MASTER THESIS

AN EMBODIED CONVERSATIONAL AGENT IN A MOBILE HEALTH COACHING APPLICATION

J.K. Hendrix

MSC HUMAN MEDIA INTERACTION

EXAMINATION COMMITTEE
Dr. E.M.A.G. van Dijk (University of Twente, Enschede, the Netherlands)
Dr. Ir. H.J.A. op den Akker (University of Twente, Enschede, the Netherlands)
R. Klaassen, MSc. (University of Twente, Enschede, the Netherlands)
H. op den Akker, MSc. (Roessingh Research and Development, Enschede, the Netherlands)

UNIVERSITY OF TWENTE.

18-04-2013

Abstract

Sedentary lifestyles are increasingly common in modern day society. This is becoming a serious problem, as a lack of physical activity can not only lead to overweight and obesity, but also increases the risk of many other health problems. In order to increase their physical activity levels, most people can benefit from some form of guidance or coaching. Many digital systems designed to provide such coaching are currently being developed, including systems for mobile hardware platforms. In an attempt to increase the effectiveness of such mobile health coaching systems, we have integrated an Embodied Conversational Agent (ECA) into one such system.

A user experiment was performed where participants used the system for two weeks, with the system employing plain text messages to deliver feedback during one of the weeks, and delivering feedback through the ECA during the other. Participants completed several surveys and an interview, and activity data was recorded and stored by the system. This data was then analyzed to try and find differences between the two feedback delivery methods.

Analysis of the collected data did not reveal a significant advantage of the ECA feedback version over the text feedback version. In fact, the text version received significantly higher scores on several items. Participants' responses during the interviews indicated that the lack of glanceability of the ECA feedback, combined with the predictability of the feedback message contents, had a strong negative influence on the evaluations of the ECA feedback version.

Preface

Before you lies the report detailing my final project, the culmination of my time as a Human Media Interaction (HMI) student. It focuses on the subject of Embodied Conversational Agents. This is not a subject with which I have had a lot of prior experience, nor is it a subject that I intend to specialize in. In fact, when starting this project my attitude towards ECAs was slightly skeptical (and I can not say that I have really been convinced in the process). So why then did I choose this assignment? Because I was quite interested in two of the other aspects of the project: the field of health behavior change, and working with the Android platform.

While I can not say that I intend to continue working in the field of ECAs (or health behavior change), I am quite happy with the decision to choose this assignment. I have learned a lot about several fields that were relatively new to me and that I found very interesting. Of course, tackling such a large project (relative to other study-related projects at least) alone, albeit with a healthy dose of support, was also a very valuable learning experience in and of itself. Looking at the entirety of the HMI study program that I have followed, I must admit that it did not feel like much of a specialization to Computer Science, but rather like it has broadened my horizons by incorporating elements of other fields. I have however very much enjoyed almost all of it.

Now to thank all those that have contributed in some way to this report lying before you today. First of all, I would like to thank my supervisors Betsy van Dijk, Rieks op den Akker, Randy Klaassen and Harm op den Akker for their guidance and support in every aspect of carrying out this project. I would also like to thank Dennis Reidsma for the support on the subject of Elckerlyc during this assignment and the previous. Thanks go to Roessingh Research and Development for providing the hardware needed to carry out the user experiment. More generally, I want to thank my parents and brother for all their moral, practical and tangible support throughout my time as a student. And finally, since the end of this project also marks the end of my time on the University of Twente campus and in the city of Enschede, I would like to thank all the friends I have made here over the years for making my time in Enschede unforgettable.

Jordi Hendrix
Enschede, April 2013

Contents

Chapter 1

Introduction

Overweight [1] and sedentary lifestyles [2] are global problems that are starting to receive more and more attention. One way to try and combat these issues is through the use of health coaching software [3]. This type of software is also making its way to mobile platforms [4]. While this type of software has already shown promise, much can still be done to improve its effectiveness and lasting appeal. One potential way of improving these aspects is through the use of Embodied Conversational Agents (ECAs), which could improve a user's view of the system on aspects such as trust, liking and respect [5].

1.1 Background

Examples of mobile physical activity promotion systems that make use of ECAs are still scarce, and while this project does not aim to develop a fully consumer-ready version of such a system, it will attempt to find evidence that this is indeed an area of promise. This project is being carried out at the Human Media Interaction (HMI) group of the University of Twente, and is a continuation of the work described in [6]. There are also parallels with other research being carried out by the HMI group as part of the Smarcos project, which is described in [7].

Aside from the HMI group, this project is also being supported by Roessingh Research & Development[1] (RRD). RRD is a research center for rehabilitation technology associated with the Roessingh rehabilitation center in Enschede. They are the developers of the mobile health coaching system that is used in this project. They have a potential interest in using ECAs in their products and supply software and hardware for use in user experiments, and are involved throughout the overall process.

1.2 Goals

This research aims to assess the benefits of using ECAs in mobile health coaching systems in general, and physical activity promotion systems in specific. This goal is formalized in the following research question:

Does an ECA offer a valuable addition to mobile health coaching systems?

This question is very general and cannot be answered directly by a single quantifiable measure. In order to determine whether or not ECAs are of value to mobile health coaching systems, we will attempt to find an area or areas in which a mobile health coaching system with an ECA has clear advantages over such a system without one. To do this, we will assess several areas for which there are indications that such advantages may be found. For these areas, we define the following subquestions:

Does the addition of an ECA to a mobile health coaching application...

1. lead to an increase in users' physical activity levels?
2. lead to an increase in users' evaluations of the user experience?

[1] http://www.rrd.nl/

3. lead to an increase in users' evaluations of the quality of coaching delivered by the system?
4. lead to users continuing to use the system for longer?
5. lead to an increase in users' perceived credibility of the system?

Each subquestion deals with a separate area of potential benefit. While it may be unrealistic to assume that we can reach conclusive (affirmative) answers to all of these questions, significant results on even one of them can provide enough information to answer the main research question.

1.3 Approach

In order to find an answer to the questions posed in the previous section, this project will integrate existing ECA software into an existing mobile health coaching application that focuses on physical activity promotion. A detailed description of this system and the existing software involved in it can be found in Chapter 3. This system will then be put to the test in a user experiment lasting 2 weeks and involving 14 office workers as test subjects. The users will be presented with 2 versions of the system, one version where the ECA delivers feedback messages, and one where these messages are presented as simple text messages. Obtained usage data and questionnaires filled out by the test subjects will then be analysed in order to answer the research questions posed in the previous section.

1.4 Structure of this Document

The remainder of this document is structured as follows. Chapter 2 discusses the theoretical background for our research, including an overview of relevant psychological theory and an exploration of other research in the same domain. Chapter 3 describes the elements of the final system that is used in our experiment, as well as the work that has been performed to tailor this system to the requirements of the experiment. Chapter 4 contains the research methodology that is used in the experiment, including the overall setup, a discussion of the data collected, and an outline of the procedure followed. Chapter 5 shows the results of the experiment in the form of an analysis of all the collected data. In Chapter 6, we end by presenting and discussing our conclusions, reflecting on the different aspects of the experiment, and providing recommendations for future work.

Chapter 2

Background: Health Behavior Change Support Systems

> A behavior change support system (BCSS) is an information system designed to form, alter or reinforce attitudes, behaviors or an act of complying without using deception, coercion or inducements. [8, p6]

This is the definition of a Behavior Change Support System, as given by Oinas-Kukkonen. We preceed this with the word *'Health'* and use the term Health Behavior Change Support System (HBCSS) throughout this text to refer to the type of information systems we are concerned with. This chapter presents an overview of the field of HBCSSs. We will start with a general outlook, and then focus more and more on the type of HBCSS we will use in our experiments. We will do this by first discussing the reason HBCSSs are needed, then explaining some of the psychological processes and theories involved in health behavior change, then reviewing some of the more traditional E-Health systems and research, and we conclude with a discussion of several existing E-Health systems that make use of ECAs.

2.1 Reasons for Health Behavior Change

This chapter is about Health Behavior Change Support Systems. So, if we want to support a change in behavior, something must be wrong with the current state of the behavior. There are plenty of patterns of behavior related to health that require change and can benefit from support, for example substance addiction. In light of the assignment however, we will focus on a single form of health behavior change: physical activity promotion. Again, if we want to promote physical activity, we must have a reason to assume that people need to be more physically active. This first section presents that reason.

2.1.1 Overweight and Obesity

In 2008, over 1.4 billion adults (age 20 and over) were overweight worldwide, and around 500 million of those were obese, according to the World Health Organisation (WHO) [1]. This makes it clear that overweight and obesity are a very serious global problem. Overweight has been identified as a major risk factor in a number of diseases such as diabetes, heart disease, stroke, osteoarthritis and several forms of cancer. Fundamentally, the cause of overweight and obesity is consistently consuming more calories than one burns. Therefore the first problem is the prevalence of unhealthy eating habits in our society. Much of the food we eat these days contains large amounts of energy, fat, salt and sugars and has a lack of vitamins and minerals. While this is obviously a problem that needs to be addressed in order to get the overweight epidemic under control, we focus here on the other side of the coin, the lack of energy expenditure caused by increasingly sedentary lifestyles.

2.1.2 Sedentary Lifestyles

In today's world, we have a lot of modern technologies to make our lives easier and more pleasant. Unfortunately, a side-effect of this is that a lot of us are becoming couch potatoes. We commute to

work by car, spend the day behind our computers, and once we get home, we watch TV. Obviously this scenario does not hold true for everyone, but it shows us some of the main causes of the increase in sedentary lifestyles. According to the WHO, 31% of the world's population was insufficiently active in the year 2008 [2]. Being inactive may not seem like such a big problem at first, but leading such a sedentary lifestyle brings with it many serious health risks, even without being overweight. The WHO estimate that 3.2 million deaths per year are attributable to a lack of physical activity [2].

2.1.3 Benefits of Physical Activity

Instead of focusing on the health risks that stem from a sedentary lifestyle, let us take a more positive approach and focus on the health benefits that an increase in physical activity can bring. This is not just to provide a more optimistic outlook, but also because increasing physical activity can have benefits even to those people that are not overweight and/or do not lead a particularly sedentary lifestlye. According to an extensive overview by the UK Department of Health [9], physical activity has shown significant benefits to health in the following areas:

Cardiavascular Disease Physical activity helps protect against coronary heart disease. It also reduces the risk of stroke and reduces risk factors for cardiovascular disease in general.

Overweight Physical activity along with a healthy diet is the best way to lose weight. Physical activity also reduces risk of mortality and morbidity in people who are already overweight.

Diabetes Physical activity significantly reduces risk of developing type 2 diabetes. Patients with type 2 diabetes can reduce risk of mortality with enough physical activity.

Musculoskeletal System Specific forms of physical activity can reduce risk of esteoporosis. Physical activity can also improve health for people suffering from osteoarthritis and lower back pain, but care must be taken not to be too active and make the problem worse.

Mental Health Physical activity can be used effectively in the treatment of clinical depression. Physical activity also has general benefits on mental health, such as reduced anxiety and stress.

Cancer Physical activity can protect against colon cancer, and reduces risk of breast cancer in women after menopause. Overall risk of cancer is also reduced by physical activity.

The report also concludes that even a moderate level of physical activity already offers a high level of protection. All in all, this clearly shows that physical activity can have a tremendous beneficial effect on people's health and general well-being, and that, for most people, it would certainly be worthwhile to become more physically active.

2.2 Psychological Frameworks

There are numerous psychological theories that have relevance in the field of health behavior change. We have chosen to focus on two of them that we believe are most commonly used and most relevant to the domain of HBCSSs. We will discuss the *transtheoretical model*, a model that focuses on the stages that an individual goes through to achieve a lasting change in health related behavior, as well as the theory of *persuasive technology*, which focuses on the ways in which technology can be designed to incite behavior change in people. Of course, these theories are much broader in scope than just physical activity promotion, so we will focus on what is most relevant to our specific application domain.

2.2.1 Transtheoretical Model of Health Behavior Change

Behavior change is a complex psychological process. The most commonly accepted and used model to describe this process is the *Transtheoretical Model of Health Behavior Change* by Prochaska [10]. Work on this model started around 1980, and has continued ever since, resulting in many small changes over the years and many versions of the model, each with slight differences. Also referred to as the *stages of change*, this model describes the process of behavior change in terms of the stages a person goes through, seen in Figure 2.1. The stages found in every version of the model are: *precontemplation*, *contemplation*,

preparation (or *determination*), *action* and *maintenance*. Some versions of the model also include relapse as a separate state, whereas others describe relapse as the transition to an earlier stage. Another stage that is not always included is the *termination* stage, mainly because most forms of behavior change require some measure of maintenance for a very long time, if not the rest of a person's life.

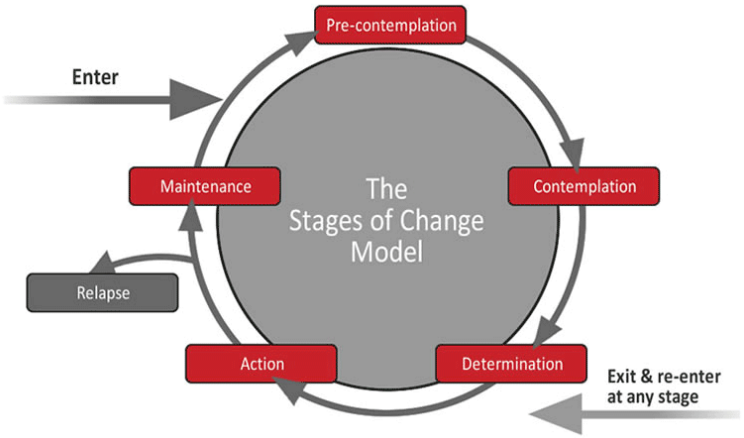

Figure 2.1: The stages of change and the transitions between them

The stages of *precontemplation* and *contemplation* are of little relevance to this project. Clearly it is important that people in these stages are being properly informed on their negative behaviors and persuaded to move towards the next stages, but use of an HBCSS would ordinarily not yet occur in these stages. Assuming that the usage of an HBCSS is voluntary, any person using such a system is already past these stages of change. The one way in which an HBCSS could be part of these stages is by informing potential users in the contemplation stage of its existence and benefits. Knowing about available support systems can help motivate people to proceed to the preparation stage.

As for the actual users of an HBCSS, it stands to reason that they are mainly in the *action* stage of change. Trying out an HBCSS and getting familliar with it can still be construed as being part of the *preparation* stage, but once somebody starts using an HBCSS seriously and making the changes in behavior that come with it, he or she is clearly in the *action* stage. In the *action* stage people are already making serious changes to their behavior and are committed to achieving a stable situation in which their actual behavior matches their desired behavior. An HBCSS can have a great deal of impact in this stage, not only offering a user tools for achieving desired behaviors, but also giving stability and helping the user make the desired behavior into a habit.

Once somebody has fully achieved the desired behavior, he or she enters the *maintenance* stage. An HBCSS could still prove very valuable in this stage, although the focus may need to shift somewhat. Where offering tools and ways to change behavior is a valid tactic during the action stage, people in the maintenance stage should already be comfortable with the routines and habits they have built up. Offering more (different) ways to change behavior in this stage may not only be ineffective, but could also be annoying to the user. The focus of an HBCSS in the maintenance stage figures to be more about monitoring the user and intervening when required to avoid relapses to negative behaviors.

In order to progress through the stages of change, the theory presents 10 processes of change. We will discuss the 4 we believe to be most relevant to the field of HBCSSs.

Counterconditioning is about learning healthy behaviors to substitute for unhealthy ones. There is clearly a role for HBCSSs here; for example by building a habit of using more active modes of transportation (walking, bicycling) instead of passive ones (car, bus).

Stimulus Control is about replacing cues for unhealthy behavior with ones supporting healthy behavior. While an HBCSS generally cannot remove cues for unhealthy behaviors, it can provide stimuli that support healthy behaviors; for example by suggesting parking a bit further from the office, or taking the stairs instead of the elevator.

Contingency Management is about the reinforcement of attitudes. While this can be both positive and negative, it has been shown that positive reinforcement is more effective in people trying to change on their own accord. An HBCSS could implement this by offering positive feedback whenever a user exhibits desired behaviors.

Helping Relationships are about offering the subject a place to turn to for support. In performing other supporting roles, an HBCSS can already build a relationship with the user. While relationships are traditionally formed with other people and not with computer systems, research has shown that people also tend to treat computers as social entities [11]. This should certainly hold true for a system that interacts with a user through an ECA.

One additional theory that has been integrated into the *transtheoretical model* is the theory of *self-efficacy* developed by Bandura [12]. *Self-efficacy* is the confidence people have in their own ability to deal with specific situations without returning to old negative behaviors. While *self-efficacy* is strengthened naturally through success, an HBCSS could reinforce this process by explicitly making users aware of their successes, either by comments or by showing them monitoring data that indicate progress. Some versions of the model describe the termination stage as the point where the subject has achieved 100% *self-efficacy*.

2.2.2 Persuasive Technology

> Persuasion involves one or more persons who are engaged in the activity of creating, reinforcing, modifying, or extinguishing beliefs, attitudes, intentions, motivations, and/or behaviors within the constraints of a given communication context. [13, p34]

While this seems like a rather detailed definition, the field of persuasion and influence is very broad and contains numerous different theories and models. Since we are dealing with persuasion by computer systems, we will focus on the area of *persuasive technology*, which deals with exactly that subject. An overview of this field is given by Fogg [14], and we will base the rest of this section on the thoughts posed in that work.

In terms of persuasion, computers, and thus HBCSSs, have several important advantages. Compared to traditional media, HBCSSs have the advantage of interactivity. In the context of persuasion the most important use of interactivity is being able to adjust strategy according to user input or other feedback. Compared to people, HBCSSs have a larger list of advantages:

Persistence HBCSSs have the ability to continuously keep on performing whatever persuasive acts are needed, even while also performing other functions. A person may get tired of trying after a while or simply have limitations on the time he or she has to persuade the subject.

Anonimity HBCSSs can be used in absolute privacy if desired. Also, the act of sharing information with an HBCSSs is often less threatening than sharing information with another person.

Data Processing An HBCSS has the capacity to process large quantities of data in short time, which allows it to give faster feedback on analyzed data than a person would be able to.

Modalities HBCSSs can make more extensive use of different output modalities than people. Where a person would normally have to describe any information he or she wants to convey, an HBCSSs could make use of graphics or sounds in order to present complex information in an effective way.

Scalability HBCSS software can be easily distributed in large numbers and across large distances virtually without cost. Even hardware, such as computers or smartphones, can be mass-produced. A person serving as a persuasive agent can only affect a very limited number of people at once.

Ubiquity With the rise of mobile computing devices, a lot of people have a computer, and thus potentially a HBCSS, with them at almost every moment. This allows the HBCSS to perform persuasive tasks in places and at times that would not be available to another person.

Fogg identifies three distinct roles a persuasive system can play: tool, medium and social actor. A tool system supports the user in practical ways, for example a wizard guiding a user through some process. A system playing the role of medium provides the user with experiences, for example a Virtual Reality system. The social actor role is for systems that support users socially, for example by rewarding them for desirable behavior. Since the medium role is of little relevance to either HBCSSs or ECAs, we focus on the other two.

Fogg defines a persuasive technology tool as: "an interactive product designed to change attitudes or behaviors or both by making desired outcomes easier to achieve" [14, p32]. There are several different tactics available that computers can use to perform this function. We describe those most relevant to HBCSSs and omit the others:

Tailoring Tailoring works by presenting users with only relevant information, allowing them to get to know what they need quickly and with little effort. Because of an HBCSS's ability to process large amounts of data and its knowledge of the user, it should be able to utilize this tactic quite effectively in certain situations. An example would be to provide a user with local weather data, or by being aware of users' activity preferences.

