, possibly resulting in a lower accuracy. However, the plugin that was used for monitoring the user's location is able to provide an accuracy estimate in meters[22]. In addition to physical activity verification by location, we could justify physical activity by means of observing interactions between the participant and Alexa. If, for example, a

[22] `https://github.com/mauron85/react-native-background-geolocation`

participant shows an increase in physical activity during the Alexa reflection phase while he did not receive a notification and does not converse with Alexa, it is likely that there is a different motivation for performing physical activity. On the opposite, if there is a lot of interaction with Alexa for a specific participant, and at the same time they perform little or zero movement, it is likely that Alexa does not stimulate them to undertake action to reach their daily step goal.

In section 2.4, we discussed several studies who investigated user experience of IPAs. Assessing the 'personal' aspect of intelligent personal assistants provides context to our results, and especially to the 'why' questions such as 'why did physical activity levels (not) increase for a particular participant?'. Such qualitative results were collected via three surveys. Prior to the study, we distributed a survey requesting basic contact information from each of the participants, such as name, age and gender. After the passive tracking phase, a second survey was distributed, requesting participants to elaborate on personal aspects such as the duration of their workday, use of activity trackers, sports in their spare time and knowledge on the subject of physical activity & health. As post-experiment survey, we asked participants to share their experiences with Alexa and DeskstApp and to reflect on their own behaviour, for example whether they were interested in performing more physical activity and whether they were carrying their phone with them all day.

Taking considerations concerning IPAs, activity trackers and the user study into account, we built a system able to track and stimulate physical activity for office workers. In the next chapter, we discuss how Alexercise works, which components are involved and the challenges we faced during the development process.

4

Alexercise

Reflecting on physical activity and being stimulated to perform exercise is facilitated by a variety of interacting components, referred to as Alexercise. It should fulfil several tasks in order to achieve its goal of making the participants reflect on their physical activity. From the perspective of data flow, the tasks of the system are as follows:

T1 *Collect raw activity data & interactions between participants and their IPA*

Before any physical activity reflection or stimulation technique could be applied, the system should be aware of the participant's activity. Thus, their motion and location data should be acquired. In addition, interactions with the IPA could give an indication of how one perceives physical activity information from an IPA and whether it results in more exercise.

T2 *Process primary data and store 'clean' data*

Collected data such as a series of GPS coordinates and an array of raw sensor data is not yet useful for getting insight into performed physical activity. For participants, physical activity information is more glance-able when it is quantified and visualised in some way. To achieve that, primary data should be processed and shaped such that it can be used for presenting it to them. An example of such is converting the accelerometer data collected from a participant's phone into the number of steps that they have taken on that day.

T3 *Present activity data*

As a third step, clean data should be expressed to the participant, either through Alexa or through the app. This way, the system allows participants to get insight in their daily activity. Participants should then be able to decide whether they should perform more physical activity.

Three components together form the system: (1) **Alexa**, (2) **DeskstApp** and (3) the **Back-end**. Each of aforementioned tasks are assigned to one or more components of Alexercise. T1 is fulfilled by DeskstApp and Alexa, which respectively collect accelerometer data and voice interactions. The second task is solely executed by the back-end, which calculates the number of steps by means of an algorithm that takes accelerometer data as input. This 'clean' data is then requested by the participant via either the app or Alexa, as described in T3. An overview of all components and their interactions is visualised in Figure 4.1. The following sections will elaborate on Alexercise their components in more detail.

4.1. Back-end

This component is a part of the system that is not directly used by the participants. particularly for our system, this is hardware and corresponding software that transforms raw data and stores transformed data, such that it is fit for presentation to the participants. More specifically, this resulted in the setup of a Model-View-Controller (MVC) project and a Web API, both in ASP.NET Core[1], using Microsoft SQL server[2] as relational database. The Web API was used to serve requests from Alexa and DeskstApp. We used the View component of

[1] `https://github.com/dotnet/core`

[2] `https://www.microsoft.com/nl-nl/sql-server/default.aspx`

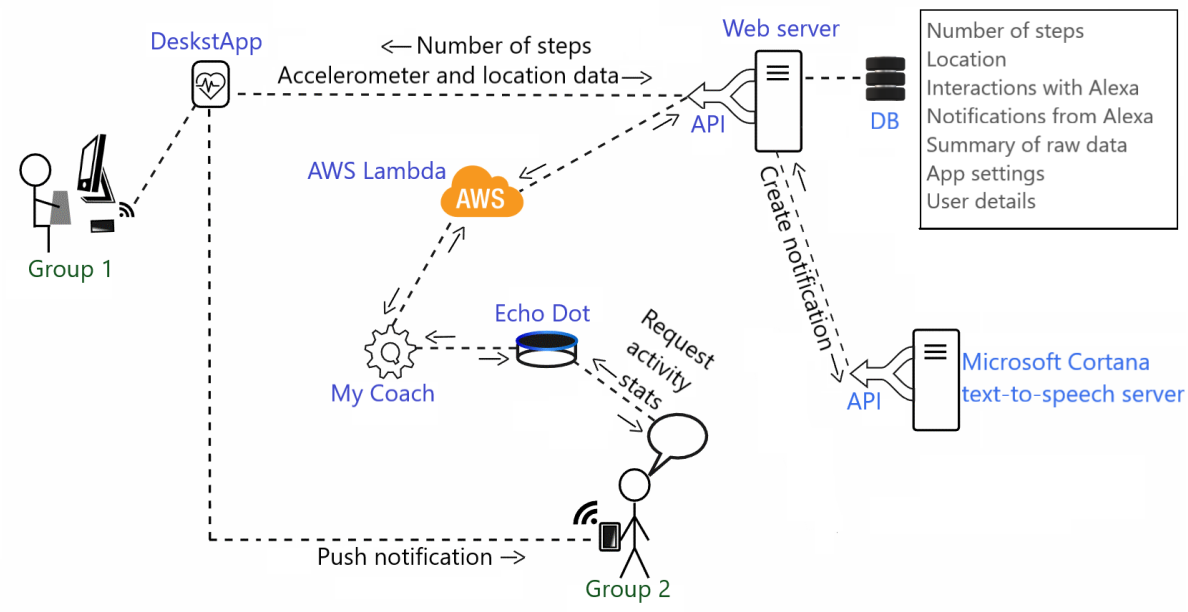


Figure 4.1: System components & interaction

this project to visualise the collected data. Code management was done via GitLab[3] using a private repository. In addition we subscribed to the text-to-speech service, part of Microsoft Cognitive Speech Services[4], further explained in section 4.3.6.

As discussed in section 3.3, multiple participants will use the system simultaneously, thus the system must be able to distinguish different participants. In other words: if there is an incoming request for any kind of data, the system should know which participant is asking and whether they are allowed to receive that specific data. To achieve this distinction, participants should be authorised in such a way that they only have access to their data. In section 4.1.1 this is explained in more detail. In section 4.1.2 it is explained how data from DeskstApp and Alexa is 'cleaned' and how the number of steps for a single user is calculated. Section 4.1.3 elaborates on how the system is being monitored via the back-end.

4.1.1. Authorisation

For authorisation, OAuth 2.0's JSON web tokens (JWT) were used, or more specific the JWT bearer token authorisation grant type [32]. A client provides his credentials to the identity provider, in our case a username and password. As a result, he will obtain an access token for authorisation, a refresh token for obtaining a new set of tokens and an expiration date of the obtained token. The resource provider is the party that provides the client the requested data (in our case the back-end). By means of an access token, the resource provider is able to distinguish which user requests the data and whether that user has access to the requested data. When a token is expired, the refresh token can be used to obtain a new set of tokens from the identity provider. Other system components which will be discussed later in more detail will also conform to the OAuth 2.0 standard.

4.1.2. Data processing

Using OAuth authentication, the back-end is able to make a distinction between users, and ensures data provision for the authorised user only, providing a safe environment to process incoming data. Such data will consist of accelerometer magnitude values, GPS coordinates, several app settings and spoken sentences by Alexa to the participants, used as follows:

Accelerometer - participants were able to reflect on their physical activity, based on the number of steps they have taken during their workday. For that, DeskstApp collected accelerometer values, computed the magnitude of each of the accelerometer vectors and sent a series of these to the back-end. This is explained

[3] https://about.gitlab.com/
[4] https://azure.microsoft.com/nl-nl/services/cognitive-services/speech-services/

in more detail in the sections 4.2 and 4.3. From plain accelerometer magnitude values it can not immediately be determined how and if a person is moving. A certain data transformation was required to be able to convert it to a single number. For that, we built an algorithm that takes the values representing the magnitudes of the accelerometer vectors as input and outputs a single integer [41]. A simplification of the implemented version of the algorithm is described in algorithm 1.

Location - DeskstApp kept track of the participant their location, discussed in detail in section 4.2. The back-end received a coordinate, a time stamp on which this coordinate is generated, the provider of this coordinate and the perceived accuracy in meters, as of table 4.1. A record containing these properties is stored directly in the database, and is used for post-experiment analysis of the results.

Interactions with Alexa - as we will see in section 4.3, the custom skill for Alexa requires prepared string sentences as a response. These sentences are made in advance and stored at the back-end, because by doing so, all data is stored in the same place which provides a clear overview for result analysis. However, this means that the Echo Dot required access to the back-end in order to facilitate a conversation with the user. Establishing an authorised connection does not require any additional effort, because access to the back-end is already required to obtain physical activity information. Section 4.3.5 explains how an Echo Dot is able to access user-specific information. The prepared sentences were stored in a static JSON file, consisting of multiple string arrays, one for each type of custom skill function (or 'intent', see section 4.3.1). Strings contained in these arrays represent the sentences that were spoken out loud by the Echo Dot, sometimes containing variables. An example of a sentence with a variable is "You have taken {1} steps today", where {1} should be replaced with the current number of steps during run time. If the back-end receives an authorised request from Alexa, it will randomly take one of the sentences made in advance from the corresponding array, exchange the variable placeholders (if any) with the proper values, store the final sentence in the database together with the participant their ID, and send back the response to Alexa.

4.1.3. System Monitoring

Alexercise has its vulnerabilities, as it consists of multiple components that should constantly be interconnected, simultaneously processing data produced by multiple participants. In addition, many different types of phones were used, which could not be tested beforehand, possibly resulting in non-uniform data. Additionally, we asked much of our participants in terms of discipline as mentioned before in section 3.4.1. In advance of the user study, we thought participants could possibly forget to execute their daily tasks. To mitigate potential problems, we ensured that there are several types of feedback. At first, we implemented the back-end in such a way that any irregularity in the data would be mentioned in log files. In addition, if there was a critical error, an e-mail would immediately notify us, containing a description of the error to focus the view for a possible solution. We also regularly checked if the system was still up and running, especially looking for recently added records to the database and whether a participant started his workday. At last, participants were urged to immediately notify us of any irregularity in any way.

4.2. DeskstApp

A key component of the system is the mobile application that was handed out to participants to monitor their movement. In section 3.2 it was determined that DeskstApp should work both for iPhone and Android phones. To achieve this, we built our application in React Native[5], which is a platform that allows app development for both Android and iOS in JavaScript and React[6]. The latter is a JavaScript library that is especially useful for building interfaces. An advantage of this approach is the large number of plugins available for tapping into smartphone sensors, which we needed for our data collection process. Before any data could be shared with the back-end, the app should provide user identification. Section 4.2.1 will explain how this is achieved. Section 4.2.2 describes how data was collected over a secure connection. To accomplish any form of reflection on physical exercise, the raw collected data required quantification and visualisation in an appealing way. On that, section 4.2.3 will elaborate.

[5] `https://facebook.github.io/react-native/`
[6] `https://reactjs.org/`

4.2.1. User identification

Earlier in section 4.1.1 we introduced the OAuth 2.0 protocol for authentication via JWT, including the promise that other system components would conform to this standard. The first step in the authorisation process is to provide credentials as a client (the app) for the identity provider (back-end). Therefore, we build two screens for registration and authentication, as depicted in figure 4.2. Upon logging in, credentials are sent to the identity provider over a secure (HTTPS) connection. If the credentials are incorrect, the app will display a detailed error message. Otherwise, the app will be given an access token. This token is included in the header of a resource request, as a means of authorisation. The next section will elaborate what data is collected by which part of the system.

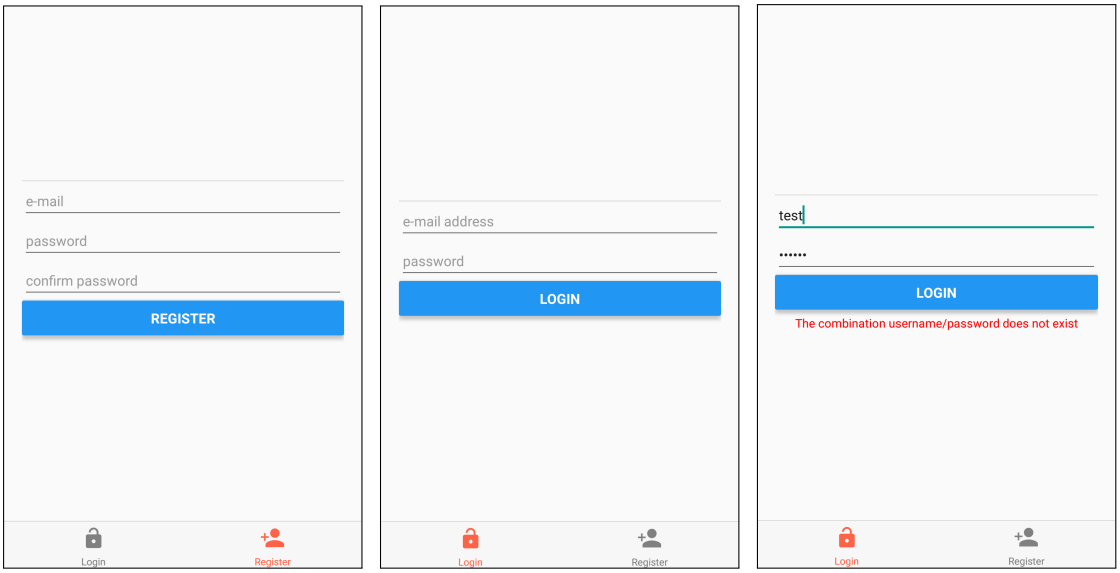


Figure 4.2: Visualisation of the register and login screens for Android, including the login screen displaying an error message

4.2.2. Data collection

In section 3.4.1 of the methodology chapter, we decided to emphasise the number of steps as single metric to present to the participants. To obtain this number, we tapped into the phone's accelerometer sensor. Accelerometer data went to a process of transformation until it resulted in the number of steps. The weight of this transformation process could be divided in two ways:

1. The *smartphone* transforms sensor data immediately, applying the step-detection algorithm
2. The *back-end* receives raw sensor data from the server and processes these data by means of the step-detection algorithm

Processing most of the data at the back-end side has several advantages. At first, it reduces battery depletion of the smartphone, because most of the processing happens on the server-side. Secondly, it allows storing not only the number of steps on the server side, but also (parts of) the raw data that is sent to the server, which might give an indication of the quality of the sensors. Therefore, we decided to host the step-count detection algorithm on the server-side. In section 4.1.2 we described the algorithm that takes acceleration magnitude values as input and outputs the number of steps. As discussed before in section 2.1, magnitude of the acceleration vector is calculated via the formula $mag = \sqrt{x^2 + y^2 + z^2}$. The variables x, y and z express the acceleration in three directions, like depicted in figure 4.3. Thus, the 'magnitude' of the vector consisting of $[x, y, z]^T$ describes the net acceleration of the phone. Both iPhone and Android phones have a built-in 3-axis accelerometer. iPhone claims that "All iOS devices have a three-axis accelerometer"[7]. For Android, accelerometers are available since API level three[8] - which means participants require a phone running an

[7] `https://developer.apple.com/documentation/coremotion/getting_raw_accelerometer_events`
[8] `https://developer.android.com/reference/android/hardware/Sensor#TYPE_ACCELEROMETER`

Android version newer than Cupcake (1.5)[9] released more than nine years ago[10]. Google keeps track of the percentage of Android versions that are active on the Play Store. The earliest version mentioned on their dashboard[11] is Android Gingerbread (2.3.x), used by 0.2% of all Android users. From this we concluded that almost every iPhone and Android phone currently used contains a built-in 3-axis accelerometer.