Suggestion A tool can persuade simply by making suggestions to the user. The key to effective suggestion is to make relevant suggestions at opportune times. This idea of the opportune moment was prominent in ancient Greek rhetoric, signified by the word 'kairos' and embodied by the god of that same name. This principle of kairos is the key advantage of mobile HBCSSs, for two reasons. First, since a mobile HBCSS is almost always on or around the user, it can choose virtually any moment to give a suggestion. Secondly, a mobile HBCSS equipped with sensor technology can be aware of the user's state and thus determine the opportune moment to offer suggestions. As an example, having an HBCSS suggesting that a user go for a run while it is raining heavily is unlikely to have any effect, whereas suggesting that a user take a walk when the weather is nice and he or she has no imminent appointments is far more likely to result in succesful persuasion.

Self-monitoring Self-monitoring is the process of providing a user with information about themselves that he or she can not perceive themselves (with the same precision). Any HBCSS equipped with sensor technology can aid in self-monitoring by providing the user with an overview of collected sensor data. Self-monitoring tends to be more effective when the presented information is more up-to-date, with real-time updates obviously being the most effective situation. A prime example of a self-monitoring tool in an HBCSS context is a step counter.

Surveillance Surveillance is the monitoring of other people's actions. Monitoring can actually work as a persuasive tool in two ways. Being able to watch the actions of others lead to desirable outcomes can have a persuasive effect on the observing party. This principle, commonly referred to as *modeling*, is a staple of *social learning theory* as initially posed by Bandura [15]. A user being aware of someone else watching their actions can also persuade them to perform certain behaviors. One way surveillance can be used in an HBCSS is to share results between users so that they can make comparisons. Since the specific HBCSS we will use in this project does not (currently) use surveillance tactics, further discussion of these principles is omitted.

Conditioning A conditioning tool is a system that uses the principles of operant conditioning to persuade its user(s). The principle behind operant conditioning is to present the user with positive reinforcement whenever desirable behavior is performed, in order to make this behavior a habit. An HBCSS could use conditioning simply by providing the user with positive feedback whenever a positive behavior is detected, or possibly even by implementing some sort of scoring system.

It has been known for a long time that people tend to respond to computers in a social way [11]. While this effect exists in virtually any computer application domain, it can be further exploited by actually making the computer application present itself as a social entity. This appearance the system presents to the user can be referred to as a persona. Through this persona, a computer can function as a persuasive social actor. One of the main ways for a computer system to play the social actor role is to give the user emotional feedback, such as praise and encouragement. A persuasive social actor attempts to build a relationship with the user by showing emotion and executing other social behaviors. This relationship can strengthen the user's attachment to the system in general, and can make any forms of

persuasion attempted by the system more effective. In order to give the appearance of a social entity, Fogg identifies five primary types of social cues:

Physical The illusion of a physical presence. For example, an HBCSS could display a face that moves occasionally to give the impression that the system is a person.

Psychological Expressing human characteristics such as feelings and personality, generally through language. To show this type of social cue, an HBCSS could apologise to the user when something goes wrong.

Language Interactive use of language beyond just offering text messages. Examples would be spoken text or speech recognition in an HBCSS.

Social Dynamics Adhering to the standards of human social behavior. This includes waiting for the user to finish speaking before replying, and praising the user when he or she succeeds.

Social Roles The adoption of a specific role that carries some social implications. For example, an HBCSS could play the role of a doctor or physical therapist.

Furthermore, a large number of factors can help to increase the persuasiveness of the persona projected by the system. Aspects such as general attractiveness [14] and affiliation with the user [16] can be exploited. It has also been shown that praise from a computer has a similar effect on people as praise from other people, and it can be used to make users more susceptible to persuasion. Even the social dynamic of reciprocity applies to the human-computer relationship [17]. When a computer has been helpful to a user, that user is more likely to "return the favor" and accept persuasion by the system.

2.2.3 Ethics

Obviously, in order to actually support a user, an HBCSS has to influence that user in some way. Since influence over people can be exploited relatively easily, it is important to keep an eye on the ethical side. Of course, inciting behavior change in people is not neccessarily unethical. In fact, in a typical HBCSS, the user actually wants the behavior change and freely chooses to use the system in order to achieve this. However, that does not excuse the system designer of responsibility. It is important for a designer to always be aware of the ethicality of every aspect of the system. It is not enough to only have ethical intentions and use ethical methods, if this nevertheless somehow results in (reasonably predictable) unethical effects, the designer is still to blame [18].

Some forms of influence are more likely to work in an unethical way than others. One area of concern is influence that is not obvious to the user. When a system directly asks a user to do something, it is obvious to the user that the system is trying to encourage the proposed behavior, but this is not always the case. For example, conditioning partially works on the user's subconscious, by trying to build habits and making behavior instinctual. In this regard, Berdichevsky states: "The creators of a persuasive technology should disclose their motivations, methods, and intended outcomes, except when such disclosure would significantly undermine an otherwise ethical goal." [18, p2]. This also already hints at the argument that not complying with one ethical rule does not necessarily make a technology unethical, it just means that it should be scrutinized more closely on other ethical issues.

One more point to be considered is users' privacy. This issue has two particular aspects. First of all, any data collected solely for analysis by the system should be handled with care and guaranteed to remain private. Secondly, if any data is shared with third parties, there should be close scrutiny on which data is transferred and whom it is transferred to [18]. In the context of health, it stands to reason that some data may be shared with a professional such as a therapist or physician. While this may benefit the user in the end, care should still be taken to not share anything the user would not be comfortable with, and to make sure the user is aware of information being relayed.

2.3 E-Health Systems for Physical Activity Promotion

Much research has already been done into HBCSSs, and many such systems have already been developed in attempts to help people trying to achieve behavior change through the use of technology. A lot of

these systems make use of some sort of application or website combined with the internet. This field is often referred to as E-Health. Since health is a very broad topic, E-Health has a broad range of applications. This goes from things like overcoming substance addiction or managing a chronic disease to more common problems such as healthy eating and proper exercise. In this section we will discuss some general attributes of E-Health, and focus on research related to the promotion and support of physical activity where applicable.

2.3.1 Benefits

Since E-Health systems generally fall into the field of persuasive technology, the potential benefits of E-Health systems are strongly related to the advantages of computers over people in the field of persuasion, as discussed in section 2.2.2. In this section however we will look at these benefits from a slightly different angle, and differentiate between two areas of benefits: intervention effectiveness and practical aspects. The former concerns the measured effect of behavior intervention through E-Health systems as compared to more traditional forms of intervention. This has been the main focus of most research into the area of E-Health. The latter area pertains to any practical aspects that arise in the real-world use of E-Health systems. Since research is generally done in controlled testing environments, less has been written about these aspects, although they may well be at least as important factors in the actual realization of consumer-ready E-Health systems.

Unfortunately, results of E-Health research do not generally show that E-Health is more effective than other forms of intervention. An overview of studies in the physical activity domain can be found in [3]. While this overview can not conclude that E-Health solutions are generally superior to more traditional forms of intervention, it does find that they are not significantly worse in any of the examined studies. While this may not seem like a particularly encouraging finding, we should realize that being at least roughly equal in effectiveness to traditional intervention methods is already quite an achievement, and may still prove to be enough if there are sufficient benefits on other fronts. Such benefits are obviously dependent on the specific application used and the traditional intervention method used for comparison, but typically there are a number of key areas in which E-Health can offer practical benefits over more traditional intervention methods.

One of these advantages is the ease with which someone can access and use an E-Health system. Since almost everybody has a computer at home these days, visiting a website or using a computer application is far easier than for example going to see a physical therapist. Even finding information tends to be easier on a computer system than in a self-help book or brochure. This makes it far easier for someone to reach an E-Health system than for example a therapist. This can also lower the hurdle of taking action to change one's behavior. This is related, but certainly not equivalent, to the concept of ubiquity mentioned in the context of persuasive technology, which pertains to a system being available wherever and whenever the user needs it.

Closely related to this is the advantage of anonymity and privacy. While this may be a more pressing concern in areas such as substance abuse than in physical activity behavior, people generally like to be able to keep things to themselves, especially when they have a problem and are looking for help. E-Health offers users anonimity in two separate ways. First of all, it means they do not have to share their problem and desire for help with another person such as a therapist. Aside from that however, it also makes it easier to keep the fact that they have a problem and need help hidden from those in their direct environment. For example, there is no chance of someone overhearing a call from a therapist, or spotting a self-help book in a bookcase. Again, while seeking help to increase one's physical activity level may be perfectly socially acceptable, a lot of people do place great value on their privacy.

There are also advantages for the service provider. A website or computer program is developed once and can then be distributed at almost no cost to any number of users around the world, so it is much easier and less costly to reach much larger groups of people within a single project. This is an example of the scalability advantage seen in persuasive technology. It should be clear that this in particular can potentially have a massive impact on the viability of deploying consumer-ready E-Health systems. While the initial costs of an E-Health program could conceivably be higher than other forms of intervention, the ability to reach a much larger amount of people without significant additional costs could mean the cost per user turns out far lower than in other forms of intervention. Of course, the ability to reach a large audience is a significant advantage even without a potential cost benefit. An intervention program that has great effectiveness still has little value in the big picture if it can only help a handful of people.

2.3.2 Mobile E-Health

With the rise of smartphones, more and more people now carry around devices that are essentially computers. This opens up another dimension for E-Health systems to use. Even before the rise of the current day smartphones that support the use of all kinds of apps, research was already being done into mobile E-Health systems using programmable mobile phones and PDA devices. Examples of such systems include applications that keep track of users' activity levels and share these with their friends (if desired) [19] [20] and systems that try and coach people during physical exercise [21] (see Figure 2.2). Of course E-Health systems running on mobile devices do have the added restriction that the user needs to own such a device, but there are also additional benefits to mobile E-Health systems over E-Health systems based on websites or desktop PC applications.

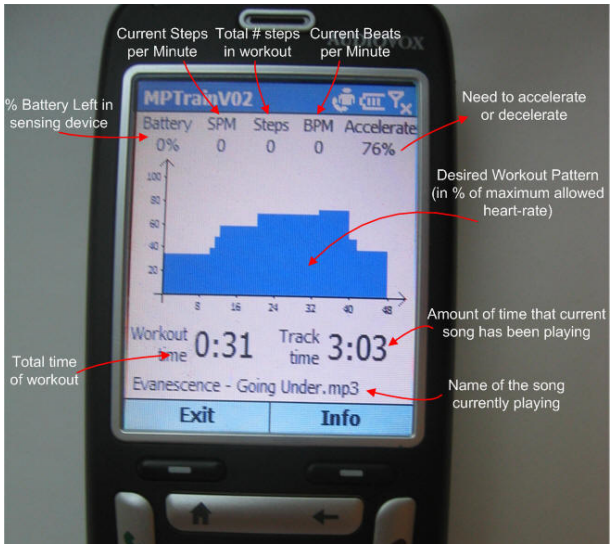

Figure 2.2: MPTrain, a mobile E-Health system

These advantages are again strongly related to those of persuasive technology in general. In the previous section we discussed effectiveness by looking at research outcomes and then looked more closely at practical aspects. Since the advantages of using a mobile platform over traditional PC-based E-Health systems are mostly in potential intervention effectiveness, we will focus here on that aspect. Also, mobile computing technology is still relatively young, and we can expect that many of its capabilities have yet to be fully exploited. This means that research on the effectiveness of mobile E-Health applications is still too sparse to draw a general conclusion.

The first major area where potential benefits can be achieved is related to the data processing advantage of persuasive technology. Modern smartphones often include several types of sensory hardware, such as a GPS and accelerometer. It is even possible to attach additional external sensor devices through bluetooth or other connectivity technologies. This can allow an E-Health system to collect valuable data about the user's activities. In itself, this is data collection and not data processing, but the two are clearly closely related. Automatic collection of data is only useful if this data can be processed and analysed in a timely manner. The idea of using sensor data to build a picture of the user's actions and situation is often referred to as *context awareness*.

The second area of potential benefits is essentially described by the ubiquity advantage of persuasive technology. Because users tend to carry their phones on them most of the time, interaction can occur whenever desired, and not just when the user is at his or her computer. This works two ways, the system can get the user's attention at any time by signalling him or her, and the user can turn to the system for information or guidance whenever he or she wants to.

While each of these can provide a valuable advantage, the most crucial benefit can be achieved by combining the two. This results in the system potentially being able to fully exploit the principles of kairos that were discussed earlier in section 2.2.2. Being able to automatically collect data about the user's activities and situation allows the system to determine what the opportune moment for persuading the

user is, and the fact that the system is carried around by the user allows the system to actually execute its persuasive behavior at just that time, which is crucial to succesful use of the *suggestion* tactic of persuasion [22].

2.4 Embodied Conversational Agents

An ECA is a computer generated character that is capable of interacting with a user through the use of language. The simplest ECAs just consist of a few different images and some text output, and the most advanced ones feature fully animated 3D-rendered bodies which allow the ECA to have a natural, realistic look and communicate nonverbally through gestures and body language, and can also speak with the user through speech synthesis and speech recognition. Research involving 3D-based ECAs started as early as the late 1990's, with systems such as Olga [23], Gandalf [24], and Rea [25].

Most early ECA research focused more on the ECAs itself than on any specific task domain. The focus was often on the interaction between the user and the system through the ECA, and to test this the user was given a trivial task domain in which the ECA was specialized, such as information on the planets of the milkyway [24] or real estate listings [25]. This line of research confirmed that users did interact with the ECAs in a social way. More recent studies generally focus on the use of ECAs in a specific task domain, in our case physical activity promotion.

Aside from a more task-oriented approach, the development of ECAs for mobile devices has also been getting more attention. Research into this area already started well before smartphones became commonplace, back when PDAs were the most prominent mobile computing devices. At that time the limitations of mobile devices in areas such as computing power were even greater than now, so most applications had to find creative solutions for this problem. Some solved this by using a remote server to perform most of the heavy lifting and then communicated the results back to the mobile device using a network connection [26]. Some systems do actually render 3D images directly on a PDA, but are limited to character models with very low detail and no textures [27]. Of course, it is also possible to largely avoid the problem, for example by using a tablet PC instead of a PDA [28], or by using a mock handheld system that is not truly mobile [29].

2.4.1 Advantages

There are still a lot of questions surrounding the effectiveness of using ECAs in user interfaces, whether in E-Health systems or otherwise. Back in 1997, Lester [30] posed the *persona effect*, the idea that a lifelike agent in a learning environment has a strong positive effect on the user's perception of their learning experience. This research also concluded that the lifelike character improved learning performance, but later research [31] questioned that claim and failed to find similar effects.

In general, most studies on the use of ECAs in the field of coaching and behavior change did not find significant improvements in coaching effectiveness when using an ECA. This does not mean that there is no use for ECAs in this domain. The use of an ECA has been shown in multiple cases to have a significant positive effect on user experience. While this may not be the most important aspect of an HBCSS, it can certainly have an impact, especially in the long term. While there is a lack of long-term studies on the subject, there are indications that a more pleasant user experience leads to, for example, users being motivated to use the system more frequently and over a longer period of time [32].

Another aspect of ECAs that potentially has a significant impact in long term use of a system is the development of a relationship between the ECA and the user. Building such a relationship can improve the user's evaluation of the ECA on points such as trust, respect and liking, and also result in the user being more interested in continued use of the system [5]. Use of a mobile platform offers even more possibilities to strengthen the user-ECA relationship. The fact that the ECA is always available potentially increases the perceived reliability and trustworthiness of the system, and the mere physical proximity and amount of interactions can cause the ECA to become significantly embedded into a user's everyday life [33].

2.4.2 ECAs in Physical Activity Promotion

The primary focus of this project is on applications running on handheld devices that deal with physical activity promotion and make use of ECAs for interacting with the user. While this already seems like a fairly restricted domain, there are still several different approaches, as the following examples will show.

The first system we discuss is MOPET [34]. MOPET is designed to support the user throughout exercise sessions, by guiding the user through fitness trails that alternate running with physical exercises. It tracks the user's position on the trail and shows the user's speed, and also tries to motivate the user through messages. It uses an external sensor device that collects heart rate and accelerometer data. When the user comes to an exercise point along the route, the system recognizes this and demonstrates the exercise to the user. The ECA is presented as a full-bodied animated 3D character that is rendered in real-time. While this system can make exercise more effective and more enjoyable for users, it does not actually motivate users to start exercising.

The second example is a mobile adaptation of the FitTrack system [35] (seen in Figure 2.3), which is very different from MOPET. Instead of supporting a user that is explicitly exercising, it monitors the user throughout the day and tries to motivate him or her to walk more. It uses the PDA's internal accelerometer to determine the user's steps walked, and then provides the user with feedback once a walk has been completed. The ECA itself is presented here as a closeup of a face, which allows it to use facial expressions and lipsync. This system is already more like the system we will use in that it is designed to be with the user at all times and monitors activity that the user regularly performs of their own initiative (walking). In this it does make users aware of their level of activity and rewards activity with praise, but it does not use *suggestion* or other explicit tactics to try and make users more active.

Figure 2.3: A mobile health counseling system by Bickmore

The last example we discuss here is the system developed in the Health and Fitness Companions project [36]. This system is actually a combination of multiple ECA-based systems that support the user in living a healthy lifestyle. We focus on the mobile companion, which runs on a PDA. In the way it works, the physical activity promotion element of the companions system contains elements from both of the systems that were already discussed. It is meant to be carried around by the user throughout the day, but it does focus on explicit exercise sessions like the MOPET system. The system can suggest and consequently track the user in different forms of exercise such as walking, running and cycling. The ECA presented on the PDA application is much less advanced than the other two systems as it consists of a static image and a text bubble. It does however include both Text-to-Speech (TTS) and speech recognition capabilities for interacting with the user.

Chapter 3

System Design

In order to find answers to the questions posed in Chapter 1, we will perform a user experiment. However, to be able to carry out a user experiment we need software (and hardware) to perform the experiment with. This chapter discusses the existing software and hardware systems used and how these are integrated into our final testing system.

3.1 The Continuous Care & Coaching Platform

The Continuous Care & Coaching Platform [37] (C3PO) is a mobile physical activity coaching system developed by RRD. While C3PO is being developed for different sets of users, its main focus is on patients suffering from medical conditions that require a fairly strict regulation of physical activity levels in order to be managed effectively. Examples of conditions that are targeted by C3PO are Chronic Obstructive Pulmonary Disease (COPD) and chronic lower back pain.

3.1.1 Hardware

A typical C3PO setup, such as the one used in our experiment, includes two pieces of hardware: a mobile Android device and an activity sensor node. The Android device is a smartphone, specifically an HTC Desire or HTC Desire S. This device is equipped with bluetooth in order to communicate with the sensor, and also has wireless data capabilities (3G and Wi-Fi). It also has access to the Google location service, which uses available mobile networks and/or (if enabled) GPS to estimate the device's geographic location.

The C3PO software is equipped to handle different kinds of activity sensor nodes, but the one used currently (and in our experiment) is the ProMove 3D motion sensor node developed by Inertia Technology[1], pictured in Figure 3.1. The ProMove 3D uses an array of sensors, including an accelerometer, a gyroscope and a magnetic compass, in order to capture movements by the user. Its size is roughly equal to that of the smartphone, and it can be attached to a user's belt using a belt clip or an elastic band clip.

3.1.2 Software

The main software component of the C3PO system is the Android application running on the smartphone. This application is currently designed to run as a homescreen replacement, meaning it is active at all times. This also renders the smartphone unusable for anything other than the C3PO application. The application has a modular structure, allowing for flexibility in design and use of the software for different target user sets. The modules used in our experiment are described here.

One of the basic elements is the status bar, which displays a digital clock, a speaker icon used to mute or unmute sound, an icon indicating the status of the connection to the sensor node, and a battery level indicator. This status bar appears at the top of every screen in the application. Several other basic

[1] http://www.inertia-technology.com/

Figure 3.1: The ProMove3D sensor node, pictured with belt clip

modules, such as the Bluetooth module, work in the background to facilitate the communication with the sensor node.

For the main screen a GUI (Graphical User Interface) module is used that displays an activity graph. This graph plots activity levels against the time of the day. A green curve indicates the (preset) target activity level for the user, and a blue curve is drawn based on the actual activity measured by the sensor node. The GUI module includes a number of customization options, such as showing the percentage of deviation from the target level and hiding the activity graph altogether. An example of the main screen (including the status bar at the top) can be seen in Figure 3.2.

Figure 3.2: Main screen of the C3PO mobile application

The module that is at the core of our experiment is the user input module. This module is used to perform scheduled interactions with the user. These interactions can take the form of questionnaires to be filled out by the user, or consist only of text being presented on the screen. Our experiment uses the basic feedback module, which combines with the user input module to present the user with feedback based on the measured activity level. This feedback consists of an evaluation message which indicates whether the user is below, at, or above the target activity level, along with a feedback message chosen randomly from a list of messages applicable to the current activity status. For example, a user who is below the target level may receive a message suggesting a short walk, a user who is around the target level may receive a message of praise, and a user who is above the target level may receive the suggestion to sit down and do some reading.

Several additional modules are available, such as a location module that tracks the user's location

through the Google location service, a weather module that uses online services to find the current weather at the user's location, and a synchronization module that periodically communicates with the central server to upload all collected data (activity and otherwise) and download modified settings information. These modules are not used in our experiment.

3.1.3 Further Information

Aside from the elements described above, the overall system also includes a framework for quickly setting up web portals. These web portals can be used (together with the central server and the synchronization module) to give users, researchers and health care professionals access to an easy to use and frequently updated overview of all relevant data, as well as an easier way for users to fill out questionnaires. Because of the short time frame of our experiment as well as the focus on the mode of feedback, this web portal system was not used.

Additional information on C3PO and related systems can be found in [37].

3.2 Elckerlyc

Elckerlyc [38] is a platform for the realization of virtual humans (or ECAs) developed by the HMI group at the University of Twente. In a nutshell, it takes a specification of behavior as input, and delivers an audiovisual representation of an ECA performing the specified behavior as output. This section discusses the key elements of the Elckerlyc platform.

3.2.1 A Behavior Markup Language Realizer

Elckerlyc is most commonly known as a Behavior Markup Language (BML) realizer. BML [39] is a language that allows for the specification of the form and timing of behavior that is to be executed by a virtual human. It was developed as part of the SAIBA framework [40], and is also used in a number of other behavior realizers such as SmartBody [41], EMBR [42], Greta [43]. BML Realizers such as Elckerlyc take behavior specified in BML and project it onto an embodiment.

The execution of behavior specified in BML may at first seem like a straightforward, although not simple, process. However, the nature and intended use of behavior specified in BML requires Elckerlyc to contain a powerful and complex scheduling component. Because BML does not require behavior to be specified sequentially, the Elckerlyc scheduler needs to resolve all the specified timing and synchronization constraints before being able to execute anything through the ECA. This is made even more difficult when considering that BML code can be supplied to the Elckerlyc system at any point in time, even when other BML code is already being executed. This requires the scheduler to determine which behavior elements can coincide, which can be sped up, which can be delayed, et cetera, while at the same time not only adhering to the demands specified by the BML code, but also maintaining a realistic and natural looking form of interaction.

3.2.2 Modular Design and Embodiments

The Elckerlyc platform is built to offer maximum flexibility in use and extension of the system. One of the hallmarks of this approach is the loader system, which uses XML specified definitions to dynamically load components of the platform. This allows applications based on the platform to use only those parts they need, or even additionally developed custom parts.

The most important elements that offer multiple different implementations that can be easily swapped in or out using this system are the embodiments. Embodiments are the components that make up the actual ECA. For example, the standard embodiment for the Elckerlyc system is a 3D scene rendered in real-time featuring a high resolution full body ECA (seen in Figure 3.3). But this embodiment can be replaced by, for example, an embodiment that controls a real-world robotic head.

In order for the other Elckerlyc components to be able to function independently of the specific embodiment used, each embodiment has its own binding. A binding forms the link between the BML elements used in Elckerlyc and whatever units of execution are used in that binding's embodiment. It contains definitions for which BML elements are to be translated to which types of embodiment behaviors.

Figure 3.3: Elckerlyc's default 3D embodiment

For example, the BML element for a smile may be bound to a certain animation file in the standard 3D embodiment, and to a set of motor instructions in the robotic head embodiment.

3.2.3 PictureEngine

Because the standard 3D embodiment is far too complex to run on the mobile Android device used in our experiment, we need a different embodiment for this. The PictureEngine is a lightweight graphical embodiment that uses a collection of 2D images in order to display the ECA. While this potentially reduces the realism and expressivity of the ECA, it also greatly reduces the system requirements to a level that is achievable on a mobile device. The PictureEngine was originally developed for use on a standard PC, but later modified and expanded to work well on the Android platform in a research project [6, 7] that served as a prologue to this project.

In order to generate a dynamic ECA from a collection of images, the PictureEngine uses a layer-based approach. Different parts of the ECA are displayed on different layers of the final image, and can thus be in different states. For example, one layer may contain the eyes, while another contains the mouth. By using this layer based approach, different parts of the ECA can be manipulated independently and combined in order to generate different expressions. It also allows the ECA to do several (connected or unconnected) things at once, such as blink while also speaking and pointing at something.

While single images may suffice for portraying expressions in many cases, there are other cases where an ECA simply has to display some motion in order to come across as believable. To make this possible, the PictureEngine also allows the use of animations instead of single images. The nature of the BML scheduler allows the duration of animations to be adjusted according to the BML code that is being executed, causing the animation to play faster or slower depending on the timespan determined by the scheduler. Animations can also include synchronization information that allows other behaviors to be made to coincide with specific elements of an animation.

In order to visually display the fact that the ECA is speaking, the PictureEngine provides a rudimentary lipsync facility. However, where the lipsync in the standard embodiment provides a full mapping from visimes to animation units, the PictureEngine lipsync currently does not make use of such a mapping, so it simply displays a "speaking" animation whenever the ECA is speaking.

3.2.4 Mobile Application

Since the Elckerlyc system is implemented almost entirely in Java, all of its core elements run on Android without any modification. However, since Android has its own environment for visual and audio output, the mobile version of Elckerlyc contains Android-specific versions of several of these subsystems. Unfortunately, this presents some additional limitations in the area of Text-To-Speech (TTS). The Android TTS facility does not offer some of the functionality that Elckerlyc commonly uses in the PC version, causing functions such as in-utterance synchronization points and viseme-based lipsync to become unavailable.

On a mobile device, there is a strong possibility that the user is unable to hear text spoken by the TTS. This could be caused by anything from environment noise to having the device muted. In order for the ECA to still be able to communicate with the user in such situations, the mobile application offers what is essentially subtitling functionality. Any spoken text is also simultaneously printed to a standard graphical text output area.

3.3 Integration

The sections above describe an activity coaching system (C3PO) and a system for displaying an ECA (Elckerlyc). Our user experiment will use the C3PO system as a basis, but while one of the conditions features the default text feedback, the second condition features an Elckerlyc ECA that presents all feedback messages to the user. This setup requires us to combine the two systems in some way.

A link between these two software systems has been made in the past [44], although that was in a different context. That work also used a different approach, where the Elckerlyc system was used in the form of an external application which was launched when required. After some consideration, we decided that it was preferable to fully integrate Elckerlyc into C3PO in order to make a valid comparison between C3PO's standard text feedback and feedback delivered by an Elckerlyc ECA. A seamless integration will allow the ECA to operate in the exact same graphical environment as the default text feedback, and will also eliminate any other possible differences that could influence the comparison.

The rest of this section discusses all of the programming work that was done in the context of this research. This mainly consists of the integration of the Elckerlyc and C3PO systems, but also includes several small adjustments needed to tailor the final application to office workers. First, the goals and requirements for the final application are presented. This is followed by a structural overview of the C3PO system where the elements that are changed have been highlighted. After that the additions and modifications made are discussed one by one, focusing on the changes that were made and the problems that were encountered.

3.3.1 Non-Functional Requirements

The main goal is simple: to combine the default C3PO feedback screen and the Elckerlyc ECA in order to obtain a new feedback screen that displays an ECA. While discovery and analysis of the requirements for the final application were performed in a more informal method than this document may suggest, we can still present a list of requirements that were considered during the development. First we discuss the non-functional requirements.

In terms of performance, the overall application must not be noticably slowed down by the integration of the Elckerlyc system. Also, the Elckerlyc ECA itself must be animated without any noticable hiccups or slowdown. This is an important point since the previous work on integrating Elckerlyc and C3PO encountered some problems in this area.

In terms of reliability, it is important that the Elckerlyc component remains fully operational even after long periods of running the application. Since C3PO has been designed to run continuously throughout the day (or even longer), it is of great importance that the Elckerlyc component is reliable and does not affect the execution of the main application even if it were to encounter internal problems.

Since we are only concerned with the integration of Elckerlyc, other areas of non-functional requirements, such as safety/privacy, usability and supportability are not discussed here because they are unaffected by the Elckerlyc system. There is one small exception in the area of usability. The new ECA feedback screen should offer the same usability as the original feedback screen, and must not provide additional intrusion on the user's activities.

One last non-functional requirement is that the original C3PO code must be changed as little as possible. Of course changes will have to be made, but it is preferable to try and limit the amount of different places in the software where code is modified. Doing this not only insures a clean technical design, but also makes it easier to merge the Elckerlyc integration into future versions of C3PO if that is ever desired.

3.3.2 Functional Requirements

Because C3PO is already a fully functional application, the functional requirements presented will focus only on the changes that need to be made, both to properly integrate Elckerlyc and to facilitate the use of the application by office workers. These requirements can be divided into and will be presented in a number of different categories.

The first category is the way in which Elckerlyc is integrated into the C3PO system, and the additional functionality required to make this integration useful. In order to be used in our experiment, the application must be capable of presenting both the traditional text and the new ECA feedback screens because we want participants to subsequently use both versions. Of course, we want the ECA to do at least something more than just read out the message, so some emotional elements should be added. In order to allow participants to make a valid comparison between the two feedback methods, we want them to be similar in order to minimize interfering factors. All this results in the following requirements:

FR 1. An ECA feedback screen must be available alongside the standard text feedback screen.

FR 2. Feedback messages must be presentable by the ECA feedback screen.

FR 3. Feedback presented by the ECA must include emotional behavior in the form of facial expressions.

FR 4. The GUI configuration must contain an optional flag for choosing between ECA and text feedback screens.

FR 5. The graphical presentation of the ECA feedback screen must be as similar as possible to the text feedback screen.

The second category is the actual functionality of the ECA feedback screen itself. While this should be as similar as possible to the text feedback screen, obviously the difference between text and a combination of animated images and sound means there will be some significant differences. We do not want practical concerns to interfere with the evaluation of our ECA, so participants should be able to mute the sound and end playback whenever desired. Also, participants will most likely have the smarthphone put away at the moment a message is generated, so playback should not simply begin immediately. Finally, we want it to be clear from the ECA feedback screen that the ECA is present (and about to say something), so it should always be displayed there. This results in the following requirements:

FR 6. The ECA feedback screen should be affected by the global mute setting of C3PO.

FR 7. Playback of ECA feedback should not start until the user initiates it.

FR 8. Users should be able to end playback before it finishes.

FR 9. The ECA should be visible in the ECA feedback screen at all times, even before playback begins.

The last category of functional requirements pertains to any other changes and/or additions that need to be made to the C3PO system in order to make it suitable for use by office workers and for the parameters of our experiment. Because the activities that are reasonable to suggest to a user differ depending on whether that user is at home or at the office, feedback message specifications need to have some way of indicating suitability for either situation. Also, we anticipate that participants may occasionally mute the system (during meetings for example), and this should not result in them missing a feedback message. This leads us to these requirements:

FR 10. Feedback messages need to have an optional flag to indicate their suitability for home and office situations.

FR 11. The system must vibrate upon receiving a feedback message, even when muted.

3.3.3 Structural Overview

Figure 3.4 shows the overall architecture of the C3PO system. This clearly shows the modular structure, where all the different modules are combined through the hub to make the total application. Of course, several of the modules displayed are not used at all in our application (such as the location and weather modules), and others are not affected by the changes (for example the bluetooth and IMA data modules). This already plays into the non-functional requirement of keeping the changes limited to a small number of places, but in order to comply with this requirement it is also important to leave the hub unchanged.

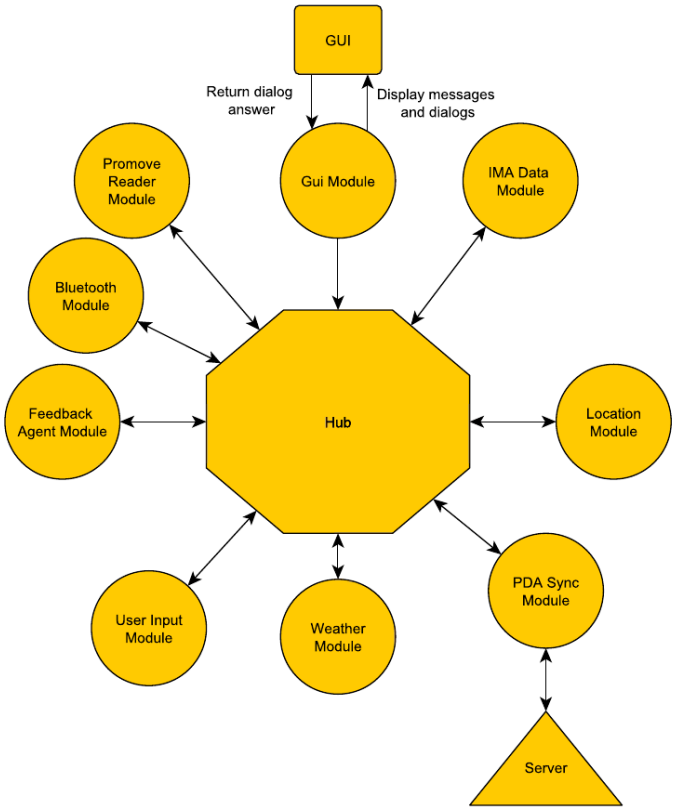

Figure 3.4: Overview of the C3PO architecture

The bulk of the changes are made within the GUI module, which is the module that contains the original feedback screen, and which thus contains the new ECA feedback screen. The only other area that is affected is the feedback module, which handles the selection of applicable feedback messages based on the activity data.

3.3.4 Integrating the Elckerlyc Mobile Packages

The first goal was to connect the Elckerlyc code to the C3PO code. In the starting situation, there were separate Android apps for both Elckerlyc and C3PO. Simply maintaining this separation and importing the required Elckerlyc classes from C3PO was not an option, because Android does not support importing from other apps. One way to get around this was to turn the existing Elckerlyc app into a library project. A library project is a form of Android application which cannot be run on its own, but which can be imported and used from within other apps. The problem with this approach was that library projects cannot contain assets. Because the images and configuration files used by Elckerlyc are assets, this would require those files to be included in the C3PO package. In the end it was decided that if the assets were to be included in the C3PO app anyway, it would be more practical to simply include the Elckerlyc Java classes in the C3PO project as well.

After the code was integrated in this way some modifications were made to the Elckerlyc classes. Because it would now function only from within the C3PO application, the original Activity (an Activity is essentially an executable class in Android) was removed. This removal required some small additional changes because of dependencies within other parts of the application, but these changes proved to be minor. Other than this, no modifications needed to be made to the Elckerlyc code at this point.

3.3.5 Feedback Screen

Now that the Elckerlyc code was available within the C3PO app, the next step was to actually include an Elckerlyc ECA in the feedback screen. The way this was done is by essentially merging the code of the original Elckerlyc Activity and the C3PO feedback screen (which is also an Activity). This process was fairly straightforward, and provided no significant problems.

The last step in the process of graphically incorporating the ECA into the feedback screen was to decide on the final graphical layout. Of course it was kept as similar to the original feedback screen as possible, in keeping with FR 5. However, since the original text took up a large portion of the screen, space had to be made for the ECA's graphical representation. In the end the ECA was displayed in the center of the screen, with the accompanying text above it and the OK button remaining in the bottom. In order to accommodate the smaller space for text, only the sentence currently being spoken by the ECA was displayed, instead of displaying all of the feedback at once. The font size was also decreased slightly. Additionally, the background image (a grayed-out closeup of part of the RRD logo) was removed. This was done because the ECA's base images were not transparent and thus caused an awkward-looking square cutoff in the background image.

During initial testing of the ECA feedback screen, several serious performance concerns were encountered. Initially, this seemed to be quickly solved by disabling the logging system used by Elckerlyc. The sheer amount of debug information generated was causing significant slowdown in overall execution. This problem was already encountered in earlier work so this was a simple fix. While this appeared to solve the problem, further testing showed that Elckerlyc was still subject to slowdown over time, as each new message was played back slower and with more hiccups than the last. Unsurprisingly, some debugging revealed thread leakage.

Further investigation into the matter showed that Elckerlyc itself was the culprit. Because the original Elckerlyc Android app was called as a standalone Activity, it was fully destroyed upon completion. Within C3PO however, the Elckerlyc system was continuing to run even after the feedback screen was closed, and with each following message, another Elckerlyc instance was spawned. This problem had two possible solutions: either properly shut down the Elckerlyc instance when the feedback screen is closed, or recycle the Elckerlyc instance and use it again for each following message. The first solution was eventually selected, for several reasons. First, according to FR 8 the user must be able to stop playback at any point. The easiest way to implement this is by simply exiting Elckerlyc. Second, the developmental nature of Elckerlyc makes it risky to assume that it can run for extended periods of time without any unexpected problems occurring. This clashes with the non-functional requirements. Third, closing Elckerlyc after use ensures that no system resources are being occupied unnecessarily.

Unfortunately, properly shutting down Elckerlyc turned out to be a challenge in itself. In principle, functionality was available for the clean shutdown of Elckerlyc and all its subsystems. However, after changing the application to make use of this functionality to shut down Elckerlyc after the playback of each message, threads were still being leaked. Further inspection revealed that there were three remaining problems. The easiest to fix was the emitter engine (the Elckerlyc subsystem responsible for generating eye blinks). In order to shutdown the emitter engine, a block of BML for ending the blinking behavior was injected before shutdown of the overall system. Also not difficult to solve was a small design flaw causing an additional timer thread to be created by the feedback screen itself. The last problem was the most difficult to solve. It took significant effort and some assistance from the original Elckerlyc author to find out that this was simply a bug within the core Elckerlyc code. Once it was discovered however, the fix was simple and quickly found.

3.3.6 Feedback Messages

After the development of the ECA feedback screen, the next item of business was to supply the ECA with actual feedback messages. C3PO normally uses plain text messages, but Elckerlyc only accepts

BML code as input. Therefore, a translation step had to be made somewhere. Several ways of doing this were considered. The initial idea was to manually formulate BML statements for each feedback message, and to present these to the system in the same configuration file that contains the text messages. Unfortunately, this approach came with some significant negatives. First of all, building all the BML code by hand would be a lot of extra work. Secondly, C3PO would require more modifications in order to support the handling of BML messages alongside the text versions.

During development, a second option arose. Instead of formulating BML by hand, it could be generated automatically based on the original text messages. Aside from the message text, the generation process would also require the message context, meaning the activity levels it was based on. This information is available at all times along with the message content within the C3PO system. This allowed the BML generation to take place all the way "at the end of the line", in the GUI module, avoiding the need for more modifications to other modules and/or the C3PO hub.

Having chosen and implemented the second option, BML statements were now being generated from the plain text versions of the feedback messages. Aside from simply applying the correct BML format to the text, this process also included the generation of emotional behaviors based on the message context. Unfortunately, the behavioral repertoire of the ECA was fairly limited. Combined with the fact that the feedback messages were quite short, this meant that opportunities for generating emotional behaviors were scarce. In the end, the only behavior that was implemented was a change of expression. In the case of a message with a positive content (meaning the user's activity is above the required level) the ECA smiled, in case of a neutral message she had a neutral expression, and in case of a message suggesting a form of physical activity (meaning the user's activity is below the required level) she had a slightly disapproving expression.

3.3.7 Switching Between Feedback Screens

In order to satisfy FR 1 and FR 4 it was necessary for both the original text and new ECA feedback screens to exist alongside each other, and for the software to choose which to use based on the GUI configuration file. In initial development the ECA feedback screen had simply replaced the original feedback screen, but this now needed to be changed. First, the original feedback screen was reinserted as a superclass of the new ECA feedback screen. This allowed other parts of the system to remain relatively intact since the original feedback screen was still intact in the exact same way it used to be. This design later turned out to be a mistake as it caused the leaked timer thread discussed in 3.3.5. This situation was then changed so that both the original and ECA feedback screens were extensions of an abstract class.