However, the accelerometers are not necessarily uniform across iPhone and Android phones. An important difference is the unit for expression of acceleration. For that, Android uses the SI units for acceleration[12] (m/s^2) and iPhone uses g-force[13] (g). The default gravitational acceleration equals to 9.80665 m/s^2 [69]. To achieve uniformity across the system's components, iPhone accelerometer axes were multiplied with 9.80665 individually. The rate for which accelerometer values were tapped from the sensor was once every 150ms. These values were then immediately converted to magnitudes of the acceleration vector. To reduce bandwidth, magnitude of the acceleration vector is already calculated in the app before it is sent to the server. To ensure the right order for which these values are received, a time stamp is appended to each value representing the magnitude of the accelerometer vector. If we would sent each of this values to the server individually, the number of requests would be equal to $\frac{1000}{150} \cdot 60 = 400$ per minute. Therefore, magnitude values were stored in the phone's local storage for ten seconds and bundled in a single request to the server.

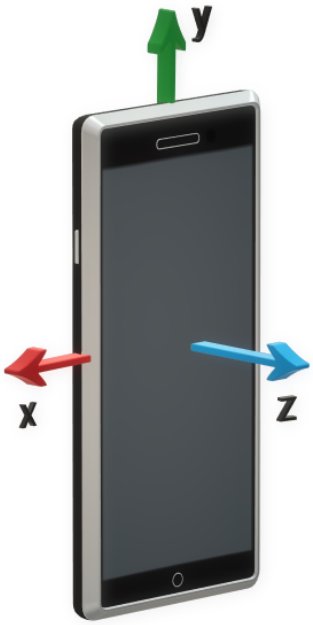


Figure 4.3: 3-axis accelerometer

In addition to aforementioned accelerometer values, as a means of verifying the validity of accelerometer values, the participant their location is frequently determined. For this, we used a background geolocation service[14]. Various properties of location are obtained as explained in table 4.1. Each location record is directly sent to the back-end, at which it is stored for evaluation in a later stage as discussed in chapter 5.

4.2.3. Presentation of physical activity

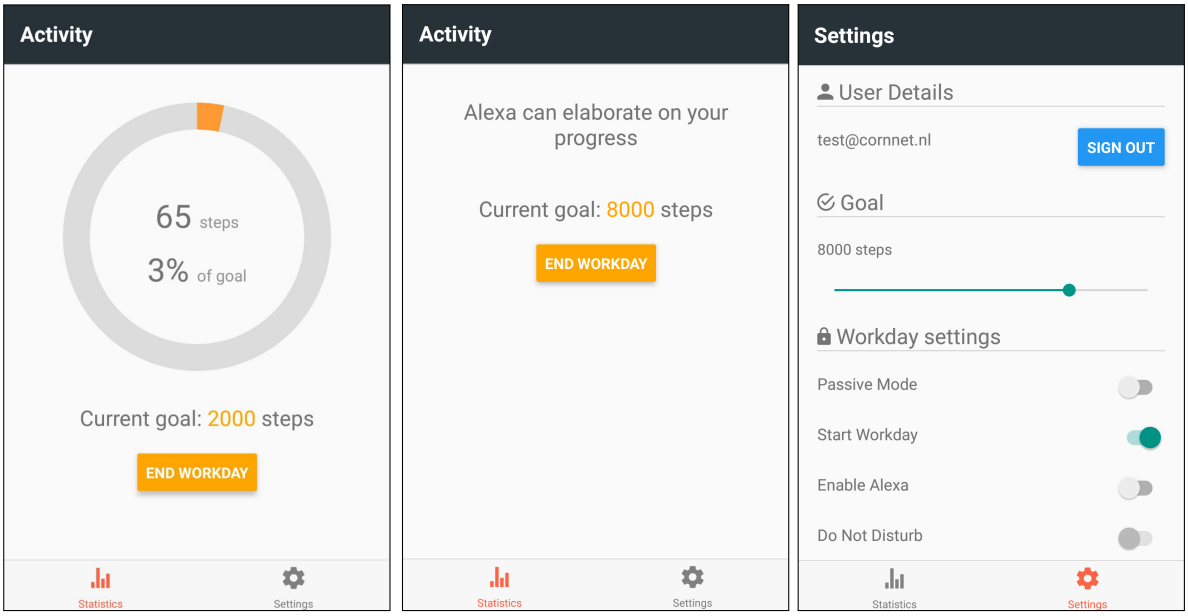
Visualising physical activity in an attractive and informative way is a crucial aspect in the design of DeskstApp. Participants should be able to get insight in their daily physical activity levels. In section 2.2 it was questioned whether gamification elements in promoting physical activity resulted in a significant improvement in movement. As it turned out, gamification elements are not necessarily more effective than quantification. In ad-

[9] https://source.android.com/setup/start/build-numbers
[10] https://android-developers.googleblog.com/2009/04/android-15-is-here.html
[11] https://developer.android.com/about/dashboards/
[12] https://developer.android.com/reference/android/hardware/SensorEvent#values
[13] https://developer.apple.com/documentation/uikit/uiacceleration
[14] https://github.com/mauron85/react-native-background-geolocation

Table 4.1: Specification of the location record

Geolocation	Coordinates expressed in latitude and longitude
Time stamp	Time at which a coordinate is recorded
Provider	Value representing how the coordinate is determined. The provider can take the values GPS, network, fused[15] and passive[16]. Fused is a combination of location technologies and for example calculates location via combining GPS and Wi-Fi. Passive is a provider that does not actually request the location, but will return the values generated by other providers
Accuracy	Presumed accuracy of the coordinate in meters

dition, it was described that a glance-able display is well-perceived by participants. This inspired us to keep our application interface simple and to highlight the participant's physical activity in a quantified way, but on a glance-able display. In addition, it was discussed in section 3.3 that participants should determine their own step goal. Having access to both the participant's goal and his number of steps creates the possibility to compare them in a meaningful way. Therefore, to visualise the physical activity of the participant for the participant, we used a circular progress widget plugin for React Native[17] that provides an indication of step progress of the current day compared to the daily goal. Because the user study will consist of two groups (see section 3.3), DeskstApp requires two versions of this screen: one for participants in the app reflection phase, and another for participants in the Alexa reflection phase. Participants in the app reflection phase require visualisation through the app, thus the widget is shown to them. Participant reflecting with Alexa are given a message that Alexa is able to provide feedback on their physical activity. Screen shots of both versions of the app are depicted in 4.4a. Participants could adjust their goal in the screen containing the app settings shown in figure 4.4b. In addition, they could tweak settings to configure the right experimental phase, as explained in section 3.3.



(a) Visualisation of the app for Android

(b) Settings screen for Android

[17] https://github.com/bartgryszko/react-native-circular-progress

4.3. Alexa

Transmission of physical activity information from the back-end via Alexa to the participant is needed functionality that meets the requirements of the Alexa reflection phase. Conversing with Alexa requires the use of sentences with a fixed syntax. This fixed syntax is not necessarily unnatural to human beings, as can be observed in the first section. The custom-made Alexa application used for our system is explained in section 4.3.2. Alexa itself consists of different parts that transfer prepared sentences from the back-end to the Echo Dot device, where they are spoken out loud. How these internal parts interconnect will be explained in the architecture section, with an example use case in section 4.3.4. Similar to the other components, Alexa should be able to connect to the back-end and fetch the necessary information. Section 4.3.5 elaborates on how the user is identified by Alexercise and how data between Alexa and the back-end is transferred in a safe way. At last, the notification feature of our system is explained in section 4.3.6.

4.3.1. Syntax

An important feature of Alexa is the Custom Skill[18]. Through the developer portal of Amazon one can add custom functionality - a skill - to the existing set of applications available for Alexa. The concept is similar to apps on a smartphone: one is able to install existing skills (apps in the analogy) via the Alexa App. Such a skill consists of several components defined by Amazon (table 4.2).

Table 4.2: Explanation of custom skill components

Custom skill component	Explanation
Intents	Actions that are requested by the users (e.g. 'order a pizza')
Sample utterances	A specific set of words the users may use to invoke an intent. For example, if the user would like to order a pizza (intent) he might say 'get me a pizza', 'i want pizza', 'order a pizza now'. These utterances all map to the same intent.
Invocation name	Name that is unique to the custom skill and that should distinct it from other skills.
Cloud-based service	Handles the intents. In the pizza analogy, a cloud-based service should make sure that the pizza company receives an order and deal with the payment.

Every request for or conversation with Alexa starts with the wake-word 'Alexa'. This will make sure that any Alexa-enabled device such as the Echo Dot is listening to the words that follow after the wake-word. Alexa could have more than a single skill installed, thus the user should make sure they direct their request to the right skill by mentioning the skill invocation word. At last, the sample utterances follow to further specify the user's request. A spoken request to Alexa typically has a structure as defined in table 4.3. More structures are defined on the website of Amazon for developers[19]. Words such as 'ask' and 'tell' and connecting words (by, from, in, using, etc.) are automatically recognised by Alexa as such.

Table 4.3: Syntax of Alexa requests

Example	Structure
Alexa, ask *My Coach* for my steps	Alexa, ask [invocation name] [sample utterance]
Alexa, get my steps using *My Coach*	Alexa, [sample utterance] [connecting word] [invocation name]
Alexa, tell *My Coach* to get my steps	Alexa, tell [invocation name] [connecting word] [sample utterance]

[18] https://developer.amazon.com/docs/custom-skills/understanding-custom-skills.html
[19] https://developer.amazon.com/docs/custom-skills/understanding-how-users-invoke-custom-skills.html

4.3.2. My Coach

Physical activity statistics are retrievable via My Coach, the self-made custom skill that is part of our system. We defined five intents for participants to use:

1. Retrieve Nº of steps
2. Fetch remaining Nº of steps intent
3. Ask for a motivational quote
4. Sing a stimulative song
5. Start the current workday

Intent (1) and (2) are merely informative and could respectively be invoked with sample utterances such as 'ask for my progress' and 'get my remaining steps'. Intent (3) and (4) are implemented as slightly mocking attractive features and can be triggered with phrases such as 'ask for a motivational quote' and 'sing a stimulative song'. The last intent is a functional one. Every workday, participants were asked to trigger the start workday intent, which started a small conversation between Alexa and the user. The participant would ask Alexa to start the workday, on which she would reply with 'What time do you expect to go home today?'. They would then reply with an expected end time, and if this was a valid end time, Alexa would thank the participant. Alexercise would then be aware of the expected end time of the participant and whether they were on schedule in terms of physical activity. This functionality is useful for sending notifications as discussed in section 4.3.6. In the next part, we will elaborate on how My Coach works behind the scenes.

4.3.3. Architecture

Alexa and the Echo Dot sometimes seem interchangeable expressions of the same concept. However, when talking about the architecture of the system it is important to emphasise the difference. When talking explicitly about the Echo Dot, the device itself is the subject. Alexa is the personification of the whole set of components that together are the system, thus including the Echo Dot. That being said, we can identify the software components involved in creating a custom skill for Alexa, not to be confused with the internal components of a custom skill as discussed in the previous section. Up until this point, we have an Echo Dot that has access to the custom skill 'My Coach' that is able to receive commands from a participant, triggering an intent if done correctly. However, without any external information supply, My Coach is not able to provide any useful physical activity information. Therefore, we require a Cloud-based service. According to Amazon's documentation there are two options for such a service: a custom web service[20] or an AWS Lambda function[21]. The latter is, as they call it, a serverless solution, meaning that a developer is able to directly create a service without the need of a self-owned server. Because AWS Lambda is well-documented, cost-free for the first one million requests and because there are many Node.js samples available (one of the languages of choice on AWS Lambda) we decided to build the web service in Node.js on AWS Lambda. While faster developing of the web service is an advantage, a disadvantage of this approach is that an additional connection is required with the web server containing the physical activity data. In this case, the advantages outweigh the disadvantages because of the Alexa Skills Kit (ASK) SDK for Node.js[22] which greatly simplifies the way requests are handled.

Initiative for interactions with Alexa is always with the user. Once a participant speaks the wake-word out loud, following with a command for Alexa matching an existing intent, a request is composed for the AWS Lambda function. Alexa handles the speech-to-text processing and is able to recognise special types such as numbers and dates in a sentence via slots[23], embedded in the the sample utterances as in figure 4.5. Words recognised as slots can be retrieved from the Alexa request by the Lambda function in the correct data type. In turn, the Lambda function requests physical activity information of the participant from the back-end and injects the obtained information in the response to Alexa. Embedded in this response is an object written in the Speech Synthesis Markup Language[24] (SSML). SSML can for example be used to define how fast text is read out loud and to embed audio files in the response. Alexa then converts this response to speech, which is in turn read out loud by the Echo Dot. A use case of information exchange is explained in the next section and visualised in figure 4.6.

[20] https://developer.amazon.com/docs/custom-skills/host-a-custom-skill-as-a-web-service.html
[21] https://developer.amazon.com/docs/custom-skills/host-a-custom-skill-as-an-aws-lambda-function.html
[22] https://github.com/alexa/alexa-skills-kit-sdk-for-nodejs
[23] https://developer.amazon.com/docs/custom-skills/slot-type-reference.html
[24] https://www.w3.org/TR/speech-synthesis11/

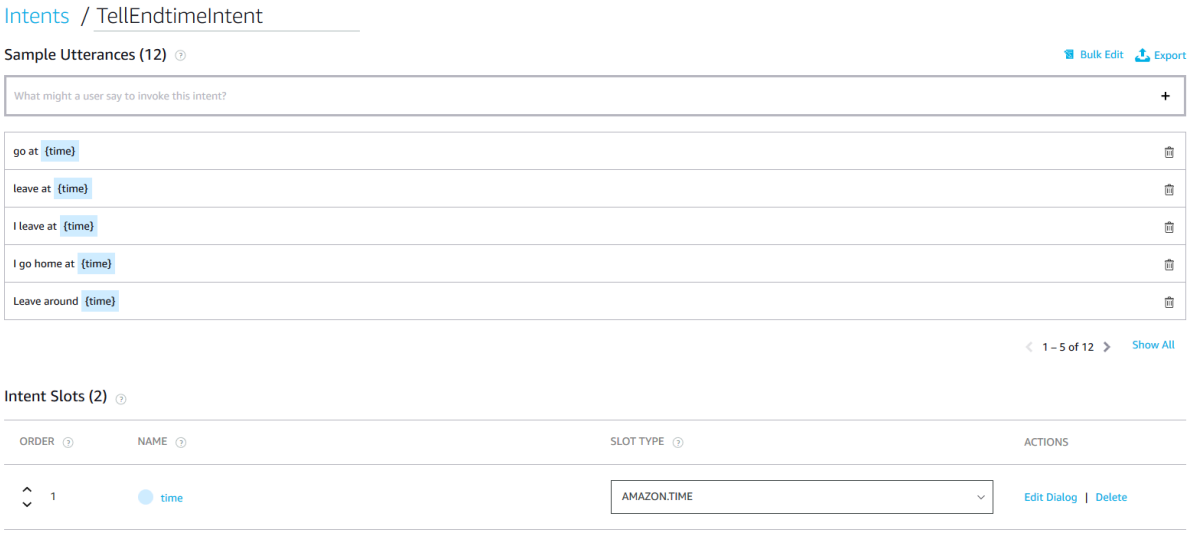


Figure 4.5: Example of slot types embedded in sample utterances, used when participant tell Alexa when they expect to go home

4.3.4. Use case: request remaining steps

Bob, one of the participants, is curious how much physical activity is still required to reach today's goal. Bob decides to ask Alexa how many steps remain for the current workday. With the sample utterance 'how many steps remain for today' Bob triggers the intent 'get remaining steps'. A request is composed and sent to the AWS Lambda function. The Lambda function receives the incoming request and determines which intent is triggered using the ASK SDK. It recognises the 'get remaining steps' intent, and therefore requests the remaining number of steps from the web server containing Bob's physical activity information collected by DeskstApp. This specific request is saved in the database for later analysis. A phrase containing the number of remaining steps for today is sent back to the Lambda function as a string and, if done correctly, forwarded to the custom skill captured in an SSML request, which processes the request and converts it to a speech result. The Echo Dot speaks the response out loud, informing Bob about his remaining steps for today.