To allow the type of feedback screen to be selected through the GUI configuration, a new boolean flag was introduced. This required some additional code in the GUI configuration reader in order to read this flag. Additionally, the feedback handler in the Android GUI modules was modified to query the global GUI configuration for this flag and to use this to select the type of feedback screen created once a feedback message is received. This has been implemented in such a way that the default text feedback screen is used if the flag is not present at all in the GUI configuration file, ensuring that old configuration files do not suddenly result in errors.

3.3.8 Text-To-Speech Generator

Once the main functionality of the ECA feedback screen was implemented, some problems surfaced in relation to the TTS system. One problem was the fact that the TTS was not affected by the C3PO mute button, which was FR 6. The other problem was more serious, as the TTS system was frequently generating errors and crashing.

Looking at the mute functionality first, this turned out to be caused by the fact that the C3PO mute function is not linked to the main Android volume setting, but instead C3PO maintains its own private mute toggle, which prevents its notification sounds from being played. The easiest solution for this problem was to introduce a check within the TTS generator that queries the GUI module for the current mute setting whenever it is trying to play a sentence. Unfortunately, this introduced a dependency on a C3PO class within the Elckerlyc code. While this is not a big problem, it means the Elckerlyc code cannot be merged back into the original Elckerlyc app without modification.

An additional change was made involving the mute functionality. Originally, the mute function in C3PO disabled the notification sounds, but also disabled the vibration signal. While this functionality was fully intended and documented, it was decided that it would be more practical to have the vibration active even if sound was muted. Otherwise, users would not be alerted of new feedback messages in any way if the sound was muted. A very small modification was made to the C3PO system to allow vibration signals to occur even when the system was muted.

As for the TTS system generating errors and/or crashing, this was a problem that had already arisen earlier in the development of the original Elckerlyc Android app. In order to ascertain the duration of an utterance, the utterance had to first be synthesized to a sound file. This sound file is then read by the Android MediaPlayer facility, which in turn can determine and return its duration. This combination between the TTS and MediaPlayer facilities was the source of many different problems, all of which were mainly related to file permissions on the generated temporary sound files.

After TTS problems started to surface again during the development of the final application, the decision was made to remove all non-essential functionality and to modify the TTS generator to be as robust as possible. Originally, each utterance was generated and saved to a different temporary file, which was registered with the TTS facility afterwards to allow it to be used directly if the same utterance was encountered again. This process was removed altogether in order to avoid file access conflicts. In the new situation, a temporary file is generated, the utterance is synthesized to it, read by the MediaPlayer, and then immediately deleted again. Also, since both the TTS and MediaPlayer facilities are potentially subject to system resources being unavailable and other possible unexpected problems, a fall-back was introduced that set the duration of an utterance to an arbitrary value (3 seconds) in case the process failed. This allowed execution to continue uninterrupted, albeit with erroneous timing information. However, playback of a message with incorrectly synchronized visual animations is preferable to the entire system crashing.

3.3.9 Feedback Configuration

At this point the ECA feedback screen was fully functional, but one additional modification to the overall system still needed to be made. Because our experiment focuses on office workers, it uses feedback messages that are applicable to an office environment. However, because physical activity outside work is also important, and to maximize the participants' exposure to the system within our limited testing period, they are also required to use the system outside office hours. This is the main reason behind FR 10, which requires separate sets of feedback messages for home and office environments.

Feedback messages are defined within the configuration file of the basic feedback module of the C3PO system. This configuration file is in an XML format, and several parameters are associated with each feedback message. This made it possible to simply add an extra (optional) parameter indicating the environment in which a message is applicable. Implementation of this additional parameter required some changes to the basic feedback module. The configuration reader was altered to allow it to read the new property from the configuration file, and this was then saved as a new property of a message. Additionally, the message manager (the subsystem responsible for choosing an applicable message from the list) had to be altered to consider the new property when selecting messages. In the current implementation, messages classified as "home" are only included in the selection process after 17:00 and in the weekends, whereas messages classified as "work" are only included on weekdays before 17:00. Any message not classified as either home or work is eligible for selection at any time and day.

3.4 Additional Configuration

The previous section describes all the programming work done to develop the final application. After the software was finished however, there was still work to be done in setting up the exact parameters and configuration for the experiment. Note that this section only describes the remaining work in configuring the software. The general setup and parameters of the experiment are discussed in Chapter 4. The configuration and setup of the software was done with the following thought in mind: while each aspect brings with it important considerations in regards to physical activity promotion, the focus of the experiment is on evaluating the difference between the ECA and text feedback, and everything else should be designed so as to provide minimal interference with this evaluation.

3.4.1 The ECA

The actual ECA used in the software, by which we mean the images and corresponding Elckerlyc bindings, was left mostly unchanged from the original Elckerlyc Android app. The visual representation was that of a (drawn) female doctor (seen in Figure 3.5), and her behavior repertoire included (among others) laughing, speaking, and several hand gestures. While a medical professional does have relevance to the field of physical activity promotion, the stethoscope in the original images was removed to make the ECA's appearance more neutral. The only other change was the addition of the slightly disapproving expression to the binding. While the image showing this facial expression already existed, no binding had been made yet.

Figure 3.5: The graphical representation of the ECA that was used

3.4.2 GUI Setup

In C3PO, the GUI module has several parameters that can change the appearance of the main screen. The most important one of these is the display of the activity graph. Because part of persuading users to be more active is making them more aware of their level of physical activity, the activity graph was enabled. This also makes the system generally more interesting, because the alternative would be to show a main screen that is effectively blank.

One additional display option concerns the status panel. The status panel is a small colored panel in the top left of the activity graph that displays the percentage a user's activity level deviates from the reference level over a set period of time. This panel was disabled after the pilot test (discussed in detail in Section 4.6) revealed it to be confusing in our setup.

3.4.3 Activity Reference

In order to compare a user's activity level to a reference level, we obviously need to somehow determine what this reference level should be. This turned out to be a bit of a problem. Since the goal of the system is to persuade users to be more active, the most logical and effective choice would be to have a personalized reference level for each user. This level could be computed by first measuring a user's average normal activity level and then raising this by a small amount. Unfortunately though, the limited time available did not allow for the use of this approach.

Another option would be to use some accepted standard recommended level of physical activity. An example would be the recommendations on physical activity for adults by the WHO [45]. The problem with this approach is that there is currently no conversion scale available for relating the measurements taken from the ProMove 3D sensor node to other quantifiable measures of activity. This means that measured units of activity can only be compared to each other, and not to number of calories burned, steps taken, minutes of moderate activity, or any other such units.

Since the only way to put the units of measurement output by the sensor node into perspective is to compare them to levels measured earlier, there was only one option left for determining a suitable reference level. Previous experimental data would have to be used. Because the development and testing of the C3PO system has generally focused on patients in rehabilitation, most of the existing data does not apply to office workers. Fortunately, a recent experiment had been performed that focused on taking a baseline measurement for physical activity in office workers. The average amount of physical activity found in this experiment was used to base our reference level on. This average was increased by 15% (later increased to 30%, see Section 5.1.2) in order to obtain a level which was supposed to provide users with a challenge, without being so difficult to achieve that it would become frustrating.

3.4.4 Feedback Message Content

Possibly the most important part of the system configuration is the list of actual feedback messages from which the system chooses when providing the user with feedback. These messages are defined as part of the configuration for the basic feedback module.

C3PO defines three main categories for feedback messages: encouraging, neutral and discouraging. Encouraging messages are selected when the user's activity level is below the reference level by more than the (configurable) threshold, and generally include suggestions of suitable physical activities. In our experiment, this threshold is set to 10% deviation. Neutral messages are selected when the user's activity level is within the threshold of the reference level, and normally consist of praise and compliments. Discouraging messages are selected when the user's activity level is above the reference level by more than the threshold, and contain suggestions for activities involving little or no physical activity. This last category exists because it is important for rehabilitation patients not to strain themselves too much. In office workers however, such a strict upper limit to physical activity makes little sense, so the categories have been redefined in our experiment. Encouraging messages retain the same content, neutral messages still contain praise, but also emphasize the need to continue to be physically active, and discouraging messages now contain strong praise and compliments.

Aside from the three categories mentioned above, messages in both the encouraging and discouraging categories are subdivided into major and minor subcategories. The distinction between the selection of major and minor messages is based on an additional (configurable) threshold, which is set to 20% deviation in our experiment. As explained in Section 3.3.9, an additional parameter is also available to indicate messages that are suitable only to a home or work setting. This makes for a final total of 9 categories of messages, as can be seen in Figure 3.6.

The strategy used for formulating the actual messages was based primarily on the approach adopted by RRD in previous usage of the C3PO system. The tone is neutral to friendly as negatively phrased messages are generally less likely to achieve the desired response from users [10]. *Suggestion* is used as the primary persuasive tool in case of an encouraging message, and in case of a neutral or discouraging message conditioning through positive feedback is used. In an attempt to minimize repetitiveness each general message has been phrased in a number of distinct ways. In formulating the encouraging messages (which contain suggestions), inspiration was taken from previous RRD experiments. Additionally, we came up with several activities involving light to moderate physical activity that could be easily performed in either a home or office setting (or both) which would be suitable for most people. This process resulted in the list of messages that can be found in Appendix A. Note that because all participants in the experiment were capable of reading Dutch, the messages are all in Dutch.

Also related (although technically defined in the configuration of the user input module) is the schedule for delivering feedback to the user. Based on previous experience from RRD, the decision was made to provide the user with feedback every hour, starting at 09:00 and ending at 22:00.

3.5 Final System Summary

In conclusion, this section presents an overview of the final system as it was used in our experiment.

The system consists of two pieces of hardware, the HTC Desire S smartphone and the ProMove 3D sensor node (Figure 3.1). The sensor is worn on the user's hip, and the smartphone can be carried by the user in whatever way he or she prefers. The software running on the smartphone is a modified version of the C3PO application developed by RRD. Its main screen shows an activity graph depicting the activity

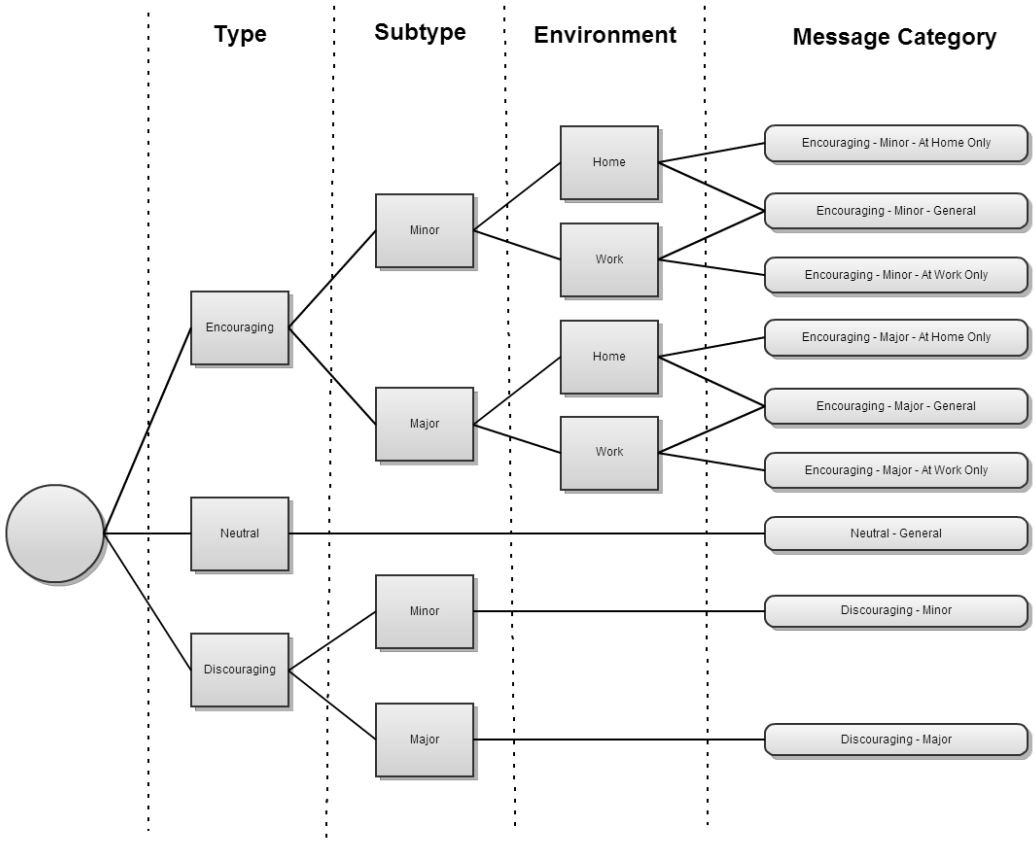

Figure 3.6: The different types of messages and the situations in which they apply

level measured by the sensor node as well as a reference level (Figure 3.2). Each hour (from 09:00 to 22:00) the user receives a feedback message, which is delivered either by a plain text feedback screen or by a feedback screen featuring an ECA (Figure 3.5). The actual message delivered by this feedback screen is selected from the list found in Appendix A based on the measured activity level and whether or not the message is delivered within standard office hours.

Chapter 4

Methodology of Evaluation

This chapter describes the design of the experiment performed to find answers to the research questions posed in Section 1.2. The experiment is performed using the equipment and software described in Chapter 3. This chapter starts with the general setup of the experiment, then discusses each of the sources used to collect data in more detail, continues by describing the full procedure for a single participant, and ends with a description of the pilot test that was performed prior to the main experiment.

4.1 General Outline

As mentioned in previous chapters, the basic idea behind the experiment is to supply participants with the system described previously, and to let them use this system on their own for a period of time, during which their activity data is collected. Afterwards, they will fill out a survey and answer some additional questions in a short interview. All this is done to be able to make a comparison between the software version featuring the plain text feedback messages and the version featuring the feedback delivered by an ECA. The rest of this section describes the most important aspects of the experimental design in more detail.

4.1.1 Target Group

The target group for the experiment is generally referred to as office workers. We use this term in a very general sense. It may be more accurate to say that the target group consists of people with a sedentary profession. That essentially means that anyone who works behind a desk falls within this category.

There are multiple reasons for choosing this group as our target. At first glance, the more obvious choice may have been one of the RRD's patient groups for which the C3PO system was developed. However, letting rehabilitation patients participate in any kind of experiment requires special licenses and permissions which are difficult to obtain. Also, RRD is currently exploring options for entering the general consumer market. Looking at this population of general consumers and combining this with the original intention of improving physical activity levels in people with sedentary lifestyles, office workers make sense as a starting point. Because a sedentary profession is often a significant contributor to a sedentary overall lifestyle [2], attempting to motivate office workers to be more active (even at work) could contribute to overall health.

Within this population of office workers, we will try to obtain a sample that is as homogeneous as possible. There are however a small number of additional limitations. Participants should be relatively healthy, to avoid any physical limitations interfering with the level of physical activity achieved. Also, participants must have at least some interest in achieving and/or maintaining a healthy level of physical activity. While this means that the sample taken may not be entirely representative of the target group (it is to be expected that there are office workers who have no interest whatsoever in their level of physical activity), it should be noted that participants who have no interest in using the system and mostly ignore it are likely to contribute nothing towards the goal of trying to compare two feedback versions. Also, the primary target audience of the system would consist of people who are in the "action" *stage of change*,

but screening for and selecting potential participants based on that criterion is not feasible considering the amount of time and manpower available.

4.1.2 Experimental Design

It should be clear by now that the experiment is meant to determine differences between the two manners of delivering feedback messages. This logically translates to two different experimental conditions: feedback by plain text and feedback by ECA.

Because the potential number of participants is quite low (as will be explained in Section 4.1.3), we have decided to use a within-subjects design for the experiment. This increases the chances of finding statistically significant results. It also allows participants to make an explicit comparison between the two conditions in the final survey and the interview, which can lead to additional insights.

To prevent the order of the conditions from influencing the results, we use a counterbalanced measures design. Thus, half of the participants will start with the ECA feedback version, and the other half will start with the plain text version. These conditions are divided among the participants at random, with one exception: participants who work closely together are assigned to the same condition in order to avoid premature exposure to the second feedback version.

4.1.3 Scale & Duration

The scale and duration of this experiment are each limited by a specific factor: the scale is limited by the amount of hardware systems available for use, and the duration by the time available for carrying out the experiment. In terms of hardware, RRD has kindly made ten systems available for use. As for time, we are limited by the availability of these systems as well as the overall time available for this project. All in all, this allows for a maximum number of 20 participants within a period of roughly seven weeks (allowing time for additional overhead tasks).

One additional issue that arose on this subject is exact definition of a week. Because the reference data was based solely on working days, it was not entirely suitable for usage during the weekend and any other days off. However, asking participants to only use the system on working days raised a number of other issues. First of all, this would either severely limit the pool of possible participants to only those that work 40-hour weeks during the testing period or lengthen the overall testing period if participants were accepted who work less. Also, the point at which the software version changes poses a problem. Because it is impossible to have all participants start on the same day, some would switch versions right after the weekend, while others would switch in the middle of a week. In the end the decision was made to simply let participants use the system for 14 days, including any days off (weekend or otherwise). This was done to maximize the participants' exposure to the system within the time available.

4.2 Activity Data

The most obvious and by far the most extensive source of data is the software. First and foremost, the software records all of the physical activity data that is received from the sensor. While it is unlikely that a significant change in behavior can be observed within the short duration of this experiment, this data can nevertheless be analyzed to possibly find patterns or to corroborate or contradict answers given by the participants in the interviews.

The software does however collect more data than just the activity levels. For one, the data shows at which points in time participants turned the system on and off. The software also maintains logs on the opening of feedback messages. This allows us to determine how often participants actually opened the feedback messages, and how much time elapsed between the participant receiving a notification and opening the feedback message. This information could potentially show a difference between the software versions.

The data collected by the software is clearly the most objective. However, it is also strongly influenced by a large number of uncontrollable outside factors. For example, if a participant shows significantly less physical activity in the second week, this could be because the weather was too bad to go outside much instead of being caused by the difference in feedback versions.

4.3 Surveys

The second source of data consists of the surveys that participants fill out. In total, there are three surveys: an intake survey for at the start of the experiment, a halfway survey for after the first week, and a final survey for after the second week. Surveys are presented to the participants through SurveyMonkey[1], an online service. This allows them to fill out the surveys without requiring additional interaction with the researcher. Complete versions of each of the surveys as presented to participants can be found in Appendix B (in Dutch).

The intake survey is simply meant to collect some background information about each participant that could be relevant. Participants are asked to fill in their age, weight, height and gender. Additionally they are asked to choose from a list of options the situation regarding physical activity that most closely resembles their own. This is done to determine the *stage of change* the participant is in. Finally, participants are asked to indicate how much experience they have in several categories relevant to the experiment. These categories are: smartphones/PDAs, ECAs, human coaching in physical activity, non-interactive self-help products, digital coaching systems and activity measurement systems.

The halfway and final surveys are identical for the most part, as they are meant to evaluate the participants' experiences with the two feedback versions, and results from both will be compared in order to draw conclusions. Both contain several standardized validated questionnaires, which were used in order to maximize reliability. Each questionnaire is directly related to one of the research questions posed in Chapter 1. All questionnaires were part of both the halfway and final surveys, aside from the explicit comparison section, which was only part of the final survey. Within each questionnaire, the order of the items is randomized for each participant. This is recommended for most of the questionnaires to avoid order bias. Each of the questionnaires used is discussed in the remainder of this section.

4.3.1 User Experience

One of the most important concepts in designing any application is the user experience. However, because this concept encompasses so many different aspects and can be very subjective and different for each user, it can be quite difficult to quantify user experience. One of the more commonly used tools for tackling this issue is AttrakDiff2 [46]. AttrakDiff2 splits the user experience concept into four separate constructs: *pragmatic quality*, *hedonic quality - identity*, *hedonic quality - stimulation*, and *attractiveness*.

The actual questionnaire is in the form of a semantic differential scale consisting of 28 bipolar pairs of adjectives which each represent two extremes of a spectrum. The participants are asked to indicate (on a scale from one through seven) which of the two adjectives they feel more appropriately describes the system, and how strongly they feel about this. The number one stands for a very strong association with the first adjective, whereas the number seven represents a very strong association with the opposite adjective. The numbers in between represent feelings that are less strong, with the number four meaning both adjectives are equally applicable.

Because this test has a strong psychological character, participants are asked not to spend a lot of time thinking about their answers, but rather to choose to give their responses based on their initial feelings and instincts. Participants are also asked to give an answer on each of the adjective pairs, even if they intellectually feel that neither of the adjectives applies to the system. In these cases their (possibly subconscious) evaluations may still be valuable.

Aside from the item order randomization applied to all questionnaires, a number of the adjective pairs are reversed, so that there isn't a "positive" and a "negative" column. This is done in an attempt to get participants to really interpret the concepts properly, and avoid acquiescence bias.

4.3.2 Credibility

One of the advantages of using ECAs in interfaces that has been reported before is an increase in perceived credibility [33]. Since credibility is an important factor in the success of persuasion attempts [14], this is an interesting aspect to investigate. In order to measure perceived credibility, McCroskey's Source Credibility Scale (SCS) [47] is used. Originally, the SCS was developed to assess the credibility of human speakers. Since then, it has been shown to be applicable to many kinds of different information sources, including computer systems.

[1] `http://www.surveymonkey.com`

The SCS uses the exact same format as AttrakDiff2, presenting the participant with pairs of bipolar adjectives and a seven-point scale to choose between them. Two versions of the scale exist, a 12-item version which contains the constructs authoritativeness and character, and a 15-item version which contains the constructs sociability, extroversion, competence, composure and character. The reason we chose the 12-item version is that we feel that the adjective pairs in that version are more applicable to computer systems in general, and to our system specifically. As with AttrakDiff2, some of the pairs have been reversed.

4.3.3 Acceptance

One of the research questions concerns the potential beneficial effect of an ECA on the long-term duration of use of the system. It is quite difficult to make a prediction on this subject. We have neither the time nor the number of participants needed to truly assess the long-term retention rates of the different versions of the system. There are however tools that allow us to measure some of the indicators of long-term usage. We have chosen to use the Unified Theory of Acceptance and Use of Technology [48] (UTAUT) for this purpose.

The UTAUT model presents a questionnaire for determining several aspects relevant to the expected acceptance and use of technologies. It is divided into a number of different constructs dealing with these aspects. Each construct consists of a small number of statements. The participant is asked to indicate his or her level of agreement with these statements on a seven-point scale.

We have chosen not to include all of the constructs in our survey. This is because some of the constructs are not quite applicable to our situation, and more importantly because we can reasonably assume that answers on some of the constructs will not be different between the two feedback versions. This means that the constructs of *performance expectancy*, *social influence*, *facilitating conditions*, and *self-efficacy* have been omitted in our survey.

Performance expectancy has been omitted because it makes little sense in the context of our system: there is no real reason to think that a physical activity coaching application will improve work performance for office workers. The *self-efficacy* construct also has little meaning in the current design of our system: the system is incredibly simple, and only provides information to the user. It can not really be used to perform some kind of specific task, nor is it really possible to fail at using it. There is also no reason to assume a difference between feedback versions on this construct, as the operation of the system is virtually identical. The argument for omitting the *social influence* and *facilitating conditions* constructs is a little different. Each of these constructs could be said to apply to our system in general. Social influence certainly plays a big role in the motivation for increasing one's physical activity levels. However, it is reasonable to assume that the feedback version has no influence on what other people think of using the system (with which they are not even familiar). This is even more clear for the *facilitating conditions* construct: feedback version certainly has no influence on the available resources, knowledge or assistance.

These omissions leave us with the constructs *effort expectancy*, *attitude toward using technology*, *anxiety*, and *behavioral intention to use the system*. Each of these constructs contain at least some questions that may reveal differences between the two feedback versions. Looking back at the original reason for using this questionnaire, the final construct (*behavioral intention to use the system*) is probably the most relevant, as it directly gages the participant's interest in using the system for an extended period of time.

4.3.4 Coaching

Because we are testing a coaching system, another relevant measure is the quality of coaching. Although both of the software versions are technically identical in the actual coaching given (the feedback messages), quality of coaching pertains to the entirety of the coaching experience received by the participant. Differences between the feedback versions are certainly possible in this area because participants may experience a more personal connection with an ECA coach, and therefore can judge certain aspects of coaching differently.

To quantify quality of coaching, no widely used tools such as the ones previously discussed exist. The questionnaire that comes closest to this goal is the Coaching Behaviour Scale for Sport [49] (CBS-S). However, this questionnaire is focused on coaching received from a human coach in the context of sports.

While physical activity is clearly related to sports, the coaching offered by our system does not target athletes, nor does it promote heavy exercise. This means that the CBS-S is not suitable for use in its original form.

Luckily, a modified version exists that targets digital coaching in the physical activity domain. This version was developed by Phillips as a part of the DirectLife project. Unfortunately, no documentation or publications exist on the subject of this modified questionnaire. We decided to still use this questionnaire, because it is largely directly applicable to our situation. It would serve us no better to create our own modified version of CBS-S.

The Philips DirectLife Coaching questionnaire consists of 21 statements (CBS-S contains 47) about the behavior of the coach. The participant is then asked to indicate, using a seven-point scale, how often the coach exhibits those behaviors. Of the 21 statements, we have chosen to omit several in our survey, because they pertained to qualities we can not reasonably assume our software to convey. For example, the statement "my coach is a good listener" simply does not apply, the software does not listen or react to the participant in any way. We have been fairly lenient in our choices however, keeping any behaviors that participants could conceivably experience, even though they are not technically exhibited by the system.

4.3.5 Explicit Comparison

In the final section of the final questionnaire, participants are asked to make a direct comparison between the two versions of the system. The questions posed here are not part of any existing questionnaire, but are instead formulated specifically for our situation. In fact, each question relates more or less directly to one of the research questions we posed in Section 1.2. In this questionnaire participants are presented with a number of statements, and asked to indicate which of the two feedback versions they feel the statement applies more strongly to. This is done using a five-point scale. The statements relate to different areas, some overlapping with the earlier questionnaires. This overlap may in fact prove interesting when comparing the results of this section to the results of the other questionnaires.

4.4 Interviews

The last source of data is the interview process, conducted during the debriefings. The main reason for conducting in-person interviews is the low number of participants. This makes it feasible to conduct an interview with each participant, something which costs too much time for larger groups. Also, interviews provide more qualitative results which are potentially quite valuable considering the possibility that the small number of participants results in not being able to find statistically significant differences between the feedback versions in the quantitative results.

The open interviews were relatively informal and loosely structured. Because the surveys already collect data very specific to the experiment's goals, the interviews are more focused on finding the aspects of the system that stand out to participants. It is mainly for this reason that the interviews are not very strictly organized. In order to adhere to this strategy, no list of questions was formulated beforehand. Instead, the general subjects that should be discussed were listed. These subjects were the following:

- General impression of the system
- Practical problems experienced
- Participant's and reference activity levels
- Feedback message content and frequency
- Difference between feedback versions
- Improvements needed for commercial viability
- Additional comments

4.5 Procedure

This section discusses the procedure of the entire experiment for a single participant. It indicates all the interactions with the researcher, all the actions taken by the participant and all documents involved. This process starts at the point where a potential participant has already expressed the intention to participate in the experiment.

4.5.1 Introductory Explanation

The introductory explanation is scheduled once someone has agreed to participate. At this point the potential participant has already received the general information sheet (found in AppendixC.1, in Dutch). This sheet contains a general explanation of the experiment, the system being tested, the data collected, and expectations that are placed on participants. First of all, the researcher asks the potential participant if he or she has any remaining questions about the experiment in general after reading the provided information.

Once any questions have been answered, the researcher starts with the explanation of the daily procedure. While usage of the system is fully explained, the participant is also given a system manual in booklet form. Explanation of the system consists of a demonstration of all the steps required to turn on the system in the morning, read the main screen, receive feedback messages, and turn the system off again at night. A demonstration of the correct way to handle and wear the activity sensor node is also given.

After the system demonstration, the participant is given a copy of the journal to be filled out daily (found in AppendixC.2, in Dutch). The journal is a compact form on which the amount of time spent in the office and any technical problems experienced can be filled in for each day of the testing period. The use of this journal is then explained, along with the surveys and the days on which they are to be completed. Afterwards, the participant is given an informed consent form (found in AppendixC.3, in Dutch). The consent form is fairly standard, indicating that the participant is adequately informed about the experiment, allows the use of collected data, and has received and will return the required equipment. The signing of the consent form by both the participant and the researcher concludes the introduction.

4.5.2 Testing Period

The testing period starts the day after a participant receives the introductory explanation, and lasts for 14 days. Each day, the participant wears the system throughout the day. Whenever technical or practical problems prohibit the participant from wearing the system for any significant period of time, the participant makes a note of this in the journal. At the end of each day the participant fills out the number of hours worked in the journal.

On the first, seventh and last day of the journal, there are notes asking the participant to fill out the appropriate surveys. Since the first survey is only about background information it is not urgent and can be completed at any time during the first few days. For the halfway survey, it is important that participants complete it before experiencing the second feedback version. In order to assure this, the journal emphasizes that this survey must be completed on day seven. The researcher also sends an email reminder to the participant on day seven explaining this. A similar situation applies to the final questionnaire. Because participants could be influenced by the debriefing interviews, they are required to complete the final questionnaire before the interview takes place. They are reminded of this by email on the final day of the testing period.

4.5.3 Debriefing

The debriefing is the final step in the experiment, which takes place after the testing period has been completed. The researcher meets with the participant in order to check and retrieve the equipment as well as the journal. After this, the interview (discussed in Section 4.4) is conducted. Finally, the participant has the opportunity to ask questions about the experiment and the research in general if desired. Once this final meeting is completed, the participant is completely done.

4.6 Pilot Test

In order to test both the software and the procedure, a pilot test was performed prior to the actual experiment. This pilot test was a shortened version of the final experiment, performed by only one participant, in four days instead of two weeks (two days for each feedback version). Because the actual results of the surveys in the pilot test are irrelevant, the halfway survey was skipped. The final survey contains all the same questions, so any problems with those questions should still have been uncovered. Results from the pilot test were used to make adjustments to the software and/or procedure as needed. The observations made during the pilot test, and any subsequent changes made to the system or procedure are discussed here.

The first notable observation already occurred after the first day of the pilot test. The tester commented that the reference activity level was far too easy to achieve. He claimed to have been only moderately active during the day, and still recorded an activity level well above the reference level. At this point the reference level was set to 110% of the original average (see Section 3.4.3), and the tester had totaled at least 150% of this reference. To remedy this, the reference level was upped to 165% for the remainder of the pilot test. This level proved too high however, as the tester's measured level was well below it for the remaining days of the test. One factor that may have contributed to the much higher measured level on the first day was the fact that the tester had worn the sensor node in the pocket of his pants, instead of on his belt. This was taken into account in the explanations in the final experiment, where it was explained to each participant that the sensor should be worn on the side of the hip.

The second important observation was regarding the display of the deviation percentage in the corner of the activity graph. The tester remarked that this percentage was confusing in two different ways. First of all, it was quite unclear what the percentage was actually based on, making it difficult to extract any information from it. Secondly, the panel that displayed the percentage changed colors based on the deviation, but an activity level well above the reference level was indicated with red. This caused additional confusion and frustration because the tester did not expect a color with negative connotations when he was in fact doing very well. As noted in Chapter 3, this issue was rectified in the final experiment by simply disabling the display of the status panel and deviation percentage entirely.

Aside from the two issues mentioned above, the pilot test revealed no significant problems. No technical issues occurred during the pilot test. Also, the tester noted that the frequency of messages was acceptable, and that the content of these messages was not so repetitive as to become annoying. He did mention that it may be preferable for the feedback message frequency to be more dynamic, depending on the deviation from the reference level. This suggestion did not result in a change to the system for several reasons. Most importantly, we wanted each participant to be exposed to a significant amount of feedback messages within the relatively short duration of the experiment in order for them to properly form an opinion on the different versions of feedback. Also, implementing such dynamic feedback frequencies would have required significant additional development time, which was not available.

Finally, the pilot tester provided some observations about the system that gave some indication of the types of comments to expect from participants in the final experiment. He noted that he started to ignore messages more frequently after some time, and that this was partially because he already had physical activities planned later in the day and therefore knew he would make up for ignoring some of the suggestions. He also noted that he would not be interested in using the system over an extended period of time, but rather that it may prove useful as a tool to sporadically use for a short period (for example one week every two months), in order to gain awareness of one's physical activity levels and build up habits where needed.

Chapter 5

Results

This chapter discusses the execution and results of the experiment described in Chapter 4. First, the overall process is described, along with any relevant general information about the execution of the experiment. After that, each of the three sources of collected data will be discussed individually. We start with the interviews because analysis of those is most straightforward. Also, observations made in the interview results can then be used to run specific tests on the remainder of the data. Next we discuss the survey results, on which we perform statistical analysis in the hope of finding answers to our research questions. Finally we present an analysis of the data collected by the software.

5.1 Process

The experiment was carried out over a period of eight weeks in total, in the area surrounding the University of Twente. Overall, the experiment was carried out in the manner it was designed, and we encountered very few significant problems. The reason that execution of the experiment took up more time than initially planned is twofold: at the start of the experiment it proved more difficult than expected to find participants, and at the end there were some problems scheduling the final few debriefings, after the actual testing periods had already been completed.

5.1.1 Participants

Participants for the experiment have been found through a variety of methods. Our main focus was on asking people in person. This allows for a better explanation of the experiment, and is also generally more effective than less personal methods. We have also sent out several e-mails asking for participants through a number of internal company and university mailing lists. Finally, several participants referred us to others who were potentially interested in participating.

Because of the personal contact required, our search was focused on finding participants based at or close to the university campus. Our main search areas were therefore the university itself and the RRD building (which is located close to the university campus). Within this constrained environment we have still found participants employed in a varied set of sedentary professions. This includes research-, management-, and administrative personnel as well as graduate students.

Most, but not all, participants had a Dutch background, and all were able to understand written and spoken Dutch. Participants were aged 22 to 61 ($M = 37, SD = 13.3$). Of the 14 participants, 8 were male and 6 were female. Participants had Body Mass Indexes (BMIs) ranging from 16.3 to 26.2 ($M = 22.06, SD = 2.78$). Seven participants were assigned to the text-first condition (meaning they experienced text feedback in the first week and ECA feedback in the second), and seven to the ECA-first condition. See Table 5.1 for more background information on the participants and the ECA-first and text-first groups.

All of the participants completed the experiment, meaning they used the system for the full 14 days (except in cases of technical or practical problems), submitted all questionnaires, and took part in the debriefing interview.

Table 5.1: Participant data

Condition		**Text-first**	**ECA-first**	**Total**
Count		7	7	14
Sex	Male	4	4	8
	Female	3	3	6
Age	Average	38	36	37
	Minimum	24	22	22
	Maximum	55	61	61
BMI	Average	20,9	23,2	22,1
	Minimum	16,3	20,2	16,3
	Maximum	26,0	26,2	26,2
Stage of Change (counts)	Precontemplation	1	1	2
	Contemplation	2	2	4
	Preparation	2	1	3
	Action	0	0	0
	Maintenance	2	3	5
Prior Experience (averages)	Smartphones	8,3	8,7	8,5
	Human Coaching	3,4	4,6	4,0
	Self-Help	2,6	5,0	3,8
	Digital Coaching	4,6	5,3	4,9
	ECAs	4,9	5,1	5,0
	Activity Monitoring	3,9	5,7	4,8

5.1.2 Problems

The experiment was executed without major problems, with the exception of one technical issue. All but three participants managed to complete the testing period without any additional assistance or support. One participant contacted us twice with minor questions which were easily resolved over the phone. No problems were reported with the operation of the software or the use of the smartphone.

The only significant technical problem occurred in the second week of one of the participants. This participant's sensor node started causing trouble. It appeared to only be charging partially, and it only lasted to the end of the morning. We were unfortunately unable to resolve this problem, so for the remaining days of the testing period, this participant was only able to use the system for a limited time until the sensor node's battery ran out. Because this participant had still managed to complete the test period for the most part, and because some periods where the system was inactive occurred with other participants as well (although not for technical reasons), it was not necessary to discard the data from this participant.

The only other issue was that one of the first few participants contacted us with the observation that the measured activity level was well above the reference level, even while this participant claimed not to have been very active. This prompted a reassessment of the reference activity level, which we subsequently decided to raise from 115% of the original average to 130% of the original average. Because the participant who had brought this to our attention had only used the system for two days during the weekend at that point, we decided to change the level for this participant, as well as for any participants starting from that point on. The three participants who had started earlier were left with the old reference level in order to avoid interference from the changed reference level on the results.

5.2 Interviews

Because of the relatively limited amount of participants, we were able to conduct a short interview with each of them during the debriefing. Most of the debriefings took place one or two days after the participant in question had finished the testing period, but in some cases the interview took place up to

about a week later because of the participant's availability. The interviews lasted roughly ten to twenty minutes.

All interviews were recorded (audio only) using a smartphone application. Afterwards, these recordings were used to produce a written summary of the interview. No full transcripts were produced, instead all relevant comments by the participants were written down in a compact form. Because of the structure of the interviews, these comments were mostly sorted by subject already, but any comments that were out of place were moved into the appropriate category. Finally, we analyzed the comments topic by topic. The findings of this analysis are presented here.

5.2.1 General Impressions

Each participant was first asked for their general impression of the system as a whole. Because this is quite a broad question, many answers given by the participants actually related to other subjects discussed in the interviews, and have been moved there for analysis purposes.

In general, the system was received relatively well: nine participants indicated that they generally found the system at least moderately enjoyable to use. One of these expressed very positive feelings towards the system, while the other eight were positive but not overly enthusiastic. One of these eight also indicated that using the system did become slightly annoying after a few days. There were also some negative opinions: three participants indicated that they did not really find the system useful. Two of them viewed this as the system not being particularly suited for their personal situation, whereas the third was of the opinion that the system was completely ineffective in general and also noted that it was not intelligent. The final two participants did not directly indicate positive or negative feelings towards the system, but rather explained that they already had previous knowledge about the C3PO system (these were both RRD researchers). One of these had also used the system before.

5.2.2 Practical Problems

The first specific question posed to the participants asked if they experienced any practical problems during the testing period. This includes both hardware and software issues, as well as problems carrying the hardware. In general, almost all participants claimed to have experienced no significant hardware or software problems, although most did have some comments on the practicality of wearing the hardware.

The few technical issues that did occur were limited to the sensor node. Three participants reported that the sensor node's battery occasionally ran out before the end of the day, one of which was the participant who had the problem discussed in Section 5.1.2. Two participants also experienced some problems with the connection between the sensor node and the smartphone, which at times resulted in the connection not being established until up to a few hours after switching the system on.

The majority of complaints about practicality concerned the sensor node, and the requirement to wear this device on the hip by use of a belt attachment. Five participants did also comment on the fact that carrying around an extra smartphone was impractical, but in general everyone was able to see past this fact since it was clearly related to the experimental situation. Nine participants made at least some comment about the sensor node being impractical or irritating. Four of these nine had experienced problems with the sensor node bumping into or being caught behind objects in certain situations. For three of the nine, wearing the sensor node just became generally annoying at some point during the evening, another noted specifically that the sensor node became annoying when sitting on a couch. Two out of the nine had solved their problems with wearing the sensor node by occasionally wearing it in their pants pocket instead of on the hip. Finally, two of the nine participants occasionally had a problem with the sensor node not because it was uncomfortable to wear, but because it was very visible when worn on the hip.