4.3.5. User identification

Before the Echo Dot is able to provide physical activity information, they should be granted access to the back-end. For that, Amazon introduced 'account linking'[25]. This allows accessing content from an external resource provider via an external identity provider. In our case, the external providers are combined in the back-end, as discussed in section 4.1.1. Account linking can be performed in both the Alexa app or the Alexa website[26]. For an Alexa account to be linked, a public web page is required to which the participant can be redirected to provide his credentials. Therefore, we created a custom login page in the back-end. Upon submitting the login form, a code is generated by our back-end and both stored in the database and sent back to Alexa - also know as authorisation code grant[27]. Using this code, Alexa performs a request to the identity provider which exchanges the code for an access token and refresh token. Resources of the authorised user can now be accessed using the access token. This procedure conforms to the OAuth 2.0 standard, discussed in section 4.1.1.

4.3.6. Notifications

Participants may not have the incentive to converse with Alexa. On the contrary: prior work indicates that people choose not to use them regularly [12]. Initially, we intended to send motivational push notifications through the Echo Dot device. However, at the moment out user study was carried out, this feature was not available. There are ongoing discussions on whether to implement this feature[28] and discussions for

[25] `https://developer.amazon.com/docs/account-linking/understand-account-linking.html`
[26] `https://alexa.amazon.com/`
[27] `https://tools.ietf.org/html/rfc6749#section-1.3.1`
[28] `https://forums.developer.amazon.com/questions/8497/push-notifications.html`

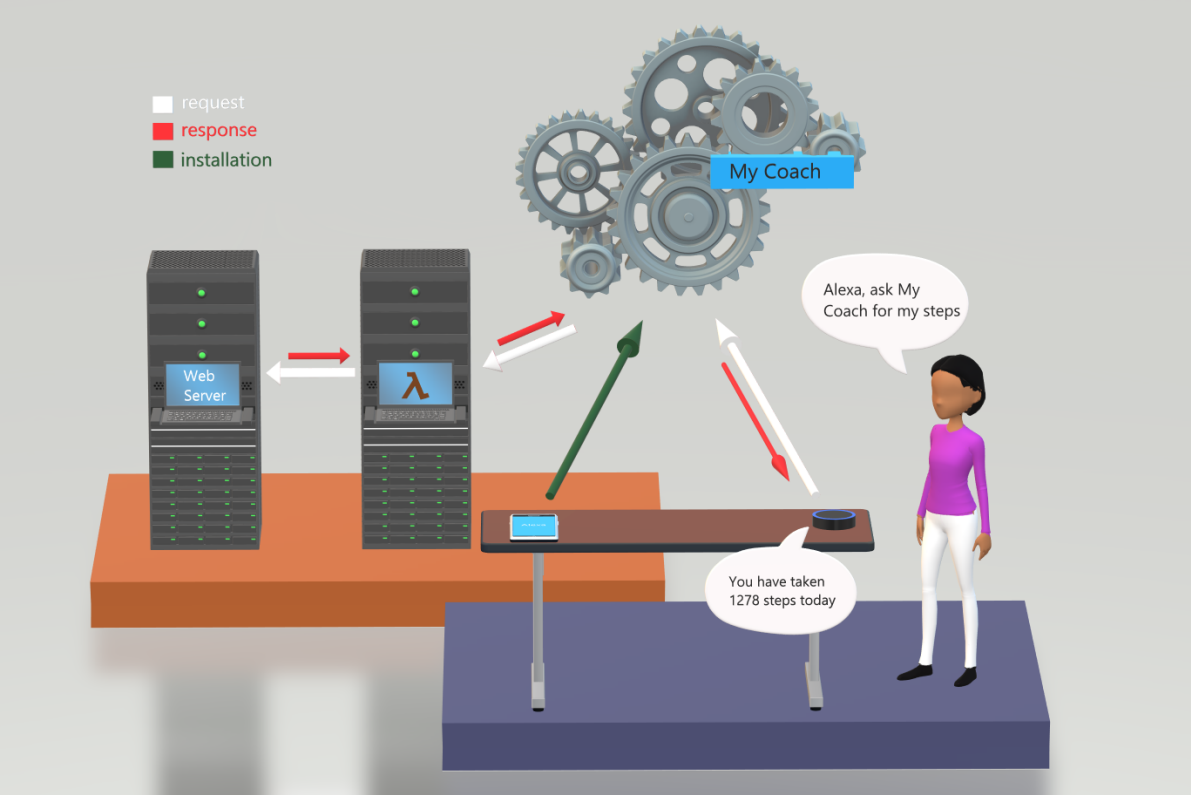


Figure 4.6: Components which together form Alexa & information exchange between the components

workarounds[29]. Since push notifications are unidirectional, we only required a text-to-speech service and not vice versa. To have compatible voice sounds, we looked into Amazon Polly[30]. However, available voices did not sound similar to Alexa. In addition, we tested Mycroft Mimic[31], which was just a single executable running on the web server, expecting a string as input and outputting an audio file. We used some of the sentences made in advance we intended to use for Alexa and converted them to speech by means of Mycroft Mimic. Unfortunately, it did not sound as sophisticated as Alexa and Amazon Polly. As a third option, there was Microsoft's Cognitive services text-to-speech[32]. Both Amazon Polly and Microsoft Cognitive services offered a free-tier service for a certain number of requests. Microsoft had an example project available in C# (our back-end programming language) which was almost immediately ready for use, so we chose Microsoft's cognitive services for our text-to-speech processing.

DeskstApp would ask the server every minute whether a notification is required for the participant. If that is the case, it would be determined whether they would still be on schedule. A notification message is composed of sentences in the following categories, if applicable to the participant:

1. Sentence for moving too little
2. Sentence for moving irregular
3. Sentence for moving sufficient
4. Sentence for completing the daily goal

Thus, an example notification for a participant moving sufficient but irregular would be "You're on the right track towards your goal. However, taking a few small breaks instead of one large break is a good start to break up your sedentary time". Irregularity is included because of Healy et al. [28], as we discussed in the Introduction. This notification was read out loud on the participant's smartphone instead of the Echo Dot. It was optional for participants to connect their smartphone to the Echo Dot, using the latter as an external speaker for the notification sound.

[29] https://forums.developer.amazon.com/questions/40799/workaround-for-push-notifications.html
[30] https://aws.amazon.com/polly/
[31] https://mycroft.ai/documentation/mimic/
[32] https://docs.microsoft.com/en-US/azure/cognitive-services/speech/api-reference-rest/bingvoiceoutput

4.4. Challenges

The development road towards the final system was full of difficulties and challenges. This section discusses the biggest obstacles that had most impact on the development and how they were overcome.

4.4.1. Ejecting from Expo

Building an app in React Native was supposed to ease cross-platform development for Android and iOS. This would probably have been the case if we did not have to deviate from the original plan. Initially, we started building our app using Expo[33], an open-source tool to allow development of pure React and JavaScript solutions without any restrictions on development OS's or app distribution. Via Expo we were able to make use of Google API's and Apple API's for accessing a participant's number of steps[34] without implementing a pedometer algorithm. However, an open feature request is to run background tasks with an application in Expo[35], functionality that is currently not implemented. Frequent updates is one of the requirements of our system, meaning that Expo is not a suitable solution. React Native has an alternative to aforementioned approach, namely building an app with native code. For that, we had to "eject" from Expo, meaning that we had the possibility to include plugins using 'native' code for Android and iOS. An advantage of this approach is that we could include React Native Background Fetch[36], a plugin that periodically allows callbacks in the background even if the app is closed, with a maximum of one callback every fifteen minutes. Disadvantages are that (1) the app no longer had access to the Expo package, including the Expo pedometer and (2) apps running native code require the installation of Android Studio for Android development and Xcode for iOS development. We will further elaborate on platform-specific development in the next section.

4.4.2. Strict Apple policies

For app development in React Native with native code, platform specific integrated development environments (IDEs) are required. For developing and testing the application, we had access to an Android phone, a laptop running Windows with Android Studio and the React Native CLI. However, for building the application for iOS devices, a development machine running MacOS is required. Initially we tried to run MacOS Sierra (10.12) on a virtual machine (VM) using Oracle VM VirtualBox, and when Xcode tools was unable to install we subsequently ran MacOS High Sierra (10.13) on a VM. Xcode tools could not be installed for those versions, because it required at least MacOS Mojave (10.14). Once running Mojave on the VM, it would not install Xcode tools because installation required a restart, and restarting would wipe the VM's applications. For testing the application for iPhone, we were able to temporarily adopt a MacBook Pro. Xcode tools offers a simulator for running the app on a virtual iPhone of choice. However, motion sensors like the accelerometer could not be enabled on the simulator, thus we needed an iPhone for the last part of the testing phase. For that, we could borrow one occasionally. Apple does not allow remote distribution of applications without a developer account of 99 USD per year. A free account is only allowed to distribute apps via physical access to the phone, with a certificate that is valid for at most seven days, which can be installed on at most three distinct iPhones. A workaround for the number of phones restriction was to create multiple iTunes accounts. Overcoming the first problem was done by giving the app a refresh installation every week for every participant using an iPhone.

4.4.3. Distribution of My Coach

In the past nine months, Amazon drastically changed the Alexa development interface. While in the previous version one had to perform some programming, the new version requires almost no programming knowledge. The new interface requires little training, but nonetheless the structure of a skill seems different compared to the old interface, which caused us to restructure our skill. However, a convenient feature in the new interface is the ability to perform beta-testing. A public custom skill is usually distributed via Amazon and available for every person owning an Alexa-enabled device. Beta-testing allows specific e-mail addresses to receive an invite code for the custom skill. Accepting this invitation will show the custom skill in the Alexa app, available for use. An advantage of this approach would also be that custom skills are instantly updated during the beta-testing phase, whereas a publicly distributed custom skill should be re-published before any changes are applied.

[33] `https://expo.io/`
[34] `https://docs.expo.io/versions/latest/sdk/pedometer`
[35] `https://github.com/expo/expo/pull/2338`
[36] `https://github.com/transistorsoft/react-native-background-fetch`

4.4.4. Installing the Echo Dots

We were in the possession of five Echo Dots - two in Estonia and three in the Netherlands. Those numbers do not cover the number of participants, thus we could not hand every participant a unique Echo Dot. This means that devices were moved around quite often. Every Echo Dot requires a link to a unique Amazon account, because if two devices are linked to the same Amazon account they can not be linked to separate back-end accounts. We created a unique Gmail or Hotmail e-mail address for each Echo Dot and created an Amazon account for each of these e-mail addresses. Echo Dots would be moved between participants either the Friday before the Alexa reflection phase or the Monday of the Alexa reflection phase, and as they moved we also unlinked the Amazon account from the back-end account of the previous participant, and linked them with the back-end account of the current participant. For Estonia, accounts were unlinked and linked remotely.

Every Echo Dot requires an internet connection via Wi-Fi. As the experiment ran mostly in university buildings, Eduroam (education roaming)[37], a world wide educational Wi-Fi network, was the intended network of choice for connecting the Echo Dots. However, this network (WPA2 Enterprise) is not supported and not recognised by the Echo Dot. As a solution, we requested Wi-Fi codes for the visitor network, a network intended for temporary guests of the university. These codes were valid for three weeks, which was sufficient for the experiment.

[37] `https://www.eduroam.nl/`

5

User Study

The collected data from the user study is the basis of our analysis for answering our research questions. This collection consists of sensor measurements, interactions with Alexa and survey reports. Sensor measurements and interactions with Alexa were provided by all sixteen participants. Fifteen participants provided feedback through surveys. This chapter will discuss how these data are analysed and how these may support us in answering our research questions. In section 5.1, the collected data is assessed on its reliability. Then, in section 5.2 we describe how participants interacted with Alexa (H1). Next, section 5.3 elaborates on the difference in physical activity levels between different phases is discussed. Finally, we describe to what extend we could provide answers on our research questions, and discuss the challenges we faced during the user study.

5.1. Data quality assessment

In order to draw conclusions based on the collected data, we ensured that it is **equal amongst participants** and **complete**. Equality of data amongst participants includes verifying that sensors of distinct phones reported similar values, i.e. that their sensors behaved in the same way. Testing completeness is about verifying whether the accelerometer-data and GPS-points are sampled sufficiently frequent, if participants enabled and disabled their workday on time and whether they carried their phone with them for every bout.

5.1.1. Sensor equality

As discussed in section 4.2.2, magnitude values of the accelerometer vector are sent to the server in batch. Storing each of these values separately would result in a relatively large claim on the available disk space. Instead, a summary of the batch data is stored in the database, reporting the mean (or average) of all values in the batch, the maximum value, the minimum value and the number of elements in the batch as in figure 5.1. As a way of assessing the quality for each of the smartphone accelerometer sensors, we compare the reported summaries between different users (thus different phones). In section 4.2.2 we have seen that accelerometer measurements are expressed in m/s^2. If the phone does not move, it is expected to report only the gravitational acceleration, because no other force is applied on it. Since our participants are office workers who are most of the time in sedentary state, we expect that the median of magnitudes of the accelerometer vector represents a phone that is not moving, thus being approximately equal to the gravitational acceleration. In addition to accelerometer data, the participant's location is supposed to be frequently reported by the app. GPS data is compared on its accuracy and frequency of reports.

accelerometer sensor analysis is carried out in the following way: we collected all summaries of accelerometer sensor reports per user and for each collection we calculated the median of average, maximum and minimum values. The median is preferred over the mean, because it ignores outliers and values generated while the participant was moving (which would result in a higher magnitude of the accelerometer). We found that the average (μ) of these median values is 9.85, which deviates 0.04 from the gravitational acceleration of ≈ 9.81 [69]. The median of average values and maximum and minimum peak values per user are depicted in figure 5.2. Figure A.1 specifically zooms in on the median of all reported average values. For minimum values ($\mu = 9.14, \sigma = 0.71$) and maximum values ($\mu = 10.71, \sigma = 0.97$) we can observe that the variation is larger than for the average magnitude values. However, this is mostly due to three Android users that report both high maximum values and low minimum values, which might indicate that their smartphone

sensors are not well-calibrated. The majority of the accelerometers report magnitude values between 8.5 and 11, which is sufficiently accurate as input for our algorithm.

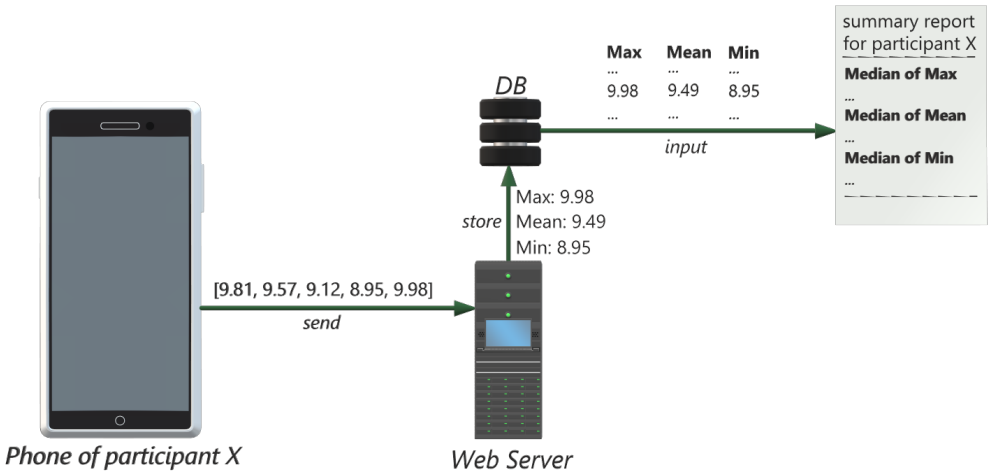


Figure 5.1: Example of capturing accelerometer data and transmitting it to the server in batch, where a record is created summarising that batch. From a collection of this summaries, we calculated $\widetilde{max}$, $\widetilde{mean}$ and $\widetilde{min}$ accelerometer values for each user.

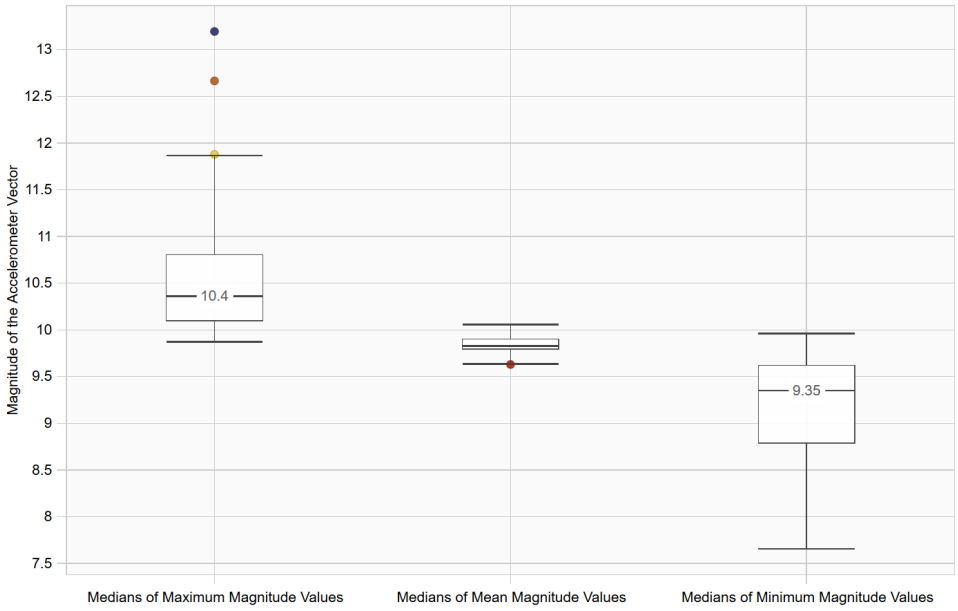


Figure 5.2: Boxplot of the medians of minimum, average (mean) and maximum magnitude values, extracted from summary reports for each participant

GPS sensors are compared on the 'accuracy' attribute reported by the participants' smartphones (see table 4.1). Thus, for all collected GPS data, we determined the variance in accuracy values by calculating the interquartile range (IQR), removing outliers that are 1.5·IQR away from the median and determining the accuracy in meters of the produced data points. In this data set without outliers, we observed that 62.39% of the location data points had an accuracy of exact 65 meters, all generated by iPhones. Additionally, 26.60% of the measured locations had an accuracy value smaller than 65 meters, of which 88.43% was generated by iPhones. These numbers indicate that Android phones are underrepresented in terms of location data reports. This is supported by figure 5.3, which shows an example of two representative participants where accumulated GPS data over three weeks are plotted on a map. It can be observed that iPhone smartphones reported significantly more location data.

(a) Location frequency for Android

(b) Location frequency for iPhone

Figure 5.3: Location frequencies for two representative participants. This is the office in Holland where most participants worked.

Summarising the observations of the mechanical differences between distinct smartphones:

- The average accelerometer values are similar between participants and slightly, but not significantly higher than we expected.
- The variance in reported values from accelerometer sensors between participants is fairly small, except for three outliers (participants)
- iPhone smartphones reported significantly more location data compared to Android phones

5.1.2. Completeness

Disciplinary behaviour of participants is involved when assessing completeness of the data, such as the habit of taking their smartphone with them when leaving their desk and whether they would toggle the start/stop workday switch in the app. According to the self-reports of participants, 26.7% (4) of the participants always carried their phone with them. 53.3% (8) of the participants often took it with them. The remaining 20% (3) reported that they carried it sometimes or occasionally.

Recapitulating from section 3.3, participants were given the task to start and stop their workday manually to set a fixed time frame for which their physical activity would be recorded. According to self-reports from participants, most of the participants never or occasionally forgot to perform these tasks as shown in table 5.1. On average, participants were especially forgetful on starting their workday. However, according to the information on the server, 73.3% (11) participants forgot to switch the workday off in the app at least once, indicating that at least three participants who filled in 'never' should be assigned to 'occasionally'. The reports for the 'sometimes' and 'often' categories are in accordance with the numerical data, as can be observed in table 5.2. In total, 79.3% of all started workdays were finished correctly by the participants.

Table 5.1: Percentage of participants reporting the frequency for which they forgot to start or stop their workday in the app

I ... forgot to	**start my workday** (% of participants)	**end my workday** (% of participants)
Never	26.7% (4)	46.7% (7)
Occasionally	40.0% (6)	33.3% (5)
Sometimes	26.7% (4)	13.3% (2)
Often	6.7% (1)	6.7% (1)

Table 5.2: Percentage of participants that forgot to end their workday according to the data available on the web server

% of participants	№ of times
13.3% (2)	1
6.7% (1)	2
20.0% (3)	3
13.3% (2)	4
13.3% (2)	5
6.7% (1)	7

During the experiment, some participants sent us feedback indicating that the app for Android phones did not always behave as expected. These had especially to do with the app reporting a lower number of steps than they expected. In response, we analysed the raw accelerometer data coming from Android phones and whether any particularities could be observed. As a result, it was found that:

1. The sample rate of the accelerometer sensor for Android phones was often too low
2. The application was sometimes shut down

Briefly, why (1) results in a lower number of steps is due to the following reason: the algorithm is built for a certain sample rate of accelerometer values. The less values are reported by the accelerometer, the lower the calculated number of steps is. This is explained in detail in 5.5.1. (2) could have multiple explanations. First, it was found in the Android documentation about life cycles of processes and applications[1] that "an application process's lifetime is not directly controlled by the application itself. Instead, it is determined by the system through a combination of the parts of the application that the system knows are running, how important these things are to the user, and how much overall memory is available in the system." If there is insufficient memory available, and another application requires memory, activities and services may be throttled or closed by the operating system. Additionally, Android will restrict sensors running in the background[2] for phones running Android 9. For some participants, the app would often only report accelerometer values +- every 15-20 minutes for a period of 30-60 seconds. This is due to the React-native-background-fetch plugin[3], which allows an app to be active in the background at most every fifteen minutes. How often an app is woken up in the background is determined by the grace of the operating system. For other participants, the app would sometimes entirely stop reporting data. Besides the array of potential technical reasons that are possible explanations for data loss, it could also be the case that participants accidentally closed the application themselves. Closing the application would remove all sensor listeners. Therefore, participants were urged to keep it open, either in the foreground or background.

Concluding, Android phones in general reported between 0-70% of the expected data. Therefore, they are unfit to serve as trusted data source for proper statistical analysis. For comparing the number of steps taken in the Alexa reflection phase and app reflection phase we therefore decided to only take into account data from iPhones. iPhone smartphones generally produced the expected number of accelerometer reports, on which section 5.3 will elaborate.

Summarising the findings related to the completeness of data:

- Approximately 80% of the participants reported they often carried their phone with them
- According to self-reports, on average participants occasionally forgot to start or stop their workday in the app
- Self-reports of participants are relatively well-aligned with data on the server
- Data from Android-phones is incomplete, which had two potential causes: (1) a low sample rate and (2) the fact that the app is sometimes closed either by the OS or participants
- Given that Android phones reported few location points and provided incomplete accelerometer data, only iPhones will be taken into account for further analysis

[1] `https://developer.android.com/guide/components/activities/process-lifecycle`
[2] `https://developer.android.com/guide/topics/sensors/sensors_overview`
[3] `https://github.com/transistorsoft/react-native-background-fetch`

5.2. Interactions with Alexa

Interactions with Alexa consist of user-initiated *conversations* and Alexa-initiated *notifications*. Only conversations with the custom skill My Coach could be recorded due to restrictions in the software of Alexa. The number of conversations differs very much between users, as can be seen in 5.4. Two of the participants conversed respectively 25 and 28 times with Alexa in a single week, whereas on the other side of the spectrum, three participants initiated zero interactions. The number of notifications varies between participants because it depends on the length of the workday and the up time of the application.

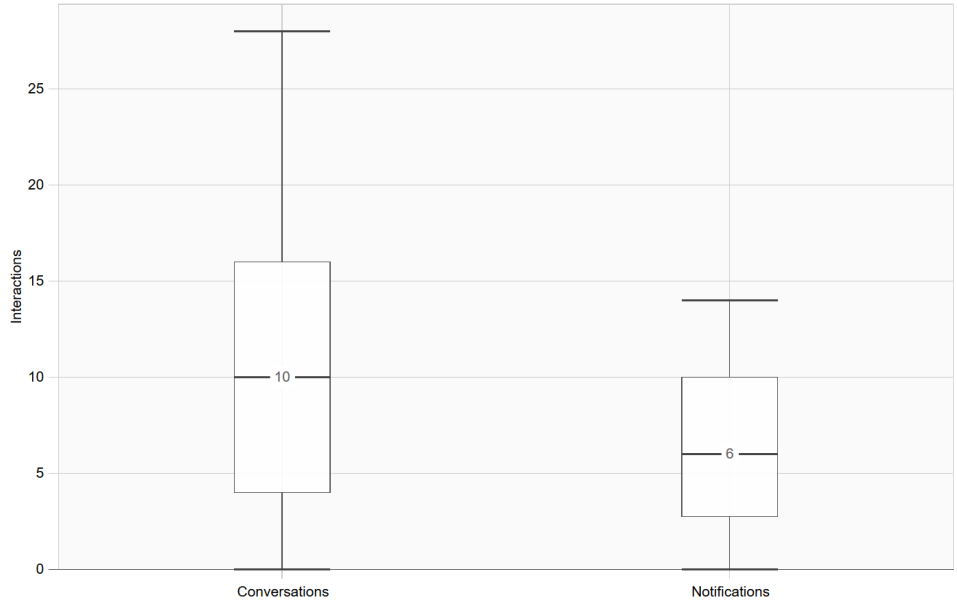


Figure 5.4: A boxplot visualising the № of conversations with and notifications from Alexa

In section 4.3.2 it was explained that a conversation starts with a voice request from a participant, triggering an 'intent'. We then categorised intents into the categories *informative*, *attractive* and *functional*. As a functional feature, participants could tell the system what time they would go home such that notifications would provide tailored information adapted to the situation, e.g.: if a participant would only work four hours, there is less time to reach a certain goal than if a participant would work for eight hours. Secondly, there are informative intents that provide the participants information about their physical activity. To make interactions with Alexa more attractive, two such features were added. Table 5.3 shows attractive features were least used by participants, requested by half of the participants. The functional intent, however, is more often used, on average approximately three times per participant. This was the only intent we explicitly asked participants to trigger. Informative intents were most often requested by participants: on average seven times per participant. The number of conversation intents resonates with self-reported data from participants: 80% (12) of them reported they used Alexa once a day, 13.3% (2) of them 2-3 times a day and one participant never used the available functionality.

Table 5.3: Distribution of conversation intents amongst participants and the total number of initiated conversations

Intent	№ of initiated conversations	№ of distinct participants	Category
Get № of steps	94	12	Informative
Start workday	44	12	Functional
Get remaining № of steps	17	6	Informative
Request stimulative song	11	7	Attractive
Get a fun fact	7	3	Attractive

On a 5-point Likert scale, 46.6% (7) of the participants found talking to Alexa moderately (4/5) or very (5/5)

intuitive. Another 46.6% found it neither intuitive nor counter-intuitive (3/5). However, 60% (9) of the participants found talking to Alexa in the office with colleagues uncomfortable, something that was also mentioned in prior work [12, 18]. One of the participants told us *"I felt a bit uncomfortable while giving voice commands to Alexa as I thought that it disturbed my colleagues (open-office). I probably would have used Alexa's other functionalities much more if I would have had it at home."* In addition, participants were generally satisfied with the way Alexa understood their commands. 66.7% (10) of them reported that Alexa always understood their commands and 26.7% (4) experienced Alexa understood them, but not always. Overall, on a 10-point Likert scale participants rated their overall experience with Alexa with a 6.4, varying from three to ten.

Recalling from section 4.3.6, notifications were sent through the mobile application. Due to partial inactivity of the Android version of DeskstApp, notifications were suppressed and therefore the number of notifications varies between participants. In total, the system registered that 86.7% (13) of the participants received at least two notifications and 46.7% (7) of the participants received ten or more notifications. However, according to self-reports 53.3% (8) participants perceived notifications. A potential reason could be that their volume was too low to hear the notification, or that they were away from their phone. Six out of eight participants would find notifications not motivational and none of the participants found that notifications stimulated them to talk with Alexa. On the other hand, five out of eight participants did not find the notifications disturbing. Most notifications were about stimulating insufficient or infrequent activity. This means that participants often were not on schedule for their goal. In the next section we take a closer look at why this could be the case and we compare the number of steps taken in each phase.

Summarising findings related to interactions with Alexa:

- The Nº of conversations with Alexa differs significantly between participants
- Participants mostly requested informative conversations (i.e. asked for their Nº of steps), in contrast to attractive features that were least used
- On average, participants found talking to Alexa intuitive, but uncomfortable when other colleagues were around
- Notifications were not experienced as motivational by most participants

5.3. Comparison of performed physical activity between phases

For comparing results of the Alexa reflection phase and app reflection phase (section 3.3) results from the Android version are excluded in this section because they reported insufficient data. Thus, this section reports data from the remaining four participants, consequently referred to as participant 1, 2, 3 and 4. For them, it was found that there are sufficient accelerometer data reports to cover some days (see figure A.2). In addition, it was perceived by these participants that the application accurately predicted their steps.

5.3.1. Evolution of self-set goals over time

Participants were able to set their own goal, as advocated by the goal-setting theory. They would start with a default value of eight thousand steps and urged to reflect on it every day, and update it if they noticed it would be too high or too low. It was found that a goal of eight thousand steps is perceived as too high. Table 5.4 shows that every participant lowered it in the course of the experiment. Some of them gradually lowered their goal and some of them changed it once during the experiment. Three of the participants found the available range of goals sufficient, whereas one of the participants wanted to set a goal lower than two thousand steps.