On the bright side, most of these issues either did not occur very frequently or simply did not have a big impact, because only one out of all the participants reported being significantly bothered by having to wear the system all day. Also, six participants reported no real discomfort in wearing the sensor node. Two of these mentioned that they actually forgot they were wearing the sensor node after a while, and one even wore the sensor node during sports.

Some small issues also existed regarding the daily usage of the system. One participant noted that it was difficult to remember to attach the devices to the charger each evening, and three participants admitted that they had forgotten some part of the procedure at least once. This included forgetting to

charge the system at night, forgetting to wear the system right away in the morning, and forgetting the smartphone altogether. Aside from the problems in the practical domain, two participants actually also reported a positive side effect of wearing the system. They found the visibility of the sensor node to be a good thing, as it prompted colleagues and others to inquire about the system.

5.2.3 Activity Levels

The next subject discussed during the interviews was the activity level. This includes both the reference level and the actual level achieved by the participants.

Unfortunately, most participants (twelve out of fourteen) indicated that the reference level appeared to be too low for them. Of these twelve, five indicated that their measured level was virtually always above the reference line. There were also five participants that suggested that this discrepancy was being caused by the commute to work (by bicycle or foot), which resulted in a significant "head start" on the reference level each morning. One of these five participants even decided to make the system more challenging by not turning it on until after the morning commute. Two of the twelve participants felt that they were not very physically active at all, even though their measured level was regularly above the reference level. It became clear that activity levels in the weekend vary greatly: one participant claimed to be significantly further above the reference level during the weekend, while another said that the weekend was the only time the measured level was below the reference. Interestingly, one of the participants, who was virtually always above the reference level, did feel that the reference level corresponded with the common standard of 30 minutes of physical activity a day.

Aside from the overall height of the reference level, there were also some comments on the curve itself. One participant found it strange that the reference line became steeper after working hours, whereas another indicated that the line should have been even more steep in the home periods. Three participants did say that they found the steepness of the reference line to be appropriate during office hours, even if it was too low in general. One of the participants commented that the reference level should not be linear at all, because people are physically active in short portions at a time, followed by periods of inactivity, especially at the office.

Several participants also reported problems with the start and end times that were used. In our configuration the system started at 07:00 in the morning and the graph ended at 23:00 in the evening. The start time proved to be too late for two of the participants, as they regularly got up and were physically active before that time. One of these participants said that this caused feelings of frustration and even slight anger towards the system. The end time proved to be too early for two other participants. One of these simply turned the system off at that time, while the other continued working with the system. This resulted in the software simply extrapolating the reference level past 23:00, which is something we were not aware would happen. The participant in question also found this behavior to be questionable.

The preset starting time also caused a few issues in another way. It turned out that when the system is switched on after the starting time, it assumes that the user has been active equal to the reference line up to the point of switching on. This is intended functionality, and is mainly used to fill in the gaps whenever the connection to the sensor node is temporarily lost. Unfortunately, we did not anticipate that this would be (unknowingly) exploited by participants who got up after 07:00. One participant actually noticed this behavior by the system and reported it as strange. Two others were not aware of it, but reported that they were only ever below the reference level when they had forgotten to switch the system off the night before, meaning they had "profited" from this functionality.

Regardless of the reference level, three participants did mention that it was interesting to be able to see their own activity level and patterns. Four participants found it reassuring to see the system confirm that their activity levels were high enough, even though two of these four had expected their activity level to be too low. This was a slightly undesirable result, because the reference level does not actually represent any known standard for a healthy physical activity level (see Section 3.4.3 for a discussion of why). Obviously we do not want participants to be under the impression that they are getting enough physical activity when they are in fact not. We have tried to remedy this by explaining to these participants that the reference level did not necessarily represent a healthy level.

Regarding the motivation to become more physically active, three participants did explicitly state that they had wanted to increase their physical activity. Four others had indicated that they felt they were already active enough before starting the testing period. When asked if they had become more active during the testing period, seven participants replied that they had not, five participants were unsure or

thought that they may have been slightly more active, and two stated that they had certainly become more active. Something that was implied from the line of questioning, and also explicitly confirmed by one of the participants, is that these increases in activity were more due to the system causing participants to be aware of their physical activity than due to the reference level and feedback. When asked what activities they had performed to become more active, responses included cycling to work, taking a walk during smoking breaks, and walking to colleagues instead of calling them.

5.2.4 Feedback Messages

In terms of the feedback messages, participants were asked for their thoughts on the contents of these messages and the frequency with which they were delivered. Unfortunately, since most participants were so frequently above the reference level, they mostly received feedback messages containing compliments. Most did receive at least some suggestions, but it would have been preferable to see more balance between the two.

When asked for a general evaluation of the feedback message contents (both compliments and suggestions), nine of the participants responded positively. When speaking about the content of the suggestions specifically, there were three participants who felt the suggestions were well chosen and suitable for an office environment. Three other participants however stated that the suggestions did not appeal to them. Explanations included that the suggestions actually were not suitable for an office environment, and that they were not interested in specific suggestions at all.

In regards to the compliments, opinions were also divided. Most participants expressed fairly neutral sentiments towards the compliments, but there were some specific comments. From the line of conversation it became clear that the messages containing compliments became fairly repetitive to most participants, but only two commented explicitly that they found the variation in the compliments to be lacking. Three participants also stated that they were unable to take the compliments seriously (after a while). One participant added that this was mostly because the system was giving a compliment while the participant did not feel like the measured activity level merited a compliment at all. There were also some positive comments on the compliments. One participant mentioned that receiving a compliment from the system did make it more enjoyable to undertake some form of physical activity. Also, one specific compliment clearly stood out to some participants: four participants mentioned that they found the message "You deserve a pat on the back!" to be very amusing. One participant even reported asking someone in his/her direct environment to give him/her said pat on the back.

When asked about the frequency of the feedback messages, nine participants replied that receiving a feedback message every hour was appropriate. The other five participants felt that it was too high. Most of these five did comment that this was at least partially because they already knew the message would contain a compliment, so the fact that the reference activity level was too low for most participants certainly played a role here. The fact that feedback messages were generated at the top of every hour had an interesting side effect: two participants mentioned that this regular notification made them more aware of the time and appreciated this effect.

While discussing the frequency of feedback messages being generated, the topic of ignoring notifications also tended to come up often. Eleven participants admitted that they had ignored some of the notifications they received, although one of them claimed to still have checked all the messages at a later point in time. Five of them pointed to being occupied with more important matters as the cause for ignoring notifications. Examples mostly included work situations such as meetings and deadlines. One participant even mentioned that the notifications were irritating when under stress from work. Another cause for ignoring notifications, mentioned by five other participants, was more related to the activity level. They indicated that they already knew they would meet the reference level for the day, either because they were already above the reference line, or because they had some form of physical activity planned later in the day that they knew would lift them over the reference line. This last comment is an important point which also came forward in other places during the interviews. People tend to have a fairly good idea of what the rest of their day is going to look like, and knowing that they will be significantly physically active later (for example because of the commute home or planned sporting activities) is often cause for ignoring the system altogether.

5.2.5 Differences Between Versions

Although the interviews did not specifically center around this aspect of the experiment, the difference between the feedback versions (text and ECA) is clearly the focus of this research in general. In the interviews, participants were asked which feedback version they preferred and why, and they were also asked for their opinions on the ECA (regardless of which version they preferred).

When asked directly which version they preferred, three participants chose the ECA feedback version. Ten indicated a preference for the text feedback, and one participant claimed to have no preference at all. The participants that preferred the ECA version gave the following explanations: it is more personal, it makes it easier to establish a connection with the system, and it is funny. The participants who preferred the text version gave reasons that almost all came down to the same thing: *glanceability*. An interface being *glanceable* can be defined as: "enabling users to understand information with low cognitive effort." [50, p1]. Because the ECA feedback screen displays the text letter by letter as spoken by the TTS, it is clear that the text feedback screen, where all the text is displayed at once, is much more *glanceable*. The fact that watching the ECA utter the entire feedback message takes more time than just reading two lines of text bothered many participants: seven of those who preferred the text version gave this as the main reason. Two others mentioned that they found the speech annoying in general, and the last participant experienced the text feedback as more convincing.

When the participants were asked what they thought of the ECA, we got a wide array of replies about different aspects. Starting with the positive comments, four participants said that they found the ECA to be fun and enjoyable. Maybe a bit too enjoyable even for one of these participants, as this participant reported finding the ECA so amusing that it caused a break in concentration during work. Aside from the fun-factor, three participants did find that the ECA made the system more personal, and that it added a sort of personality to the system, although one of them did mention that this was a very "flat" personality. One participant was also of the opinion that the ECA had more right to give the user a compliment than a simple text screen. Finally, one participant actually reported that the ECA feedback was more effective at stimulating physical activity.

There were also numerous negative comments about the ECA. The opinions most frequently heard were that the ECA does not add anything significant to the system, and that the ECA was not believable because it was not a real person. Both of these comments were given five times. Another comment, given by three participants, was that the ECA is a bit of a gimmick, and that it loses its appeal quickly. On the point of the enthusiasm displayed by the ECA we received two opposite opinions: two participants thought that the ECA does not show emotion and is not enthusiastic enough, whereas another participant found the ECA to be too enthusiastic, to the point of it not being believable.

There was one other comment about the ECA itself that was neither directly positive or negative. Regardless of our efforts to hide the fact that our ECA was originally designed to be a doctor, two participants did mention that they identified her as a medical professional. Interestingly enough, one of these two felt that this was a fitting identity for an ECA commenting on physical activity, while the other one did not, and would have preferred something along the lines of a fitness instructor.

During the discussion of the ECA, a practical matter also came up in most interviews: the sound produced by the TTS. Five participants experienced the spoken text as annoying, often stating that it is impractical in social situations. We did not specifically ask the participants if they had ever muted the sound, but four participants did mention that they had muted the sound most of the time, and two others had muted it while at work. It is possible that having the sound muted has contributed to the irritations caused by the lack of glanceability mentioned earlier. This is because the text is displayed at the same speed as the spoken text, but if there is no spoken text the speed at which the visible text appears may just seem arbitrarily slow.

5.2.6 Possible Improvements

After discussing the experiences participants had with the system, they were asked what changes they felt needed to be made to the system in order for it to become viable for commercialization. A large number of suggestions was given, of which the most frequently heard and most relevant ones are discussed here.

Clearly the most obvious area for improvement is on the hardware side. It was obvious that most participants shared the opinions noted here, even though not all of them explicitly commented on the hardware aspect of the system. Seven participants suggested that the sensor node needed to be either

smaller or built into the smartphone as its current size is simply impractical. Six participants suggested that the application should run on the user's own smartphone instead of on a separate device, which clearly would make the system more practical. Two additional comments about the hardware were made: one participant felt that the range of the connection between the smartphone and the sensor node should be increased, and another pointed out that this wireless connection should be investigated to see if it is harmful to carry around the devices for such extended periods of time.

The second area of comments was the reference activity level. Four participants explicitly suggested that the reference level should be customized for each user, although from other comments about the reference level it is not unreasonable to assume that most participants agreed with this sentiment. One participant went a bit further than just a customized reference level, and suggested introducing a level of intelligence to the system. This participant also pointed out that it may be beneficial to handle periods of intense activity differently in order to keep users active throughout the day, because the current situation sometimes leads to users simply ignoring the system all day when they have a sporting activity planned in the evening.

Also relating to the activity data, there were some comments on the way the information is presented to the user. Four participants indicated that they would like to be able to have a broader overview of the collected data, enabling them to compare several days or even weeks instead of only being able to view today's data. Another suggestion was to present the activity level in some sort of familiar quantity such as calories burned or steps taken. Two participants made that suggestion, and also commented that this would help set individual goals because it would make users more aware of what they are actually trying to achieve.

Next there were a number of suggestions regarding the appearance of the ECA feedback screen. Three participants suggested that the ECA would be more effective if its appearance was more personalized towards the user. Suggestions included making it more relevant to the target audience and even letting users customize their ECA themselves. One participant also suggested that a more visually appealing 3D-rendered ECA would have more impact. While not directly related to the appearance of the ECA, one participant suggested the option of letting users choose between ECA and text feedback themselves, because some people are simply more visually oriented than others.

While it is not a direct suggestion, four participants indicated that the system would only be effective for users who have an intrinsic motivation for increasing their physical activity, such as overweight people committed to losing weight and patients of physical therapists. The implicit suggestion here is to focus the system on an audience that does have such intrinsic motivation. Another interesting comment in this area came from a participant who suggested that the system would be more suited for short periods of use on a somewhat regular basis (for example one week every other month) as a sort of tune-up, instead of continuous use over an extended period of time. A similar suggestion was also made during the pilot test (see Section 4.6).

Finally, there were suggestions concerning the feedback messages themselves. One of the participants suggested introducing more variation into the set of messages, although this is a rather obvious area for improvement. More interesting was the suggestion by another participant to stop using a rigid time schedule for the feedback messages and let the activity level influence the frequency of feedback, so that the user is not continuously interrupted when everything is going well.

5.2.7 Additional Comments

At the end of the interviews, participants were asked if they had any other comments or questions about the system or the experiment. Most participants did offer at least some comments, but they will not be discussed here. A large portion of the comments given actually applied to one of the previous categories, and was processed there during the analysis. Any other comments were outside the scope of the current research.

5.3 Surveys

As mentioned earlier, the surveys were completed by participants through the SurveyMonkey on-line service. Every participant except one submitted the halfway survey on day seven of the testing period

(the day before the second feedback version was activated), with the one exception submitting the survey on the morning of day eight. All participants completed the final survey before the debriefing.

Because the survey was divided into five separate sections, each containing a different questionnaire, the results are presented for each section separately. In order to analyze the results from the surveys we focus on finding differences between the two feedback versions. In order to do this statistical analysis is performed on the answers collected.

5.3.1 User Experience

In order to assess the possible differences in user experience between the two feedback versions, we used the AttrakDiff2 questionnaire. After processing all the submitted data, the graph in Figure 5.1 was generated as an overview of the results and a starting point for the analysis. We can see that the evaluations of the text and ECA versions of the system are not far apart in most cases. There are some visible differences, but we can not tell whether these are significant just by looking at this graph.

What we can tell from this graph and the corresponding mean values, is that while most adjective pairs score above the median value of the answer range (4), the majority still falls within the range of average scores (3 to 5). The only strong exception is the "undemanding - challenging" pair which scores well below average.

The AttrakDiff2 questionnaire's adjective pairs are divided into four different constructs signifying different aspects of user experience. First, we check the internal consistency of each of these constructs separately to ascertain their reliability, and that of the scale as a whole. We obtain the results shown in Table 5.2. We can see that most of the constructs, as well as the scale as a whole, have strong internal consistency, with Cronbach's α values well above 0.7. We do see that the internal consistency of the *pragmatic quality* construct is not good, as well as the *hedonic quality - stimulation* construct for the text feedback version. In the case of the *hedonic quality - stimulation* construct, we can obtain an acceptable internal consistency by removing the "undemanding - challenging" adjective pair, resulting in Cronbach's α values of 0.713 (text version) and 0.855 (ECA version). The *pragmatic quality* construct is not so easily adjusted, there is no subset of adjective pairs that results in a Cronbach's α value of over 0.7 for both the ECA and text versions.

We now evaluate the scores for the scale as a whole. We established that the entire scale is reliable as a whole. Taking the average of all adjective pairs over the entire scale, we find a mean score of 4.64 for the ECA version and 4.66 for the text version. Unsurprisingly, a paired-samples t-test reveals that this difference is not significant (as seen in the first entry of Table 5.3). Therefore, we can not say that there is a significant difference between the two versions on the aspect of user experience as a whole. However, we will look more closely at the results in order to see if there are significant differences on any part of the scale.

As the next step, the scores for each of the separate constructs are assessed. For this test, only the reliable constructs are used. That means the adjusted version of *hedonic quality - stimulation* is used, and *pragmatic quality* is omitted completely since there is no way to make it reliable. We obtain the score for each construct by taking the average score of the adjective pairs belonging to that construct. Next, a paired-samples t-test is performed to evaluate the significance of any differences. The results of this test can be seen in the second section of Table 5.3. Reading from the last column (Sig. 2-tailed), we find that none of the constructs show significant ($Sig. < 0.05$) differences between the ECA and text feedback versions of the system.

Finally, we examine the individual adjective pairs. The graph in Figure 5.1 already showed that differences do exists for some of the individual adjective pairs. In order to find out if any of these differences are significant, we again use a paired-sample t-test. The results of this test can be found in the bottom part of Table 5.3. Again focusing on the last column of the table, we immediately see that most values are well above 0.05, meaning most differences are not significant. We find several values that are marginally significant ($0.1 > Sig. > 0.05$): "rejecting - inviting" ($Sig. = 0.088$), "cumbersome - straightforward" ($Sig. = 0.086$) and "cheap - premium" ($Sig. = 0.054$). In each case, it is the text version that received the higher (more positive) score. Only a single adjective pair shows a significant difference between the ECA and text versions: "complicated - simple" ($Sig. = 0.035 < 0.05$). It is also the text feedback version that scores significantly higher on this adjective pair, indicating that participants found the ECA version to be significantly less simple (and thus more complicated) than the text version. This

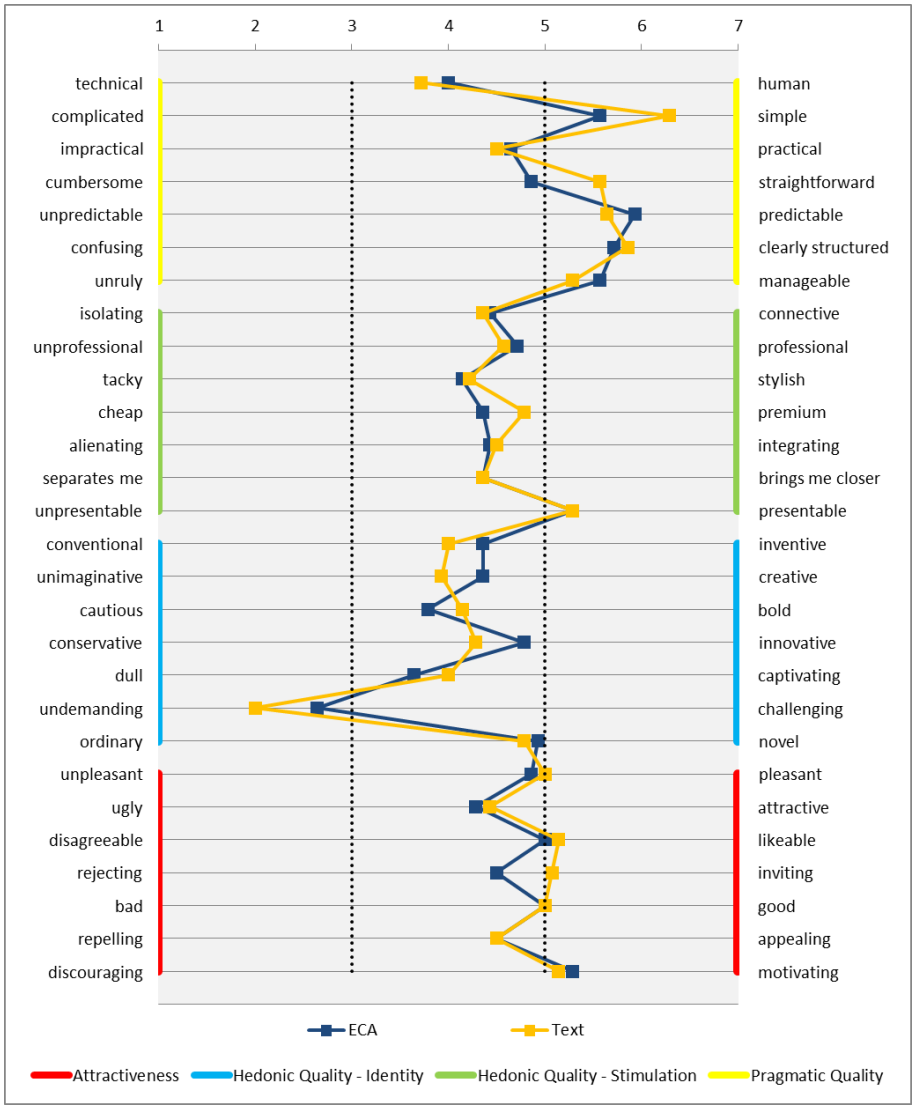

Figure 5.1: Mean values found for AttrakDiff2 adjective pairs

being said, the ECA version still received a mean score of 5.57 on the adjective pair in question, which is above average.