Table 5.4: Goal for each participant at the start and end of the experiment

Participant	Goal at the start	Goal at the end
1	8000	3000
2	8000	7000
3	8000	2500
4	8000	5500

5.3.2. Physical activity per phase

In section 3.4.2 it was determined that the most important physical activity metric is a participant's number of steps. This number should then be verified by the available GPS data, as discussed in section 3.4.3. For each of the four participants, we first determined which working days would qualify as days with sufficient data reports. Therefore, we included only working days where at least 75% of the expected number of accelerometer data was reported. In addition, workdays shorter than six hours were excluded to create a fair comparison. Activity performed on those workdays were then categorised per phase. For each day of every participant, we excluded physical activity where GPS data indicated a user would travel faster than 10 km/h. For example, participant 4 once forgot to disable their workday and left the application on when travelling by train. Our filtering method filtered out the falsely detected physical activity in the train, and kept detected activity during the short bouts from the office to the railway station of departure, and from the railway station of arrival to the presumed destination. Because participants sometimes forgot to stop their workday in the app, we also applied a distance filter on the results. If a participant would move more than five kilometres away from the office, measured physical activity would be excluded, such that only physical activity during working hours and near the office are measured.

For each phase, active days per phase were collected and the mean number of steps per participant was calculated, visualised in figure 5.5. It can be observed that for participants 2 and 4, there is no data available for the app reflection phase. This is due to the fact that these participants worked in the office for less than five days and the remaining data reports were of insufficient quality for these participants. Most of the missing data is due to troubleshooting interventions for participants with iPhone smartphones for reasons explained in section 4.4.2. In the next section we will evaluate our research questions and to what extend the results are supportive in answering them.

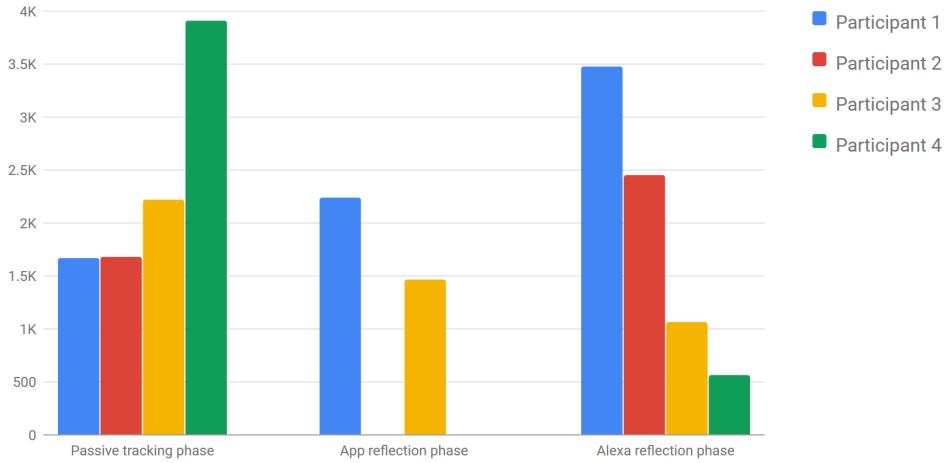


Figure 5.5: Mean absolute № of steps taken (y-axis) for days longer than six hours where at least 75% of the expected accelerometer-data was collected

---

Summarising findings related to the comparison of performed physical activity between phases:

- Only data from four participants using iPhones is included for evaluating their goals & № of steps
- Participants lowered their goal during the experiment, either gradually or at once
- Physical activity performed more than five kilometres away from the office or while travelling at high speed is removed from the final results
- № of steps taken per phase are visualised for four participants

5.4. Evaluation of the research statement

For our study, we investigated how proactive intelligent personal assistants affected a participant's physical activity levels, especially compared to activity trackers (**RQ1**). In addition, at the end of the experiment, participants were asked to indicate whether they felt the need to take more active breaks during their workday in the future (**RQ2**). They also reported how often they would perform sports outside working hours, such that we could observe possible differences in those who have an active lifestyle and those who do not (**RQ3**). In addition, it was hypothesised that participants would probably not often use Alexa (**H1**) and that activity trackers (i.e. the app reflection phase) would not lead to a significant increase in number of steps (**H2**).

RQ1 We intended to use data from all sixteen people participating in our experiment for performing non-parametric statistical tests in order to answer RQ1. Four out of sixteen participants provided sufficient data, which we consider too little for proving the effectiveness of Alexa compared to activity trackers. Due to the combination of (1) the small number of participants (2) insufficient data and (3) habits of participants such as forgetting their phone, it is impossible to draw conclusions on whether Alexa can be considered 'more effective' than activity trackers. Therefore, this question remains unanswered. However, we can still learn from the process of developing the system and executing the user study. For example, aspect (3) gives us insight in whether smartphones are suitable devices for measuring physical activity. Also, for (2) we found that accelerometers are relatively stable sensors, but not suitable for running in the foreground for a longer period of time. These are insights for future work that will be addressed in section 6.4.

RQ2 As discussed in the introduction, the duration of the experiment is too short for measuring long-term integration of physical activity in people's daily lives. However, 46.6% (7) of the participants reported they intend to use an app or activity tracker for monitoring their physical activity, based on the experience of the user study. It could be the case that some participants misinterpreted this question, as beforehand 66.7% (10) of the participants were already using activity trackers and fitness apps. In addition, 73.3% (11) of them reported they feel the need to take active breaks based on the experiment. This indicates that participants in general have a positive attitude towards taking active breaks.

RQ3 This question remains unanswered for the same reason RQ1 remains unanswered.

H1 Based on the observed number of interactions with Alexa, it can be concluded that the number of interactions varies per participant: some of them used Alexa regularly, some did not initiate any conversation at all. Depending on how one interprets 'occasionally', this hypotheses is either accepted or rejected. On average, participants had ten interactions a week, which we consider more often than occasionally. A possible explanation is that for our user study we used a headless device, whereas in Cohen et al. [9] they investigated IPAs on smartphones. Also, perhaps IPAs gained more popularity since the study performed by Cohen et al. [9] - 60% (9) of our participants were already familiar with intelligent personal assistants. Another explanation is that Alexa was used relatively often for one week, because it was perceived by participants as a novel feature. Perhaps its use would stagnate if the experiment would have lasted for a longer period of time.

H2 This hypothesis can neither be accepted nor rejected, as it depends on results from RQ1.

Besides results answering directly to our research questions, we observed elements in the behaviour of participants during the experiment that are nonetheless interesting to mention. In the next sections we will evaluate the tasks that are handed to participants, and the feedback we received on our experiment.

5.4.1. Tasks for participants

In advance, participants were requested to perform several tasks during the experiment such as carrying their phone at every bout, starting and stopping the workday in the app and telling Alexa what time they expected to go home. As presented in section 5.1.2, participants reported they often perform such tasks. Advantages of tracking methods that are 'framed' or 'labelled' by participants, is that they exactly know when their workday started and stopped, where automated methods may fail to do so. However, this could also result in unrepresentative physical activity data. Therefore, future work could address methods that do not merely rely on daily or weekly tasks by participants.

5.4.2. Feedback from participants

Communication with the participants was a significant part of the time we spent on the user study. We received over one hundred e-mails from participants. Most of them were about (1) arranging a meeting to hand over instructions and Echo Dot devices or (2) troubleshooting the application. In contrast, some e-mails included useful feedback and advice about Alexa and DeskstApp.

One participant advised us that it would be good to use text push-notifications instead of voice-notifications. The self-reports of participants indeed found voice-notifications not stimulating. In addition, it was reported by a participant through e-mail that "*until this date I have not received any notifications*", which was an indication that the Android app was sometimes suddenly 'killed' without transmitting notifications. Other e-mails for example addressed the issue of connecting Alexa to the educational Wi-Fi network (Eduroam) and one participant commented on the interpretation of the use of the concept 'playing sports' in one of the survey questions: "*E.g. play sports typically refers to games with a ball, however most sports don't. I do sports but I don't play sports.*" Providing a platform for participants on which they can provide feedback on the experiment in any way is certainly recommended for future work.

5.5. Challenges

During the user study, many challenges were faced such as distributing the right software (app) and hardware (Echo Dot), monitoring multiple participants in parallel, and conversing with all the participants during the experiment and quickly respond to their problems. However, the biggest challenge we faced was keeping the app up and running for all the participants using Android phones. We thoroughly investigated why it did not always behave as expected and therefore we assessed DeskstApp its sample rate, which we will discuss in the next section.

5.5.1. Infrequent sample rate

The React Native package that was used for tapping values from the hardware accelerometer, requested an update interval for measuring acceleration points, i.e. a number in milliseconds for which the accelerometer should provide an acceleration value in m/s^2 (Android) or g (iOS) to the x, y and z direction[4]. iPhones used for this study indeed provided accelerometer-data in the right frequency. However, when subscribing to an Android accelerometer, its sampling rate suddenly varied. We observed that, despite the variation in reporting frequency, the 'real' frequency would never exceed the requested update interval, thus if S was the provided update interval, the accelerometer would produce values with a sampling rate $\geq S$. After conversing with the creators of the React-native-sensors package[5], it was pointed out that the Android documentation tells us that "the data delay (or sampling rate) controls the interval at which sensor events are sent to your application (...) The delay that you specify is only a suggested delay. The Android system and other applications can alter this delay."[6]. This lower sampling rate of the accelerometer generally results in a lower amount of recorded steps.

To examine the step detection quality of different sample rates of the accelerometer, we tested several sample rates three times. In an indoor environment, we performed a little walk consisting of exactly 50 steps, walking the same path every time on a flat surface. The device we used was a Motorola Moto G4 plus. As depicted in figure 5.6, the average number of steps decreases as the real sample rate decreases. An odd observation is that a requested sampling rate of 100ms produces a lower real sampling rate than a requested sampling rate of 150ms. An explanation for this could be the order in which we tested each of the requested sampling rates: we started with our original sampling rate (150ms), then went up (250ms, 400ms) and at last tested a sampling rate of 100ms.

Before performing the sample rate test, we asked for verification of our problem identification. After performing this test, the developers of the React-native-sensors package commented on our question that "it might have been a problem with having multiple accelerometers overwriting each others updateIntervals"[7] and that it should be fixed in the new version of the package. However, when requesting clarification on this statement, we found that it might be a different issue that does not apply to our system.

[4] https://github.com/react-native-sensors/react-native-sensors/blob/master/README.md
[5] https://github.com/react-native-sensors/react-native-sensors/issues/163#issuecomment-433647022
[6] https://developer.android.com/guide/topics/sensors/sensors_overview#sensors-monitor
[7] https://github.com/react-native-sensors/react-native-sensors/issues/163#issuecomment-442439393

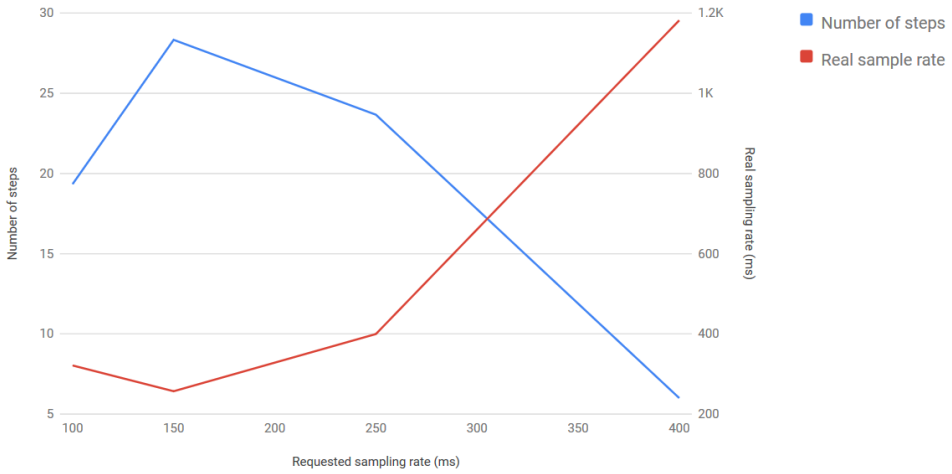


Figure 5.6: Relation between average real sample rates and number of steps in an experiment walking 50 steps

6

Discussions & Conclusion

Previously, we evaluated the results from the user study with respect to the research questions. In this chapter, we will further elaborate on questions raised by our study and the impact of the results. Additionally, we will reflect on its limitations and set out possible directions for future work based on our experiences and observations regarding the experiment. Concluding, we will once more reflect on the problem statement, our contribution to the problem and results from the user study.

6.1. Discussion

Our aim for the user study was to integrate physical activity of moderate intensity into daily routines of workers with sedentary jobs. In this section, we will reflect on the results from this user study and to what extend they contribute to the solution of breaking up sedentary time in the office.

Table A.1 shows there were fourteen different phone types in the experiment, used by fifteen participants. For assessing physical activity, we foremost relied on the values reported by the built-in hardware accelerometer for each of these phones. Since each accelerometer could behave in a different way, we analysed the *quality of the sensor measurements*. We found that the average values reported by these distinct sensors are more or less identical. However, the variance in reported values is slightly larger, especially due to three outliers (phones). This level of accuracy was sufficient for our step count algorithm. Based on the results from our study, smartphone accelerometers are considered sufficiently accurate for monitoring physical activity intensity. However, there are other limitations concerning smartphones that should be considered before adopting smartphones as main tracking device, as we will discuss in section 6.2.1.

Observing the interactions between Alexa and the participants, we found that the majority of all interactions are requests for the taken number of steps. Thus informative requests were preferred over the built-in attractive features. Also, some participants interacted with Alexa very often, and some participants never requested any information from Alexa. Because prior work indicated that intelligent personal assistants are used occasionally, or not at all by the majority of the users [9], we decided to configure Alexa such that it is proactive. For that, we used voice notifications. Participants reported they were not stimulated to either perform physical activity or interact more often with Alexa concerning their physical activity. Possibly, this is due to the fact that the notification feature was integrated in the mobile application instead of the Echo Dot, due to restrictions in the custom skill software. Qualitative results show that some participants indicated that notifications were too long, which was experienced as disturbing.

Prior work stimulated an approach based on the quantified self movement [77]. We made participants reflect on their number of steps as a single metric. However, one participant indicated that "*I missed more feedback (statistics) and information about what is considered healthy/unhealthy during working hours and why.*" Other participants indicate that "*the steps is not really help for me to do some practice, and most of my practice is done after work time*" and "*I usually take most steps from home to work and from work to home, but then the app is off...*". This indicates that breaking up sedentary behaviour is not yet a goal for some participants. Explaining participants why physical activity during working hours supports them in improving their health outcomes was done by means of notifications. Future work could address other ways of raising awareness for breaking up sedentary time, for example by presenting different metrics such as number of active breaks.

As we discussed in the Introduction chapter, we were particularly interested in measuring physical activity during working hours, because working hours are likely to contain sedentary behaviour. For that, we partially relied on the discipline of participants to carry out daily and weekly tasks. We found that participants most of the time remembered to execute daily tasks such as starting their workday and telling Alexa what time they would go home. However, for some tasks there is a small discrepancy between their self-reports and the collected data, where participants reported more executed tasks than they actually performed. From our side, it required a lot of task monitoring and we reminded participants to execute certain tasks if they seem to forget to do so. Afterwards, we filtered out physical activity that was not performed in the workplace. Future work may investigate the options of automatically determining the participant their location, to ease the burden on participants.

Concluding, we found that Alexa is generally well-perceived by participants. Talking to Alexa was experienced as intuitive, and the majority of the participants reported Alexa could understand them well. On the other hand, we also found that talking to Alexa is experienced as uncomfortable when other colleagues are around. Prior work indicated this may be caused by privacy concerns, especially when personally identifying information is involved [18]. Future work could address this issue by anonymising personal information without losing its relevance, for example by setting collective goals for the office cubicle. Besides the perceived inconvenience of talking to Alexa in the office, the majority of the participants found Alexa not necessarily more convenient than the application. Therefore, when designing for intelligent personal assistants in the office, these constraints should be taken into account. In section 6.4 we will further elaborate on potential research directions.

6.2. Limitations

6.2.1. Technical improvements

During the system development phase and user study we encountered a number of challenges, of which most issues were related to technical issues with the mobile application. In prior work, we can also observe that collecting and presenting data to participants was often troublesome: in Consolvo et al. [10], it was mentioned that a situation where the device failed to present the correct data caused "(...) frustration [which] often led to participants questioning if the device was malfunctioning or if they had accidentally broken it". Also, in Cambo et al. [3] problems concerning noisy sensing were experienced. For our study, we found that a persistent subscription on the accelerometer via the react-native-sensors plugin[1] caused the app to shut down at a certain point in time, which results in infrequent data reports and data loss. It was reported by one of the developers of this package that they detected a possible bug, causing update intervals of different accelerometers to overwrite each other. Future work using React Native software could solve this problem in several ways. At first, one could develop a self-made plugin for React Native which accesses the Google Pedometer via Google Fit their Recording[2] and Sensors[3] APIs and for Apple via the HealthKit API[4]. Another approach is to build and distribute the app in Expo, because processing tasks in the background is a feature currently in progress[5]. A third option is to make use of recently developed plugins for React Native that handle the communication between Google and Apple APIs for you such as React Native Apple Healthkit[6] and React Native Google Fit[7].

6.3. Threats to validity

For drawing reliable conclusions, one needs to eliminate threats to validity, i.e. factors beside the technical intervention that unintentionally may have caused participants to behave differently or produce biased results. We discuss the selection procedure of participants, the effect of Alexa as novel phenomenon, our experimental setup, and a small inequality in the two versions of our system, in order to provide insights on threats to validity of future work.

[1] https://github.com/react-native-sensors/react-native-sensors
[2] https://developers.google.com/fit/android/record
[3] https://developers.google.com/fit/android/sensors
[4] https://developer.apple.com/documentation/healthkit
[5] https://expo.canny.io/feature-requests/p/background-tasks-support
[6] https://github.com/terrillo/rn-apple-healthkit
[7] https://github.com/StasDoskalenko/react-native-google-fit

6.3.1. Selection of participants

The number of participants is comparable to that in similar studies [3, 10] and representative for qualitative measurements. In terms of resources we were restricted to this number of participants. For obtaining statistically relevant insights by applying, for example, the Mann Whitney U test[8], one needs more participants and more resources to provide each participant with. In addition, the majority of the participants had a technical background, and participants were between 22-46 years old. Future work could address the selection of participants in such a way that a representative set of workers is selected, covering more distinct ages and backgrounds.

6.3.2. The novelty of Alexa

The majority of the participants reported they did not use Alexa nor any similar IPA prior to the experiment. The novelty of Alexa could possibly influence the number of times it was used by participants to reflect on their physical activity. To overcome this potential threat, one could either (1) select participants who are already familiar with Alexa or similar IPAs, or (2) expand the duration of the experiment to detect whether the use of IPAs decreases over time.

6.3.3. Real-world situation versus controlled setting

For comparing activity trackers and Alexa in the workplace used by people with sedentary jobs, one could decide to carry out an experiment in a real-world setting or in a controlled setting. In our definition, a real-world setting represents a real-life situation without full control over external factors, whereas for a controlled setting, the environment is designed and heavily regulated. Compared to a real-world setting, a controlled setting allows a better oversight and regulation over all variables, and could for example be executed in an external location, which could be identical for each of the participants. In addition, identical activity tracking devices could be handed to each of the participants. In contrast, a real-world setting is an approach we chose based on prior work [77], which means the experiment is executed in a real office using the devices that belong to participants, casting little restrictions on their daily schedules and whereabouts. However, there are limitations of this approach compared to controlled settings:

- Participants sometimes forget their phone when leaving their desk
- Participants use smartphones of different types, which may behave in a different way
- The state of the out-of-office environment is not taken into account. For example, we did not verify if people would perform a different amount of exercise when it was raining outside, while it might affect the motivation for taking a walk
- Participants work different time lengths

In other prior work, it was decided to put more constraints on the experimental setup, in order to gain more control over environment variables [3, 10]. While these studies take place in a less realistic environment, controlled environments are able to overcome the aforementioned limitations. Therefore, future work could consider to design their experiment in line with studies focusing on control over environment variables.

6.3.4. Notifications

To ensure a fair comparison between the app and Alexa, notifications should be enabled for participants in the app reflection phase, because participants in the Alexa reflection phase receive voice-notifications. Our goal was to compare the effectiveness of Alexa with the effectiveness of activity trackers, and not necessarily the effectiveness of notifications. Due to time restrictions we were unable to implement push notifications for DeskstApp. There are React Native libraries[9] available for future applications.

6.4. Future work

6.4.1. Improving data collection

For our system, we explored several options for collecting physical activity data. Potential tracking devices were, for example, smart watches, smartphones or other accelerometer-enabled devices. We found that smartphones are suitable tracking devices for the reason that they are able to transmit data wireless to a web service in real-time, both indoors and outdoors, a function that most smart watches do not yet possess.

[8] http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/BS704_Nonparametric4.html
[9] https://github.com/wix/react-native-notifications

Because we were limited in resources, participants allowed us to use their smartphones as tracking devices. As software for monitoring their activity, we looked into prior work such as the Sense-IT mobile application [63]. However, this is an Android-only solution, whereas we wanted to make use of a cross-platform app for our participants in the possession of iPhone smartphones. The same holds for some other applications we investigated. Thus, we built our own application in React Native, using the hardware accelerometer sensor on iPhone and Android smartphones. However, especially data reported by Android phones turns out to be infrequent and insufficient. For Android 9 and later versions, accelerometers can not be used for long-term monitoring in the background[10].

To overcome these shortcomings, future work could consider using smartphone pedometers instead of accelerometers for tracking the amount of physical activity, however they do not report the activity intensity. How to use pedometers for React Native is explained in section 6.2.1. Because Android and iOS have different pedometer APIs, their accuracy should be compared.

We found that participants sometimes forgot their phone during bouts. If one has access to more resources, we recommend providing every participant with a dedicated tracking device worn on either the wrist, ankle, hip or thigh, or tied to a person's body in any other way. An example of such is the ActiGraph accelerometer[11], which allows tracking of multiple motion and physiological metrics, and is able to transmit data wireless nearly in real-time via Bluetooth[12]. However, a downside of this approach is that measurements are recorded all day, while participants should especially be urged to reflect on their physical activity during work time.

6.4.2. Evaluating the goal setting theory

We developed our system in accordance with the quantified self movement, advocated by prior work [77]. For physical activity quantification, we used the number of steps participants took as single metric of reflection. However, due to technical issues, some participants found this number to be inaccurate. To lead a person their attention away from technical imperfections, one could decide to abstract the physical activity metric, e.g. classify activity as 'little', 'sufficient' and 'good' [3]. On the other hand, the goal setting theory points out that specific goals are considered more effective [38]. Therefore, one should decide what level of accuracy a system is able to offer, and try to be as specific as possible to the user without losing their trust in the system.

Besides being specific, self-efficacy is an important aspect in goal setting theory. However, recent work by Konstanti and Karapanos [35] brings to light that 61% of their 81 participants adopted the default goal provided by the system. We observed that none of the participants using DeskstApp adopted the default goal. A possible explanation for this discrepancy with our results is that we urged participants to reflect on their goal on a daily basis. Additionally, all participants using our system lowered their goal over time, indicating that the default value was too high. This may have caused participants to update it. Future work could address the effect of self-set goals of different levels of difficulty on behaviour of participants, especially regarding the goal update frequency.

6.4.3. Sensing well-being

In section 6.1 we discussed the fact that more metrics could be used to make a person reflect on their physical activity and their number of breaks in sedentary time. In addition, section 2.3 elaborated on a number of studies involved in well-being assessment. Aspects of well-being such as cognitive focus, the absence of stress and attentiveness could respectively be measured by capturing audio and motion data [56], breath cycles [27] and mouse and keyboard interactions [16]. Besides this information being of interest to the worker as it improves health outcomes, well-being at work could also be in the employer their interest, as it decreases the number of workers that require sick leave [23]. Future work may investigate whether extra mental well-being metrics would result in an increased awareness of the importance of physical activity during working hours.

6.4.4. Context-awareness and interruptibility

Voice notifications sent by the system were transmitted with a fixed time interval between the notifications. An improvement of such is to make them context-aware, taking into account a person their *interruptibility* [65]. According to Iqbal and Bailey [30], interruptions such as notifications would best be served at *break points*: "moments of transition between two observable, meaningful units of task execution". Doing so, re-

[10] `https://developer.android.com/guide/topics/sensors/sensors_overview`
[11] `https://www.actigraphcorp.com/actigraph-wgt3x-bt/`
[12] `https://www.actigraphcorp.com/cdh/`

duces the cost of an interruption, such as frustration. Context-awareness is also beneficial for receiving tailored information from intelligent personal assistants. Amazon employee Sarikaya [62] tells us that the key promise of an IPA is to stitch together information from different sources, to ease the user their burden of going through different applications. In addition, proactive assistants continuously learn from feedback signals provided by the user. Future work could investigate to what extend intelligent personal assistants are able to provide the right information at the right moment.

However, for both assessing a person their context and providing the right information at the right time, one needs a lot of data. A question that is raised when systems know a lot about a person their whereabouts, is: to what extend can accuracy be exchanged for privacy? The next section elaborates on questions that future work could investigate regarding privacy of intelligent personal assistants.

6.4.5. Privacy

Using a commercial device such as the Amazon Echo Dot raises certain privacy concerns, such as the fact that it is a device that constantly listens to voice inputs until the wake-word is detected. These concerns could be overcome by a system ensuring privacy, such as the privacy-driven data model by Nogueira et al. [43]. Key aspects of such a system are transparency and data ownership by the user. Future work may investigate whether the attitude of people towards IPAs changes if such promises are made, and to what extend these are possible to ensure using an IPA.

Besides external parties collecting data, there could also be vulnerabilities in intelligent personal assistants, a realistic scenario as we have seen in the past [8]. To address that problem, IPAs should be thoroughly tested to reveal such vulnerabilities. Proper voice authentication could also provide extra assurance for privacy [21]. Future work using the Echo Dot should take into account the vulnerabilities of Echo devices. For example, unauthorised voice purchasing is a weakness in the current system, as products can be purchased without any verification and by any voice [25, 75]. In addition, the 4-digit PIN that can be added for restricted actions can be cracked by brute-force attacks. One could also decide to keep track of the development of open-source IPAs such as Mycroft[13], as these promise transparency and data-ownership by the user.

6.4.6. Future applications of IPAs

The development of IPAs is still ongoing, as we also experienced during our implementation phase of My Coach for Alexa. There is still much to gain in terms of resource consumption, as IPAs are mainly "fueled by the extraction of non-renewable materials, labor, and data" [31]. Prior work shed their light on the future of intelligent personal assistants, and which elements are important for its relevance in any system. Azaria and Hong [1] claims that IPAs should be personal, dynamic learners, supportive, affable (or kind) and at last instructable to be of use in a recommender system. Also, as a next step, the autonomy of an IPA could be increased if it can be ensured that actions executed by the IPA are approved by the user. However, it is believed that full autonomy of IPAs is not desired, as users want to keep some level of direct control over them. In contrast to Azaria and Hong [1], Cohen et al. [9] believes that the HCI community should especially focus on "useful, usable, delightful, caring, ethical, entertaining and educational personal assistants".

We believe that, considering the privacy issues in previous section and experiences from our user study, useful, ethical and educational aspects of personal assistants are especially important. An IPA should be *useful*, in the sense that it should have added value to a system compared to existing techniques. The strength of IPAs is the ability to tailor information together from multiple applications, which make them suitable for combining different information sources in complex systems where otherwise multiple applications are required. However, an IPA should also be *ethical*, thus taking privacy into account, act in the user their interest unless other people are harmed by it, and have limited autonomy unless ethical behaviour can be assured. At last, for our study we wanted Alexa to be especially *educational*, in the sense that participants would learn something about their own behaviour. Whereas some evidence indicates that the application of intelligent personal assistants for educational purposes are successful [15], the added value of IPAs for such needs to be investigated. Future work could take these considerations into account when building applications with intelligent personal assistants.

6.5. Conclusion

First world countries struggle with health-related issues due to people being physical inactive [36, 46]. In particular, sedentary behaviour is widely integrated in the western lifestyle. Evidence indicates that more

[13] `https://mycroft.ai/`

interruptions in sedentary time is positively associated with a reduction of metabolic risk [28]. An environment where sitting often occurs is the workplace. The World Health Organisation [47] acknowledges physical inactivity during work hours as a problem, and proposes a more healthy lifestyle for both employers and employees, respectively by allowing flexible working hours and encouraging walking and cycling. The latter are examples of physical activity of moderate intensity, which has convincing health benefits [46].

Activity trackers are often used as a means of reflecting on physical activity. However, a large-scale study indicated that activity trackers are not very effective, especially if there is no additional incentive such as cash [22]. Therefore, we aimed for a different approach, using an intelligent personal assistant (IPA). A key promise of IPAs is that they are able to combine data from different sources to provide tailored information. We used the Amazon Echo Dot as headless device serving Alexa, Amazon their personal assistant. Cambo et al. [3] inspired us to develop a solution for the workplace to break up sedentary time. The role of Alexa in our system was to proactively interrupt participants, providing them updates on their physical activity. To measure the effectiveness of Alexa for stimulating physical activity, it was compared to the effectiveness of a self-made activity tracker app (DeskstApp) for iPhone and Android phones. Zuckerman and Gal-Oz [77] advocated the quantified self approach[14], feeding a person their intrinsic motivation by making them reflect on physical activity metrics. Alexa and DeskstApp presented the number of steps as single metric, respectively by voice or as text to the sixteen participants. Participants were allowed to set a daily step goal, which has promising effects on a person their motivation, according to the goal setting theory [38]. For one week, half of the participants used Alexa, and the other half reflected on their physical activity in the app, and vice versa the week after. Besides investigating the differences between the two versions of the system, we were interested in the integration of moderate physical activity into daily routines, which we assessed in a qualitative way. In addition, we investigated whether people who perform physical exercise outside working hours would behave in a different way compared to people who do not.

At the beginning of the experiment, we faced some technical challenges. At first, the mobile application for iPhone smartphones were hard to distribute remotely. Remote distribution was a requirement, because half of our participants worked in Estonia and the other half in the Netherlands. In addition, Apple disallows free apps to be installed longer than seven subsequent days. Therefore, participants in the Netherlands using iPhone smartphones required weekly app updates. While distribution was mainly a challenge for iPhone smartphones, Android smartphones suffered from noisy data reports. We found that DeskstApp running on Android infrequently sampled data from the accelerometer, whereas this data is crucial to determining a participant his physical activity. In addition, there was much difference in the number of collected GPS values for both smartphone operating systems.

At the end of the experiment, we performed a thorough data quality analysis. We found that physical activity data for participants with Android smartphones was insufficient for drawing conclusions concerning physical activity. Prior work indicated that data collection using sensors is a challenging step, which could result in noisy sensor data and frustrated participants [3, 10]. However, we were able to observe interactions between participants and Alexa, and we found that most participants tend to make use of informative features compared to attractive features. On average, participants found talking to Alexa rather intuitive, and they perceived that Alexa often understood their requests correctly. On the other hand, using Alexa in shared offices with colleagues around was perceived as uncomfortable by 60% of the participants. Four participants with a reasonable collected amount of qualitatively good data, found the step detection algorithm to be accurate. We observed that they either immediately or gradually lowered their daily step goal over time. Regarding the performed physical activity, we were unable to draw conclusions of significance.