5.3.2 Credibility

The second questionnaire used was the SCS. This was used to evaluate possible differences in credibility between the two feedback versions. Since the SCS is a semantic differential scale consisting of several constructs just like AttrakDiff2, the analysis process is identical. We start with a simple graph representing the mean values for both the text and ECA versions on each adjective pair, as can be seen in Figure 5.2. Again, the two lines are not far apart on most of the adjective pairs, with the biggest difference being visible on the pairs “unpleasant - pleasant” and “awful - nice”.

Again, most of the values appear to be on the high side of average. Upon closer inspection, the only score below 4 is the ECA score on the pair “worthless - valuable”, which is 3.93. What we can also see from the graph is that the text version scores higher on most of the adjective pairs. Only “dishonest - honest” and “selfish - unselfish” show a higher value for the ECA version (the value on “inexpert - expert” is equal for both versions).

The SCS is divided into two constructs of six adjective pairs each. A reliability check on both of

Table 5.2: Reliability values for for each of the used scales and their constructs

Scale	Construct	Condition	Cronbach's α
AttrakDiff2	Pragmatic Quality [1]	ECA	0.649
		Text	0.649
	Hedonic Quality - Identity	ECA	0.803
		Text	0.788
	Hedonic Quality - Stimulation	ECA	0.815
		Text	0.676
	Hedonic Quality - Stimulation (Adjusted) [2]	ECA	0.855
		Text	0.713
	Attractiveness	ECA	0.854
		Text	0.841
	Entire Scale	ECA	0.930
		Text	0.917
Source Credibility Scale	Authoritativeness	ECA	0.838
		Text	0.848
	Character	ECA	0.795
		Text	0.805
	Entire Scale	ECA	0.901
		Text	0.898
UTAUT	Effort Expectancy [1]	ECA	0.778
		Text	0.588
	Attitude Toward Using Technology	ECA	0.897
		Text	0.716
	Anxiety [1]	ECA	0.460
		Text	0.651
	Behavioral Intention To Use	ECA	0.975
		Text	0.924
	Entire Scale	ECA	0.842
		Text	0.678
	Entire Scale (Adjusted) [3]	ECA	0.875
		Text	0.730
Coaching	Entire Scale	ECA	0.889
		Text	0.825

[1] Scale that can not be made reliable (for both ECA and text versions).
[2] Adjective pair "undemanding - challenging" removed.
[3] Statement 6 removed.

these constructs and the scale as a whole reveals the results shown in Table 5.2. With all values well above 0.7, we can see that both constructs have good reliability, and the scale as a whole does as well. We continue by evaluating the results for the scale in its entirety, which leads to a mean score of 4.51 for the ECA version, and 4.83 for the text version. A paired-samples t-test reveals that this is in fact a significant difference ($Sig. = 0.007 < 0.05$), as can be seen at the very top of Table 5.4. Therefore we can say that the text version was scored as significantly more credible by participants.

In order to find out more about this difference, we take a more in-depth look by considering the constructs separately. Since both constructs are reliable, we run a paired-samples t-test on the average scores for each in order to find significance. The results are shown in Table 5.4 (in the second section). We see that while the *character* scale displays a marginally significant advantage in favor of the text version ($0.1 > Sig. = 0.070 > 0.05$), neither construct shows a fully significant difference on its own. Thus we can not state that either version of the system scores significantly higher on *authoritativeness* or *character*.

Finally, we take a look at the individual adjective pairs. We saw in Figure 5.2 that two of the adjective pairs show larger differences than the others. We now perform another paired-samples t-test to ascertain

Table 5.3: Paired-sample t-test results for AttrakDiff2

	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Entire Scale	-,01531	,35569	-,161	13	,875
Hedonic Quality - Identity	-,05102	,44689	-,427	13	,676
Hedonic Quality - Stimulation (Adjusted)	,11905	,62508	,713	13	,489
Attractiveness	-,12245	,57104	-,802	13	,437
technical - human	,286	1,383	,773	13	,453
complicated - simple	**-,714**	**1,139**	**-2,347**	**13**	**,035**
impractical - practical	,143	1,834	,291	13	,775
cumbersome - straightforward	-,714	1,437	-1,859	13	,086
unpredictable - predictable	,286	1,541	,694	13	,500
confusing - clearly structured	-,143	1,231	-,434	13	,671
unruly - manageable	,286	1,858	,576	13	,575
isolating - connective	,071	1,207	,221	13	,828
unprofessional - professional	,143	1,099	,486	13	,635
tacky - stylish	-,071	1,141	-,234	13	,818
cheap - premium	-,429	,756	-2,121	13	,054
alienating - integrating	-,071	,829	-,322	13	,752
separates me - brings me closer	0,000	,961	0,000	13	1,000
unpresentable - presentable	0,000	1,177	0,000	13	1,000
conventional - inventive	,357	1,646	,812	13	,431
unimaginative - creative	,429	1,089	1,472	13	,165
cautious - bold	-,357	,929	-1,439	13	,174
conservative - innovative	,500	1,506	1,242	13	,236
dull - captivating	-,357	1,737	-,769	13	,455
undemanding - challenging	,643	1,985	1,212	13	,247
ordinary - novel	,143	,770	,694	13	,500
unpleasant - pleasant	-,143	,949	-,563	13	,583
ugly - attractive	-,143	1,167	-,458	13	,655
disagreeable - likeable	-,143	1,231	-,434	13	,671
rejecting - inviting	-,571	1,158	-1,847	13	,088
bad - good	0,000	1,109	0,000	13	1,000
repelling - appealing	0,000	1,109	0,000	13	1,000
discouraging - motivating	,143	1,099	,486	13	,635

if these (or any other) differences are statistically significant. The results of this test are shown in the final section of Table 5.4. We find that the exact two adjective pairs identified earlier as showing the biggest difference ("unpleasant - pleasant" and "awful - nice") show significant results ($Sig. = 0.040$ and $Sig. = 0.047$ respectively). This leads to the conclusion that participants found the ECA version to be significantly less nice and significantly less pleasant than the text version.

5.3.3 Acceptance

The next questionnaire was UTAUT. This questionnaire was used to measure the acceptance of the system as an indication for intention of long-term use. The part of the questionnaire we used consists of 14 statements spanning 4 separate constructs. While this questionnaire uses statements instead of adjective pairs, the analysis is still very similar to that of the previous two questionnaires as it also uses a seven-point scale. One important note: all the answer values for the *anxiety* construct have been reversed. This is done because the statements in this construct measure the amount of anxiety experienced by participants, and it should be clear that less anxiety implies more acceptance. In short: after reversing, higher scores are better, as with the other constructs.

We again start with a simple graph to represent the mean values of the answers given on each

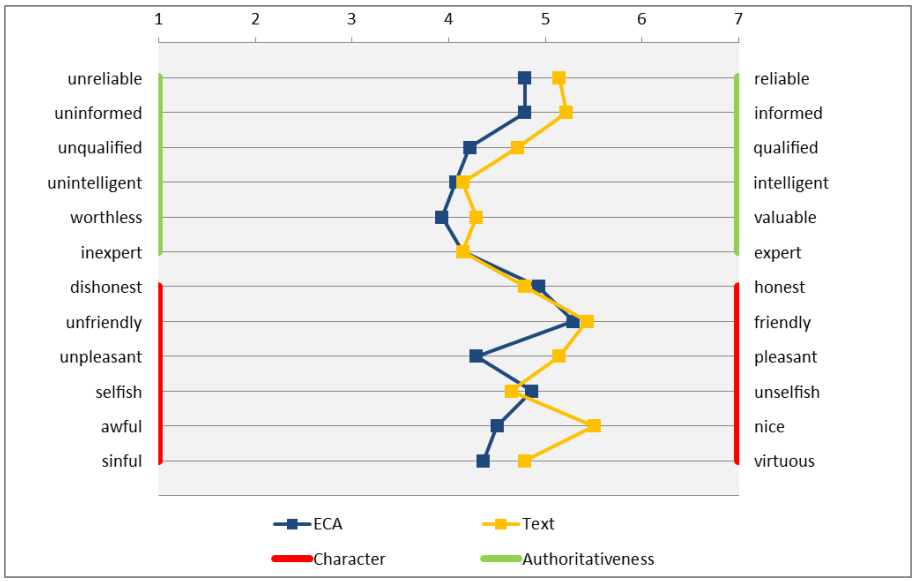

Figure 5.2: Mean values found for Source Credibility Scale adjective pairs

Table 5.4: Paired-sample t-test results for the Source Credibility Scale

	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Entire Scale	**-,31548**	**,36861**	**-3,202**	**13**	**,007**
Authoritativeness	-,28571	,65185	-1,640	13	,125
Character	-,34524	,65524	-1,971	13	,070
unreliable - reliable	-,357	1,598	-,836	13	,418
uninformed - informed	-,429	1,399	-1,147	13	,272
unqualified - qualified	-,500	1,225	-1,528	13	,151
unintelligent - intelligent	-,071	1,328	-,201	13	,844
worthless - valuable	-,357	1,737	-,769	13	,455
inexpert - expert	0,000	1,301	0,000	13	1,000
dishonest - honest	,143	1,406	,380	13	,710
unfriendly - friendly	-,143	1,099	-,486	13	,635
unpleasant - pleasant	**-,857**	**1,406**	**-2,280**	**13**	**,040**
selfish - unselfish	,214	,893	,898	13	,385
awful - nice	**-1,000**	**1,710**	**-2,188**	**13**	**,047**
sinful - virtuous	-,429	1,555	-1,031	13	,321

statement for both of the feedback versions, which can be found in Figure 5.3. What jumps out from this graph is that by far the biggest difference between the ECA and text versions can be seen on statement 11, and that statements 29 and 30 (i.e. the *behavioral intention* construct) score lower than most other statements.

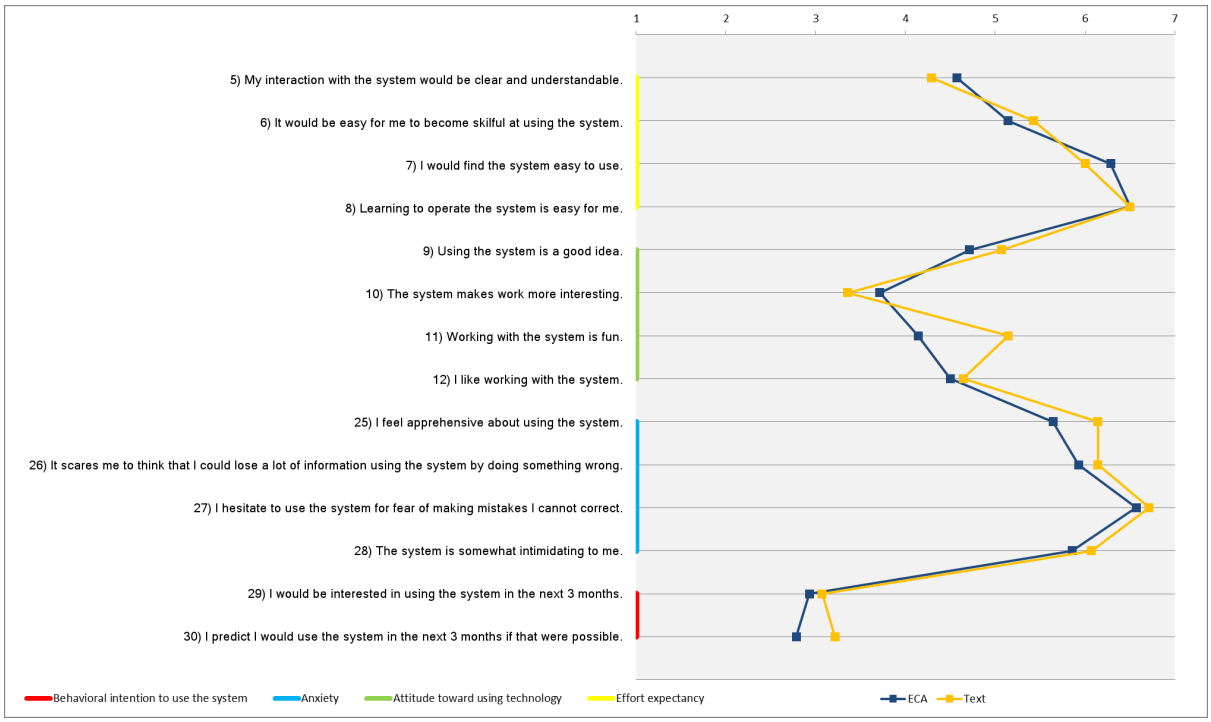

Figure 5.3: Mean values found for UTAUT statements

The results of the reliability analysis can again be found in Table 5.2. The constructs of *effort expectancy* and *anxiety* are not reliable for both ECA and text versions, and also can not be adjusted in such a way that they become reliable for both. The scale as a whole is also not reliable for both version as it is, but can be made reliable by removing the values for statement 6.

Running a paired-samples t-test for the adjusted total scale, the reliable constructs, and the individual statements, we find the results shown in Table 5.5. Unfortunately, none of the differences prove to be statistically significant. The only difference that is marginally significant is on statement 11 ("Working with the system is fun"), with $Sig. = 0.063$.

5.3.4 Coaching

The last questionnaire that was present in both main surveys was the quality of coaching questionnaire. This was used to assess the quality of the coaching delivered by the system. Like UTAUT, it consists of statements with which the participant indicates their level of agreement. Unlike with UTAUT, we used a five-point scale on this questionnaire. Because the final coaching questionnaire used is a modified version of a questionnaire which is in itself already a modified version of CBS-S, it is no longer divided into constructs.

Looking at the initial graph, seen in Figure 5.4, the two lines are fairly close to each other again. While we see some differences, none appear to be very large. Something that does stand out is the fact that scores on most statements are below or only slightly above 3, while in the other questionnaires the scores were predominantly above the median of the answer value range.

After looking at the initial graph, a reliability analysis was performed. Since there are no separate constructs, only the scale as a whole was tested. This proved to have good reliability, with Cronbach's α values above 0.8 for both ECA and text versions. Subsequently, a paired-samples t-test was performed. Results can be found in Table 5.6. Unfortunately, we again find no differences that are statistically

Table 5.5: Paired-sample t-test results for UTAUT

	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Entire Scale (Adjusted)	-,17033	,79723	-,799	13	,438
Attitude	-,28571	,98477	-1,086	13	,297
Behavioral Intention	-,28571	,95503	-1,119	13	,283
Statement 5	,286	1,899	,563	13	,583
Statement 6	-,286	1,383	-,773	13	,453
Statement 7	,286	1,139	,939	13	,365
Statement 8	0,000	,555	0,000	13	1,000
Statement 9	-,357	1,393	-,960	13	,355
Statement 10	,357	1,277	1,046	13	,315
Statement 11	-1,000	1,840	-2,034	13	,063
Statement 12	-,143	1,512	-,354	13	,729
Statement 25	-,500	1,871	-1,000	13	,336
Statement 26	-,214	1,929	-,416	13	,684
Statement 27	-,143	1,292	-,414	13	,686
Statement 28	-,214	2,045	-,392	13	,701
Statement 29	-,143	,864	-,618	13	,547
Statement 30	-,429	1,284	-1,249	13	,234

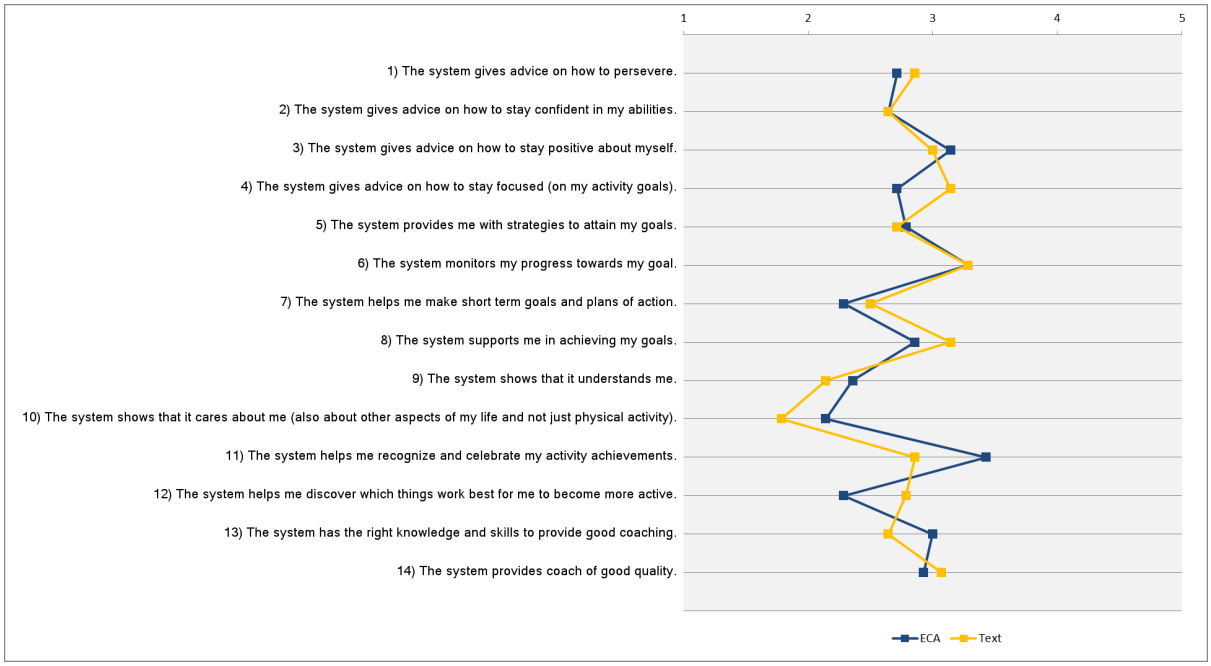

Figure 5.4: Mean values found for Quality of Coaching statements

significant. Statement 11 ("The system helps me recognize and celebrate my activity achievements.") is marginally significant ($Sig. = 0.055$). While this is not quite statistically significant, it is worth noting that it is actually the ECA version that received the higher scores on this item, while all previous significant differences showed an advantage for the text version.

Table 5.6: Paired-sample t-test results for Quality of Coaching

	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
Entire Scale	,00000	,56867	,000	13	1,000
Statement 1	-,143	1,406	-,380	13	,710
Statement 2	0,000	1,664	0,000	13	1,000
Statement 3	,143	1,562	,342	13	,738
Statement 4	-,429	1,604	-1,000	13	,336
Statement 5	,071	,917	,291	13	,775
Statement 6	0,000	1,414	0,000	13	1,000
Statement 7	-,214	,893	-,898	13	,385
Statement 8	-,286	1,267	-,844	13	,414
Statement 9	,214	1,251	,641	13	,533
Statement 10	,357	1,082	1,235	13	,239
Statement 11	,571	1,016	2,104	13	,055
Statement 12	-,500	1,286	-1,455	13	,169
Statement 13	,357	1,277	1,046	13	,315
Statement 14	-,143	1,351	-,396	13	,699

5.3.5 Explicit Comparison

The last questionnaire, which was only present in the final survey, asked participants to make a direct comparison between the two feedback versions. Participants were asked "Please indicate which of the two versions of the system you..." and were then presented with a list of fourteen statements and had to choose which feedback version the statement applied more strongly to. This was done on a five-point scale, allowing participants to indicate a strong or weak inclination towards either version, or a neutral opinion.

Presented in Figure 5.5 is a graphical representation of the mean values found on each of the statements, where values closer to one indicate stronger association with the text version, and values closer to 5 represent stronger association with the ECA version. What we can immediately tell from this graph is that the text version scores higher on all of the positive attributes (one through ten), whereas the ECA version scores higher on every negative statement (eleven through fourteen).