Future work relying on sensor-based data collection methods are advised to especially focus on obtaining qualitatively good data, for example using dedicated devices like ActiGraph accelerometers[15] that do not rely on disciplinary behaviour of participants. Additionally, one could make use of the built-in pedometer for Android an iPhone smartphones, however their accuracy should be assessed. Furthermore, one could raise awareness of the fact that breaks in sedentary time are beneficial for health outcomes. This could be done by providing additional metrics to the participant, such as number of active breaks. For applications using intelligent personal assistants, future work could investigate the effect of context-aware notifications on a person their cognitive focus and overall well-being. However, for that one needs a lot of data. Therefore, a challenge for future work is to design for privacy, without reducing the IPAs 'personality'. For example, by providing collective physical activity metrics (e.g. office cubicle) which anonymises physical activity data for individuals. We believe that future work should especially focus on useful, ethical and educational applications.

[14] http://quantifiedself.com/
[15] https://www.actigraphcorp.com/actigraph-wgt3x-bt/

A

User study and System design & validation

A.1. User Study

Table A.1: List of different phone types that were used during the experiment

iPhone	**Android**
iPhone 5c	Blackview BV7000
iPhone 6s	Huawei P10 lite
iPhone 7	Nokia 8
iPhone 7 plus	OnePlus 3
	OnePlus 5T
	Samsung Galaxy A8
	Sony Xperia X
	Sony Xperia X Compact
	Xiaomi Mi A1
	Xiaomi Redmi Note 5 pro

Table A.2: Operating systems & Nº of participants using that specific OS

iPhone		**Android**	
OS	Nº of participants	OS	Nº of participants
iOS 10.3.3	1	7.0	2
iOS 11	3	7.1.2	1
		8.0	5
		8.1	3

A.2. System design & validation

Table A.3: Sample rates of the accelerometer in DeskstApp for Android while detecting 50 steps

requested sampling rate	detected steps	real sampling rate	measurements	duration (ms)
	18/50	314ms	108	33915
100ms	17/50	349ms	97	33856
	23/50	301ms	115	34594
	25/50	259ms	218	56543
150ms	30/50	252ms	133	33566
	30/50	260ms	162	42058
	25/50	402ms	88	35345
250ms	22/50	401ms	85	34115
	24/50	396ms	84	33253
	8/50	1041ms	32	33299
400ms	5/50	1152ms	29	33421
	5/50	1353ms	25	33820

Table A.4: Count of the perceived deviation in accuracy values

Accuracy range (m)	**Count**
0-24	5268
25-49	1152
50-74	17885
75-99	149
100-124	730
125-149	279
150-174	391
175-199	26
200-224	1273
225-249	22
250-274	35
275-299	6
300-314	11

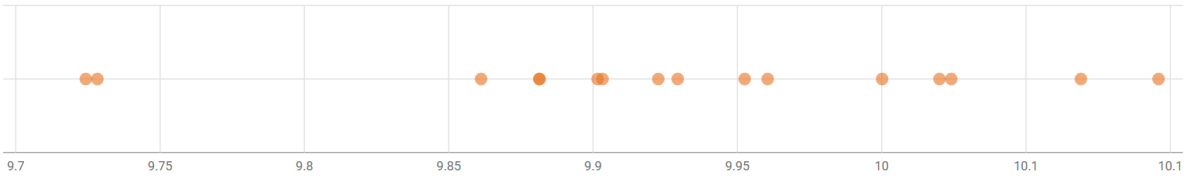


Figure A.1: Median of mean values from the summary reports, where each dot represents a participant, $\mu = 9.85, \sigma = 0.10$

Algorithm 1 Pedometer Algorithm

Require: AVG, function that calculates the average value of a set of numbers
Require: ABS, function that takes a set and returns it s.t. every value is transformed to its absolute value
Require: MINPEAKVALUE, function that determines the miminum required peak height
▷ M is a set of magnitude values (m) and their corresponding timestamps (t), ordered on timestamp

```
function STEPS(M)
    minVal ← MINPEAKVALUE(M)                    ▷ Determine minimum required peak height (see 2)
    minDiff ← 0.31                              ▷ Minimum time (in seconds) between different peaks
    M_normalised ← ABS(M − AVG(M))              ▷ Normalise the set by extracting its average
    t_lastpeak ← NULL
    peaks ← 0
    for i ← 1 to |M_normalised| − 1 do
        (m_i, t_i) ∈ M_normalised
        if m_i > minVal then                    ▷ If the current value is high enough to be a peak
            if m_i > m_{i+1} AND m_i > m_{i−1} then   ▷ If the current value peaks
                if t_lastpeak = NULL OR t_lastpeak − t_i > minDiff then
                    peaks = peaks + 1
                    t_lastpeak ← t_i
                end if
            end if
        end if
    end for
    return peaks                                ▷ Return number of steps
end function
```

Table A.5: Mean absolute Nº of steps for days longer than 6 hours with a coverage of 75%

Participant	Steps phase 1	Steps phase 2	Steps phase 3
1	1675	2245	3483
2	1686		2458
3	2226	1472	1071
4	3916		570

Algorithm 2 Helper Functions

```
function MINPEAKVALUE(M)
    M_range ← ∅
    threshold ← 0.5
    max ← 9
    for each (m, t) ∈ M do
        if m > threshold AND m < max then
            M_range ← M_range ∪ {m}
        end if
    end for
    m_avg ← AVG(SD)
    return m_avg
end function
```

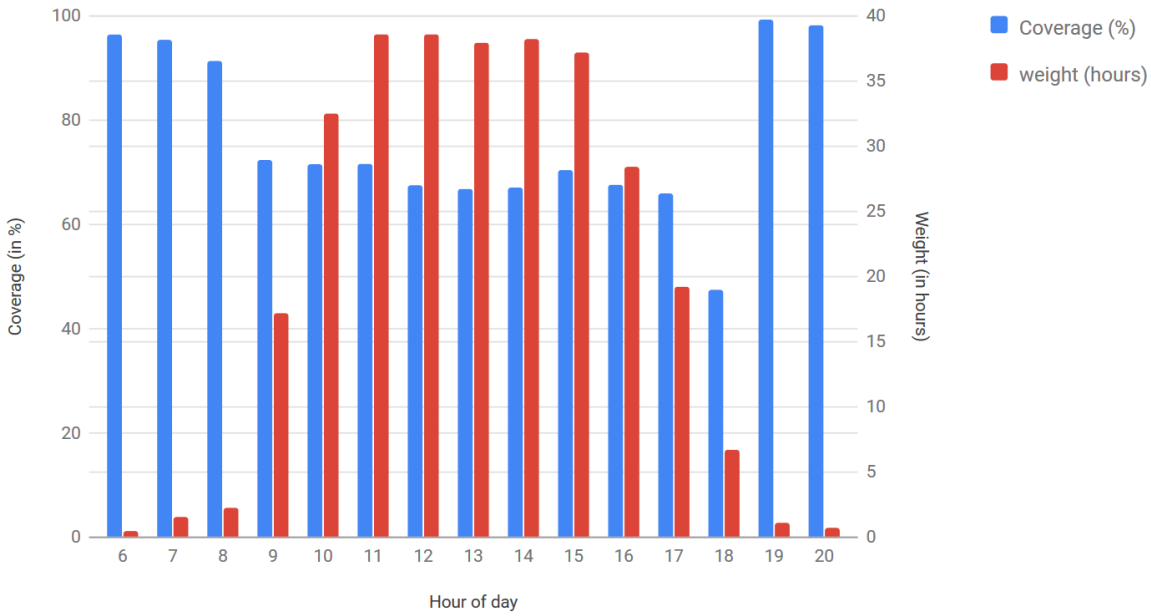


Figure A.2: Normalised coverage of accelerometer reports of every workday hour

List of Figures

1.1 Health effects of physical (in)activity for adults [46] . . . 2

2.1 40 subsequent recordings of magnitude show a sinusoidal pattern in the first half, which indicates the person is walking. From recording 25 and on, the person does not move, because the line is flattened and approaches the gravitational constant. Chart is based on figure 2 & 3 in Khedr and El-Sheimy [34], reproduced by DeskstApp, the application that monitors physical activity of participants as discussed in chapter 4 . . . 8

3.2 Different states of the app . . . 19

4.1 System components & interaction . . . 24
4.2 Visualisation of the register and login screens for Android, including the login screen displaying an error message . . . 26
4.3 3-axis accelerometer . . . 27
4.5 Example of slot types embedded in sample utterances, used when participant tell Alexa when they expect to go home . . . 31
4.6 Components which together form Alexa & information exchange between the components . . . 32

5.1 Example of capturing accelerometer data and transmitting it to the server in batch, where a record is created summarising that batch. From a collection of this summaries, we calculated $\widetilde{max}$, $\widetilde{mean}$ and $\widetilde{min}$ accelerometer values for each user. . . . 36
5.2 Boxplot of the medians of minimum, average (mean) and maximum magnitude values, extracted from summary reports for each participant . . . 36
5.3 Location frequencies for two representative participants. This is the office in Holland where most participants worked. . . . 37
5.4 A boxplot visualising the Nº of conversations with and notifications from Alexa . . . 39
5.5 Mean absolute Nº of steps taken (y-axis) for days longer than six hours where at least 75% of the expected accelerometer-data was collected . . . 41
5.6 Relation between average real sample rates and number of steps in an experiment walking 50 steps . . . 44

A.1 Median of mean values from the summary reports, where each dot represents a participant, $\mu = 9.85, \sigma = 0.10$ . . . 52
A.2 Normalised coverage of accelerometer reports of every workday hour . . . 54

List of Tables

1.1 Categories of physical activity intensity as defined in [2] and [46]. MET stands for metabolic equivalents, where one MET is the energy expenditure at rest . . . 3

2.1 Four groups of participants as in [22] . . . 9
2.2 Varieties of well-being [71] . . . 11

3.1 Benchmark tests of Mycroft & Alexa, using a Wi-Fi network of 52.0 Mbps down and 27.3 Mbps up, posing each question 5 times per device taking turns. Numbers in the first columns are approximations of average response times (RT), and the values in the second column indicate whether the IPAs provided a correct answer (CA). . . . 16
3.2 Details about experimental setup per round . . . 18

4.1 Specification of the location record . . . 28
4.2 Explanation of custom skill components . . . 29
4.3 Syntax of Alexa requests . . . 29

5.1 Percentage of participants reporting the frequency for which they forgot to start or stop their workday in the app . . . 37
5.2 Percentage of participants that forgot to end their workday according to the data available on the web server . . . 38
5.3 Distribution of conversation intents amongst participants and the total number of initiated conversations . . . 39
5.4 Goal for each participant at the start and end of the experiment . . . 40

A.1 List of different phone types that were used during the experiment . . . 51
A.2 Operating systems & Nº of participants using that specific OS . . . 51
A.3 Sample rates of the accelerometer in DeskstApp for Android while detecting 50 steps . . . 52
A.4 Count of the perceived deviation in accuracy values . . . 52
A.5 Mean absolute Nº of steps for days longer than 6 hours with a coverage of 75% . . . 53

Bibliography

[1] Amos Azaria and Jason Hong. Recommender systems with personality. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 207–210. ACM, 2016.

[2] Srinivasan Beddhu, Guo Wei, Robin L Marcus, Michel Chonchol, and Tom Greene. Light-intensity physical activities and mortality in the united states general population and ckd subpopulation. *Clinical Journal of the American Society of Nephrology*, pages CJN–08410814, 2015.

[3] Scott A. Cambo, Daniel Avrahami, and Matthew L. Lee. BreakSense: Combining Physiological and Location Sensing to Promote Mobility during Work-Breaks. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, pages 3595–3607, 2017. doi: 10.1145/3025453.3026021. URL `http://dl.acm.org/citation.cfm?id=3026021{%}0Ahttp://dl.acm.org/citation.cfm?doid=3025453.3026021`.

[4] Carl J Caspersen, Kenneth E Powell, and Gregory M Christenson. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public health reports*, 100(2):126, 1985.

[5] CBS. Internet; toegang, gebruik en faciliteiten. `https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83429NED/table?dl=12B8F`, 2017. Online; accessed 11 October 2018.

[6] Kong Y Chen and JR DAVID R BASSETT. The technology of accelerometry-based activity monitors: current and future. *Medicine & Science in Sports & Exercise*, 37(11):S490–S500, 2005.

[7] Kong Y Chen, Kathleen F Janz, Weimo Zhu, and Robert J Brychta. Re-defining the roles of sensors in objective physical activity monitoring. *Medicine and science in sports and exercise*, 44(1 Suppl 1):S13, 2012.

[8] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. Alexa, can i trust you? *Computer*, 50(9): 100, 2017.

[9] Phil Cohen, Adam Cheyer, Eric Horvitz, Rana El Kaliouby, and Steve Whittaker. On the Future of Personal Assistants. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*, pages 1032–1037, 2016. doi: 10.1145/2851581.2886425. URL `http://dl.acm.org/citation.cfm?doid=2851581.2886425`.

[10] S Consolvo, McDonald DW, T Toscos, Chen MY, J Froehlich, and B Harrison. Activity sensing in the wild: a field trial of UbiFit Garden. In: BT - Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08). pages 1797–1806, 2008.

[11] Kirsten Corder, Søren Brage, and Ulf Ekelund. Accelerometers and pedometers: methodology and clinical application. *Current Opinion in Clinical Nutrition & Metabolic Care*, 10(5):597–603, 2007.

[12] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. "What can i help you with?". *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '17*, pages 1–12, 2017. doi: 10.1145/3098279.3098539. URL `http://dl.acm.org/citation.cfm?doid=3098279.3098539`.

[13] Michael B. del Rosario, Stephen J. Redmond, and Nigel H. Lovell. Tracking the evolution of smartphone sensing for monitoring human movement. *Sensors (Switzerland)*, 15(8):18901–18933, 2015. ISSN 14248220. doi: 10.3390/s150818901.

[14] Deloitte. Global Mobile Consumer Survey 2017: The Netherlands. pages 1–29, 2017. URL `https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/technology-media-telecommunications/2017GMCSDutchEdition.pdf`.

[15] Néstor Darío Duque Méndez, Paula Andrea Rodríguez Marín, and Demetrio Arturo Ovalle Carranza. *Intelligent Personal Assistant for Educational Material Recommendation Based on CBR*, pages 113–131. Springer International Publishing, Cham, 2018. ISBN 978-3-319-62530-0. doi: 10.1007/978-3-319-62530-0_7. URL https://doi.org/10.1007/978-3-319-62530-0_7.

[16] Dalila Durães, Davide Carneiro, Javier Bajo, and Paulo Novais. Using computer peripheral devices to measure attentiveness, 2016. ISSN 21945357.

[17] Dalila Durães, David Castro, Javier Bajo, and Paulo Novais. Modelling an intelligent interaction system for increasing the level of attention. In *International Symposium on Ambient Intelligence*, pages 210–217. Springer, 2017.

[18] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. Voice activated personal assistant: Acceptability of use in the public space. In Sakae Yamamoto, editor, *Human Interface and the Management of Information. Information and Knowledge in Applications and Services*, pages 324–334, Cham, 2014. Springer International Publishing. ISBN 978-3-319-07863-2.

[19] Eurostat. *Health and safety at work in Europe (1999–2007)*. 2010. ISBN 9789279146060. doi: 10.2785/38630. URL http://ec.europa.eu/eurostat/documents/3217494/5718905/KS-31-09-290-EN.PDF/88eef9f7-c229-40de-b1cd-43126bc4a946.

[20] Rosa Angela Fabio and Giulia Emma Towey. Long-term meditation: the relationship between cognitive processes, thinking styles and mindfulness, 2017. ISSN 16124790.

[21] Huan Feng, Kassem Fawaz, and Kang G Shin. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, pages 343–355. ACM, 2017.

[22] Eric A. Finkelstein, Benjamin A. Haaland, Marcel Bilger, Aarti Sahasranaman, Robert A. Sloan, Ei Ei Khaing Nang, and Kelly R. Evenson. Effectiveness of activity trackers with and without incentives to increase physical activity (TRIPPA): a randomised controlled trial. *The Lancet Diabetes and Endocrinology*, 4(12):983–995, 2016. ISSN 22138595. doi: 10.1016/S2213-8587(16)30284-4. URL http://dx.doi.org/10.1016/S2213-8587(16)30284-4.

[23] Institute for Health and Productivity Studies. Physical activity in the workplace: a guide for employers. https://www.cdc.gov/physicalactivity/worksite-pa/index.htm. Online; accessed 30 November 2018.

[24] David Gimeno, Marko Elovainio, Markus Jokela, Roberto De Vogli, Michael G Marmot, and Mika Kivimäki. Do passive jobs contribute to low levels of leisure-time physical activity? the whitehall ii cohort study. *Occupational and environmental medicine*, 2009.

[25] William Haack, Madeleine Severance, Michael Wallace, and Jeremy Wohlwend. Security analysis of the amazon echo. 2017.

[26] T. Hall, N. A. Malone, J. Tsay, J. Lopez, T. Nguyen, R. E. Banister, and D. Y.C. Lie. Long-term vital sign measurement using a non-contact vital sign sensor inside an office cubicle setting. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2016-Octob:4845–4848, 2016. ISSN 1557170X. doi: 10.1109/EMBC.2016.7591812.

[27] Tian Hao and Roxane Chan. MindfulWatch: A Smartwatch-Based System For Real-Time Respiration Monitoring During Meditation. *Proc.ACMInteract.Mob.WearableUbiquitousTechnol*, 13(3):1–19, 2017. ISSN 24749567. doi: 10.1145/3130922. URL http://doi.org/10.1145/3130922.

[28] Genevieve Healy, David W Dunstan, Jo Salmon, Ester Cerin, Jonathan Shaw, Paul Zimmet, and Neville Owen. Beneficial associations with metabolic risk. *Diabetes Care*, 31(4):661–666, 2008. ISSN 1935-5548. doi: 10.2337/dc07-2046.Abbreviations.

[29] Jun-Ho Huh and Kyungryong Seo. An indoor location-based control system using bluetooth beacons for iot systems. *Sensors*, 17(12), 2017. ISSN 1424-8220. doi: 10.3390/s17122917. URL http://www.mdpi.com/1424-8220/17/12/2917.

[30] Shamsi T. Iqbal and Brian P. Bailey. Effects of intelligent notification management on users and their tasks. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, page 93, 2008. doi: 10.1145/1357054.1357070. URL http://portal.acm.org/citation.cfm?doid=1357054.1357070.

[31] Vladan Joler and Kate Crawford. Anatomy of an AI system. 2018.

[32] Michael B. Jones, Brian Campbell, and Chuck Mortimore. JSON web token (JWT) profile for oauth 2.0 client authentication and authorization grants. *RFC*, 7523:1–12, 2015. doi: 10.17487/RFC7523. URL https://doi.org/10.17487/RFC7523.

[33] Iftikhar Ahmed Khan, Willem-Paul Brinkman, Nick Fine, and Robert M. Hierons. Measuring personality from keyboard and mouse use. *Proceedings of the 15th European conference on Cognitive ergonomics the ergonomics of cool interaction - ECCE '08*, (1984):1, 2008. doi: 10.1145/1473018.1473066. URL http://portal.acm.org/citation.cfm?doid=1473018.1473066.

[34] Maan Khedr and Nasser El-Sheimy. A smartphone step counter using IMU and magnetometer for navigation and health monitoring applications. *Sensors (Switzerland)*, 17(11), 2017. ISSN 14248220. doi: 10.3390/s17112573.

[35] Chrysanthi Konstanti and Evangelos Karapanos. An inquiry into goal-setting practices with physical activity trackers. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pages LBW587:1–LBW587:6, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5621-3. doi: 10.1145/3170427.3188663. URL http://doi.acm.org/10.1145/3170427.3188663.

[36] I-Min Lee, Eric J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair, and Peter T Katzmarzyk. Impact of Physical Inactivity on the World's Major Non-Communicable Diseases. *The Lancet*, 380(9838):219–229, 2012. ISSN 0140-6736. doi: 10.1016/S0140-6736(12)61031-9.Impact.

[37] You-Wei Lin and Chi-Yi Lin. An interactive real-time locating system based on bluetooth low-energy beacon network †. *Sensors*, 18(5), 2018. ISSN 1424-8220. doi: 10.3390/s18051637. URL http://www.mdpi.com/1424-8220/18/5/1637.

[38] Edwin A Locke and Gary P Latham. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, 57(9):705, 2002.

[39] Ewa Luger and Abigail Sellen. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 5286–5297, 2016. doi: 10.1145/2858036.2858288. URL http://dl.acm.org/citation.cfm?doid=2858036.2858288.

[40] Cécile Boulard Masson, David Martin, Tommaso Colombino, and Antonietta Grasso. "the device is not well designed for me" on the use of activity trackers in the workplace? In *COOP 2016: Proceedings of the 12th International Conference on the Design of Cooperative Systems, 23-27 May 2016, Trento, Italy*, pages 39–55. Springer, 2016.

[41] MathWorks. Counting steps by capturing acceleration data from your android device. https://www.mathworks.com/help/supportpkg/mobilesensor/examples/counting-steps-by-capturing-acceleration-data-from-your-android-device.html, 2018. Online; accessed 22 December 2018.

[42] Charles E Matthews, Kong Y Chen, Patty S Freedson, Maciej S Buchowski, Bettina M Beech, Russell R Pate, and Richard P Troiano. Amount of time spent in sedentary behaviors in the united states, 2003–2004. *American journal of epidemiology*, 167(7):875–881, 2008.

[43] Danilo M. Nogueira, Cristiano Maciel, José Viterbo, and Daniel Vecchiato. A privacy-driven data management model for smart personal assistants, 2017. ISSN 16113349.

[44] Catherine J Norris and D Ph. Brief Mindfulness Meditation Improves Attention in Novices. pages 2803–2808, 2015.

[45] U.S. Department of Health and Human Services. Physical activity guidelines for americans. https://health.gov/paguidelines/second-edition/pdf/Physical_Activity_Guidelines_2nd_edition.pdf, 2018. Online; accessed 16 November 2018.

[46] Health Council of the Netherlands. Physical activity guidelines 2017. https://www.healthcouncil.nl/documents/advisory-reports/2017/08/22/physical-activity-guidelines-2017, 2017. Online; accessed 12 November 2018.

[47] World Health Organisation. Healthy workplaces : a model for action for employers, workers, policymakers and practitioners. 2010. ISSN 7117181788.

[48] World Health Organization et al. *Healthy workplaces: a model for action: for employers, workers, policymakers and practitioners*. Geneva: World Health Organization, 2010.

[49] André Pimenta, Davide Carneiro, Paulo Novais, and José Neves. Monitoring mental fatigue through the analysis of keyboard and mouse interaction patterns. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 222–231. Springer, 2013.

[50] Andre Pimenta, Sergio Goncalves, Davide Carneiro, Florentino Fde-Riverola, Jose Neves, and Paulo Novais. Mental Workload Management as a Tool in e-Learning Scenarios. *PECCS 2015 Proceedings of the 5th International Conference on Pervasive and Embedded Computing and Communication Systems*, pages 25–32, 2015. doi: 10.5220/0005237700250032.

[51] André Pimenta, Davide Carneiro, José Neves, and Paulo Novais. A neural network to classify fatigue from human-computer interaction. *Neurocomputing*, 172:413–426, 2016. ISSN 18728286. doi: 10.1016/j.neucom.2015.03.105. URL http://dx.doi.org/10.1016/j.neucom.2015.03.105.

[52] Bernd Ploderer, Wolfgang Reitberger, Harri Oinas-Kukkonen, and Julia Gemert-Pijnen. Social interaction and reflection for behaviour change. *Personal Ubiquitous Comput.*, 18(7):1667–1676, October 2014. ISSN 1617-4909. doi: 10.1007/s00779-014-0779-y. URL http://dx.doi.org/10.1007/s00779-014-0779-y.

[53] Suporn Pongnumkul, Pimwadee Chaovalit, and Navaporn Surasvadi. Applications of smartphone-based sensors in agriculture: A systematic review of research. *Journal of Sensors*, 2015(July), 2015. ISSN 16877268. doi: 10.1155/2015/195308.

[54] Stéphanie A Prince, Kristi B Adamo, Meghan E Hamel, Jill Hardt, Sarah Connor Gorber, and Mark Tremblay. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 5(1):56, 2008.

[55] A Purington, J G Taft, S Sannon, N N Bazarova, and S H Taylor. "Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo. *Conference on Human Factors in Computing Systems - Proceedings*, Part F1276:2853–2859, 2017. doi: 10.1145/3027063.3053246. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019559408{&}doi=10.1145{%}2F3027063.3053246{&}partnerID=40{&}md5=a155e45b9083b5fedf2215ffab18a136.

[56] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. Passive and In-Situ assessment of mental and physical well-being using mobile sensors. *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*, page 385, 2011. doi: 10.1145/2030112.2030164. URL http://dl.acm.org/citation.cfm?doid=2030112.2030164.

[57] Arsénio Reis, Dennis Paulino, Hugo Paredes, and João Barroso. Using Intelligent Personal Assistants to Strengthen the Elderlies' Social Bonds, 2017. ISSN 16113349. URL http://link.springer.com/10.1007/978-3-319-58700-4{_}48.

[58] Hongjai Rhee and Sudong Kim. Effects of breaks on regaining vitality at work: An empirical comparison of 'conventional' and 'smart phone' breaks. *Computers in Human Behavior*, 57:160–167, 2016. ISSN 07475632. doi: 10.1016/j.chb.2015.11.056. URL http://dx.doi.org/10.1016/j.chb.2015.11.056.

[59] Caroline R Richardson, Tiffany L Newton, Jobby J Abraham, and Ann M Swartz. Walking Interventions and Weight Loss. (1):69–77, 2008. doi: 10.1370/afm.761.INTRODUCTION.

[60] Rothney. Validity of Physical Activity Intensity Predictions by ActiGraph,\nActical, and RT3 Accelerometers. *Obesity*, 16(8):1946–1952, 2008. doi: 10.1038/oby.2008.279.Validity.

[61] Umair Saad, Usama Afzal, Ahmad El-Issawi, and Mohamad Eid. A model to measure QoE for virtual personal assistant. *Multimedia Tools and Applications*, 76(10):12517–12537, 2017. ISSN 15737721. doi: 10.1007/s11042-016-3650-5.

[62] Ruhi Sarikaya. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine*, 34(1):67–81, 2017. ISSN 10535888. doi: 10.1109/MSP.2016.2617341.

[63] Mike Sharples, Maria Aristeidou, Eloy Villasclaras-Fernández, Christothea Herodotou, and Eileen Scanlon. The Sense-it App. *International Journal of Mobile and Blended Learning*, 9(2):16–38, 2017. ISSN 1941-8647. doi: 10.4018/IJMBL.2017040102. URL http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJMBL.2017040102.

[64] Matti Siekkinen, Markus Hiienkari, Jukka K Nurminen, and Johanna Nieminen. How low energy is bluetooth low energy? comparative measurements with zigbee/802.15. 4. In *Wireless Communications and Networking Conference Workshops (WCNCW), 2012 IEEE*, pages 232–237. IEEE, 2012.

[65] Jeremiah Smith, Anna Lavygina, Jiefei Ma, Alessandra Russo, and Naranker Dulay. Learning to recognise disruptive smartphone notifications. *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services - MobileHCI '14*, pages 121–124, 2014. doi: 10.1145/2628363.2628404. URL http://dl.acm.org/citation.cfm?doid=2628363.2628404.

[66] Fabio A Storm, Ben W Heller, and Claudia Mazzà. Step detection and activity recognition accuracy of seven physical activity monitors. *PloS one*, 10(3):e0118723, 2015.

[67] Andreas Ströhle. Physical activity, exercise, depression and anxiety disorders. *Journal of neural transmission*, 116(6):777, 2009.

[68] Ann M. Swartz, Leah Squires, and Scott J. Strath. Energy expenditure of interruptions to sedentary behavior. *International Journal of Behavioral Nutrition and Physical Activity*, 8:1–7, 2011. ISSN 14795868. doi: 10.1186/1479-5868-8-69.

[69] Barry N. Taylor and Ambler Thompson. The International System of Units. *Nist Special Publication 330 2008 Edition*, page 52, 2008. ISSN 00465763. doi: 10.1007/BF01449764.

[70] Catrine Tudor-Locke, Susan B. Sisson, Tracy Collova, Sarah M. Lee, and Pamela D. Swan. Pedometer-Determined Step Count Guidelines for Classifying Walking Intensity in a Young Ostensibly Healthy Population. *Canadian Journal of Applied Physiology*, 30(6):666–676, 2005. ISSN 1066-7814. doi: 10.1139/h05-147. URL http://www.nrcresearchpress.com/doi/abs/10.1139/h05-147.

[71] Ruut Veenhoven. Subjective measures of well-being. In *Human Well-Being*, pages 214–239. Springer, 2007.

[72] Sven Venzke-Caprarese. Standortlokalisierung und personalisierte nutzeransprache mittels bluetooth low energy beacons. *Datenschutz und Datensicherheit - DuD*, 38(12):839–844, Dec 2014. ISSN 1862-2607. doi: 10.1007/s11623-014-0329-9. URL https://doi.org/10.1007/s11623-014-0329-9.

[73] Jacqueline Wijsman, Bernard Grundlehner, Hao Liu, Julien Penders, and Hermie Hermens. Wearable physiological sensors reflect mental stress state in office-like situations. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, (iii):600–605, 2013. ISSN 2156-8103. doi: 10.1109/ACII.2013.105.

[74] Fadel Zeidan, Susan K. Johnson, Bruce J. Diamond, Zhanna David, and Paula Goolkasian. Mindfulness meditation improves cognition: Evidence of brief mental training. *Consciousness and Cognition*, 19(2):597–605, 2010. ISSN 10538100. doi: 10.1016/j.concog.2010.03.014. URL http://dx.doi.org/10.1016/j.concog.2010.03.014.

[75] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. Understanding and mitigating the security risks of voice-controlled third-party skills on amazon alexa and google home. *CoRR*, abs/1805.01525, 2018. URL `http://arxiv.org/abs/1805.01525`.

[76] Ziwei Zhu, Sebastian Ober, and Roozbeh Jafari. Modeling and detecting student attention and interest level using wearable computers. *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks, BSN 2017*, pages 13–18, 2017. doi: 10.1109/BSN.2017.7935996.

[77] Oren Zuckerman and Ayelet Gal-Oz. Deconstructing gamification: evaluating the effectiveness of continuous measurement, virtual rewards, and social comparison for promoting physical activity. *Personal and Ubiquitous Computing*, 18(7):1705–1719, 2014. ISSN 16174909. doi: 10.1007/s00779-014-0783-2.