Since the statements do not make up a specific scale and are not divided into separate constructs, there is no reason to look at the "scale" as a whole. This is also the reason that the values for properties that convey negative qualities are not reversed. Instead they are left unchanged for clarity. Because we are not dealing with a real scale, no reliability analysis was performed. Instead, the only statistical test used was a one-sample t-test. This was done to determine if any of the mean values found was significantly above or below the neutral value, and thus if any of the tendencies towards the ECA or text versions were significantly above the neutral value (3).

The results of the one-sample t-test can be found in Table 5.7. We can see a number of significant results. A significantly stronger association with the text version can be found on statements 3, 6 and 7. This means participants were of the opinion that the text version was more pleasant to use, that it gave better advice, and that they would prefer it for long-term use. Significantly stronger associations with the ECA versions are visible on statements 11, 12 and 14. This indicates that participants felt that the ECA version was more irritating, was more cumbersome in its use, and that they ignored more of its feedback messages (something which is further investigated in Section 5.4.2).

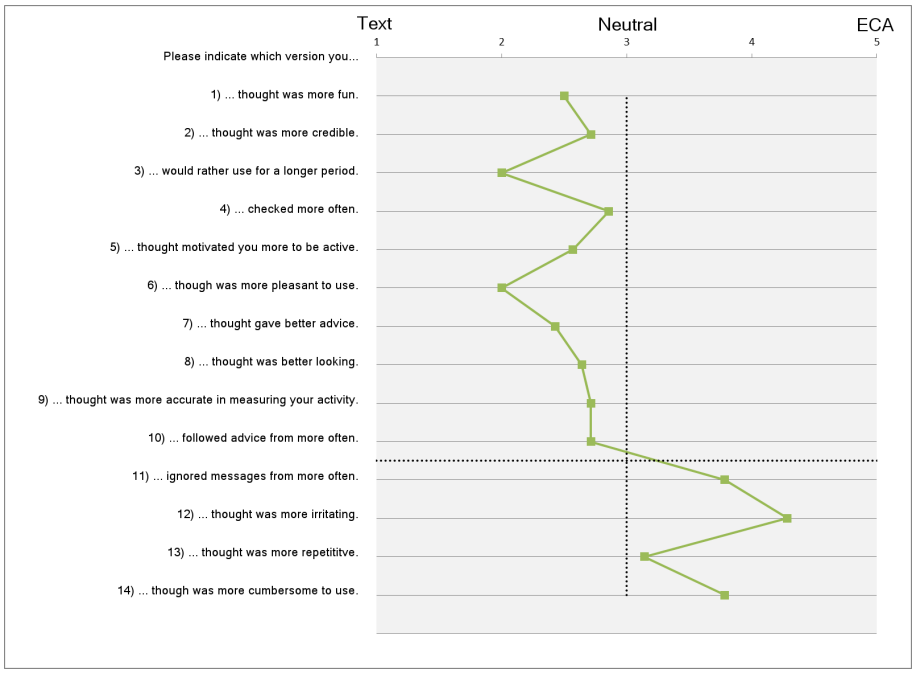

Figure 5.5: Mean values found for explicit comparison statements

Table 5.7: One-sample t-test results for explicit comparison

	Test Value = 3			
	t	df	Sig. (2-tailed)	Mean Difference
Statement 1	-1,242	13	,236	-,500
Statement 2	-1,472	13	,165	-,286
Statement 3	**-2,646**	**13**	**,020**	**-1,000**
Statement 4	-,414	13	,686	-,143
Statement 5	-1,472	13	,165	-,429
Statement 6	**-2,646**	**13**	**,020**	**-1,000**
Statement 7	**-2,511**	**13**	**,026**	**-,571**
Statement 8	-,960	13	,355	-,357
Statement 9	-1,749	13	,104	-,286
Statement 10	-1,295	13	,218	-,286
Statement 11	**2,797**	**13**	**,015**	**,786**
Statement 12	**5,264**	**13**	**,000**	**1,286**
Statement 13	,694	13	,500	,143
Statement 14	**3,294**	**13**	**,006**	**,786**

5.4 Software Data

The final source of data to analyze is in the logs compiled by the software. This includes the measured activity levels, but there is also information about the opening of feedback messages, which we will also analyze and discuss. For different reasons such as sickness, vacation days, practical issues with the sensor node and technical problems, not all participants were able to use the system all day every day during the testing period. Because of this, there are some holes in the data from this source. Wherever necessary and possible, we will discuss and try to correct for these issues.

In the analysis of this data, some general observations are presented, but the main focus remains on the difference between the ECA and text feedback versions.

5.4.1 Overall Activity

We start by looking at the collected activity data. Obviously, this is where holes in the data set pose the biggest problem: if a participant has not used the system for a certain period of time, we do not know the participant's activity level during that time. In order to cope with this problem, we have devised two different adjusted measures, which we will call Total Corrected Activity (TCA) and Qualifying Workday Activity (QWA).

The idea behind the TCA is to use all activity data collected (including non-working days), and to "fill in" any holes in the data caused by the system not being switched on. In order to calculate the TCA for one day from one participant we first determined a completion percentage for the day by evaluating for what period of time the system collected data for that day compared to the time of a full testing day (from 07:00 to 23:00). If this percentage is less than 100%, we multiplied the remainder by the reference level for a day and added this to the measured activity. We used this method for replacing missing data because this is also how the system handles any period during which no data is collected (see also Section 5.2.3). This measure can be summarized in the following formula: $(TCA = activity + (1 - completion) * reference)$.

The second measure, QWA, focuses on the activity levels that were recorded during working days with acceptable completion percentages. This means that instead of filling holes in the data, we discard any data that is incomplete or collected during non-working days. This obviously means that there will actually be more holes than in the original data set. This does not pose a real problem because the values we use to compare the two feedback versions are the participants' weekly averages, which can still be calculated even if there is no data for some days. In order to determine QWA values, a day is discarded if either the completion percentage is below 80%, or if the amount of hours worked is below four. One participant had such low completion percentages that all but one of this participant's daily values were discarded. Because no average can be established for the week with no qualifying values, the data for this participant was discarded in this method of analysis.

We start by looking at the overall average values for each day for both TCA and QWA, which are graphically displayed in Figure 5.6. What we can see immediately is that almost all of the average daily levels are above the reference level, with the QWA on day 13 being the only exception. This confirms what most participants already reported in the interviews: the reference level was easy to surpass, and thus possibly too low. Also visible in this graph is the fact that the two measures are quite similar. The biggest difference is found on day 11, which may be partially explained by the fact that on that day there are only two values that qualify for QWA. There is no real clear visual indication of a drop in activity levels over time (the Pearson product-moment correlation coefficient also did not reveal a statistically significant relation between time and activity level). We also do not see any obvious differences between days 7 and 8, which would indicate an effect of the feedback version switch.

Next, we make the comparison between the two feedback versions. Figure 5.7 shows the levels for both versions for the TCA and QWA measures. From these graphs, we can tell that the ECA version activity is generally slightly lower than the text version activity, although not by much. The overall average values also show this: TCA shows an average of 418 for the ECA version and 438 for the text version, QWA results in 393 for ECA and 433 for text. Furthermore, when looking at participants individually, we find that using the TCA measure, 9 out of 14 participants recorded more activity while using the text feedback version. Using the QWA measure, the same thing goes for 12 out of 13 (qualifying) participants. This difference between the feedback versions is explored further by performing a paired-samples t-test on the average per version of each participant. The difference does not prove to be significant for either

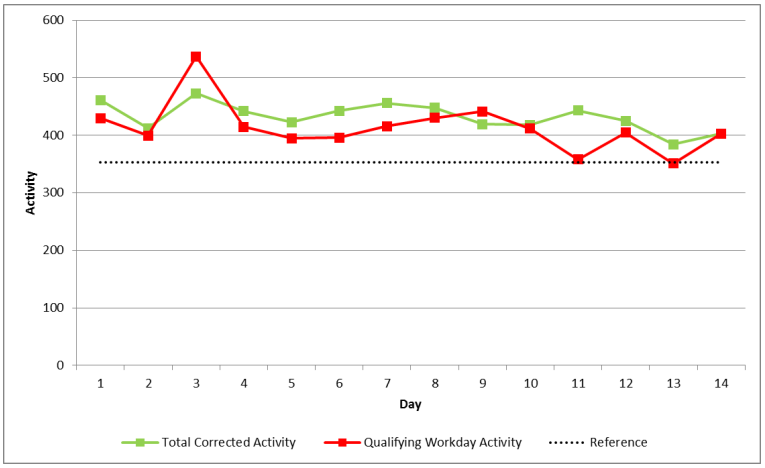

Figure 5.6: Average daily values for both activity measures

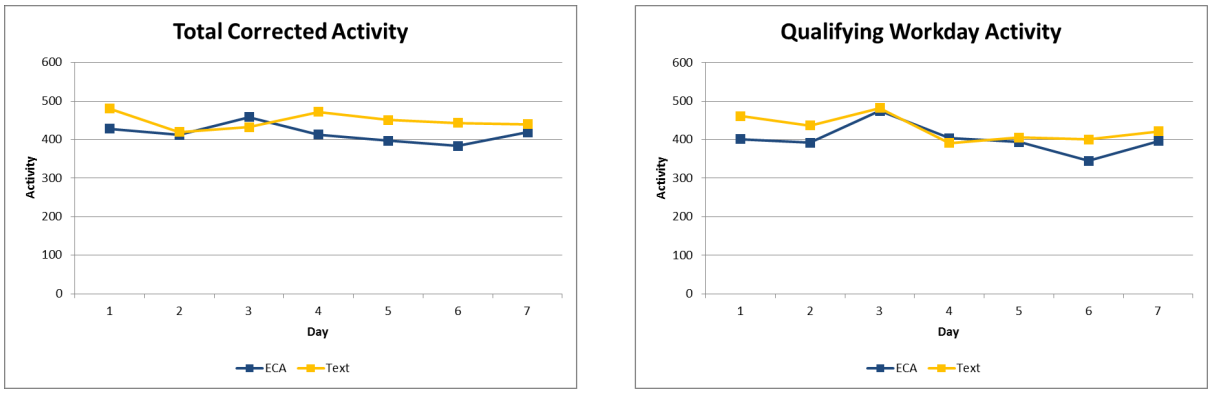

Figure 5.7: Comparison between ECA and text for both activity measures

version ($Sig. = 0.411$ for TCA, $Sig. = 0.091$ for QWA). All in all, there is some indication that the ECA version resulted in less activity than the text version, but there is no statistically significant evidence.

5.4.2 Feedback Messages Seen/Ignored

From the log files created by the smartphone application, we can determine when feedback messages were generated, and if and when they were subsequently closed by the participant (meaning the "OK" button was pressed). While this is not necessarily entirely accurate, we interpret the closing of the feedback screen as the participant having seen the feedback message. Technically, it is of course possible to view the message but not close the screen, as well as close the screen without having actually looked at the message. However, we do not assume either scenario to be particularly likely because of the way in which the application is designed.

Because not all participants have managed to use the system for the entire day every day of the testing period, the amount of feedback messages generated varies per participant per day (fourteen for a full day of using the system). Because of this, we look at the ratio of generated messages that were seen instead of the amount, since it is impossible to have seen a message that was never generated. There are participants that did not use the system at all on one or more days, resulting in zero message being generated. Since our analysis looks at the averages per day and per feedback version, this does not prove to be a problem as this is only a sporadic occurrence.

In general, the overall viewing rate came to 49.1%. There were large differences between the individual participants' averages, which ranged from 19.5% to 71.6%. Figure 5.8 shows the ratio of feedback

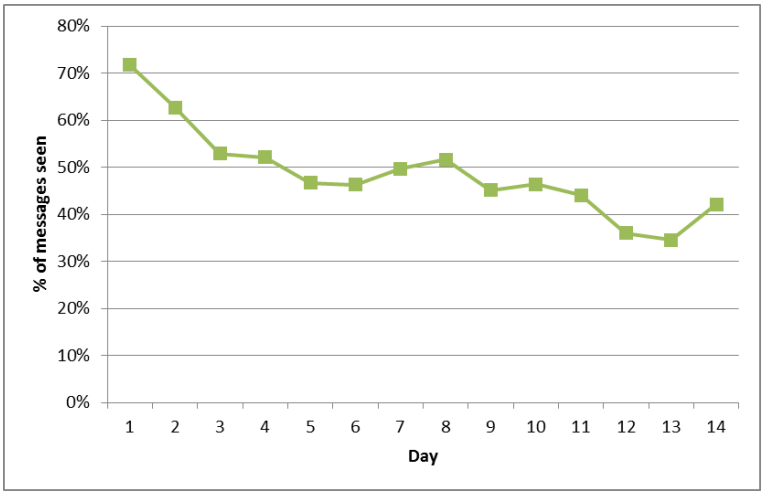

Figure 5.8: Ratio of messages seen for each day of the testing period

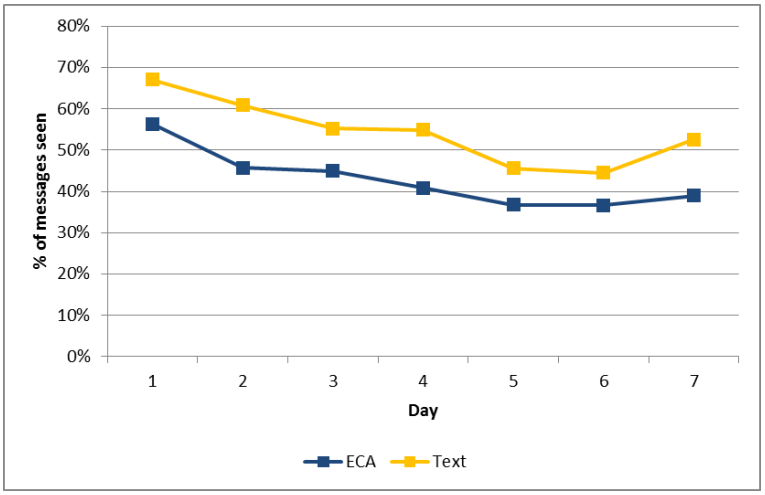

Figure 5.9: Comparison of feedback versions on ratio of messages seen

messages seen over the course of the testing period. We can clearly see that the percentage starts off quite high but quickly drops. The Pearson product-moment correlation coefficient also shows a moderate, negative correlation between the percentage of messages opened and the day of the testing period ($r = -0.334, n = 182, p < 0.0005$). This could be explained by the fact that participants reported becoming less interested in the feedback as they realized that their activity level was generally above the reference level, as well as the fact that the novelty value of a system tends to wear off quickly. While the increase between day 7 and day 8 is not very large, it could be an indication that participants' interest in the system was reawakened somewhat by the introduction of a new feedback version. There is also a rather strong increase visible between days 13 and 14. A possible explanation for this is that the knowledge of being on the last day of the testing period (strengthened by the reminder email and final survey) motivated participants to pay a little more attention to the system for one last day.

When comparing the two feedback versions (see Figure 5.9), several indicators can be found for the text version receiving a higher viewing rate than the ECA version. Out of all the participants, only three had a higher viewing rate in week 2 than they did in week 1. All three of these participants were using the text feedback in week 2. In total, ten out of fourteen participants had a higher viewing rate on the text version than on the ECA version. The numbers per version tell the same story: 43.5% of ECA messages were opened compared to 54.4% of text messages. Finally, a paired-samples t-test on the participants'

daily ratios for both conditions reveals that this difference is in fact significant ($Sig. = 0.018 < 0.05$).

5.4.3 Message Viewing Delay

The final measure we intended to look at was the amount of time that passed between a feedback message being generated and that message being seen by the participant. However, the values obtained revealed a flaw in the idea behind this measure. After processing the log data, an overall average viewing delay (not counting messages that were never viewed) of over 19 minutes was found, which appears shockingly high. This revealed the problem that in many cases, participants simply ignored or missed the notification at the moment the feedback message was generated, and then decided to check the system on their own initiative at a later point in time. We can no longer reasonably consider the elapsed time in such situations to be a delay, since it is debatable whether or not the viewing of the message can still be considered a reaction to the notification when it occurs significantly later. Also, the measure offers us no real information: if a message is checked significantly later than its notification, the actual amount of delay is highly unlikely to be influenced by any aspect of the system, as it is much more likely to depend on when the participant has a free moment to check the system.

This exercise did lead to a revised measure: the ratio of messages actually checked in direct reaction to a notification. Of course, we can not tell the cause of participants' from the software logs. We can however determine the ratio of messages that was seen within a certain (small) amount of time from the notification being played. Using a value of 5 minutes for the time limit, we can perform an analysis analogous to the one for the viewing ratios.

Overall, the rate of immediate viewing was 21.9%. The viewing ratio for each day of the testing period can be seen in Figure 5.10. While less clear than in the overall viewing graph, the ratio does appear to decrease over time as well. The Pearson product-moment correlation coefficient shows a small, negative correlation between the percentage of messages opened immediately and the day of the testing period ($r = -0.334, n = 182, p = 0.002$). A comparison of the feedback versions can be found in Figure 5.11. Of the fourteen participants, twelve had a higher immediate viewing ratio using the text version than they did using the ECA version. A similar difference also exists in the overall numbers: 17.3% of all ECA messages were viewed immediately, compared to 26.1% of all text messages. Finally, we performed another paired-samples t-test on the participants' daily ratios, again indicating a significant difference ($Sig. = 0.020 < 0.05$).

While it is unsurprising to see that these results closely resemble those found in the overall viewing ratios, it is nevertheless valuable to observe that the majority of messages that were seen were not looked at immediately, and thus most likely not in direct response to a notification being played.

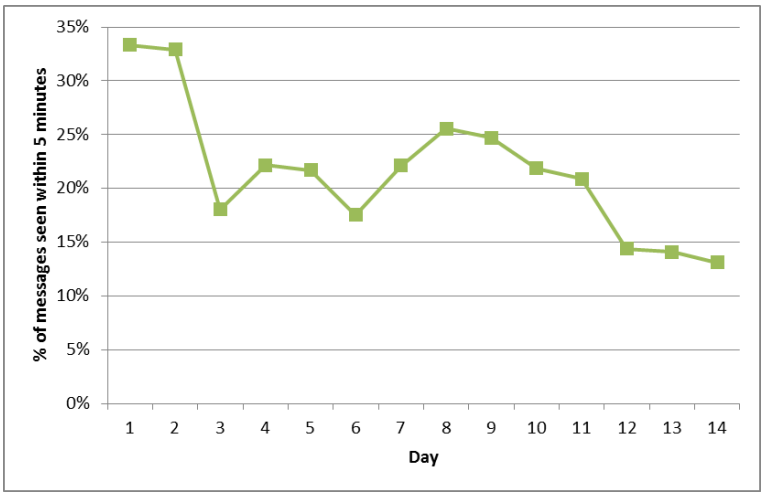

Figure 5.10: Ratio of messages seen within 5 minutes of notification for each day of the testing period

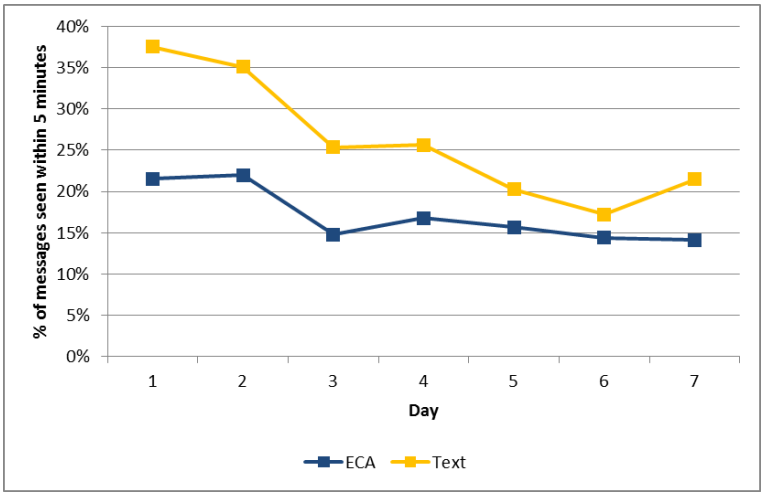

Figure 5.11: Comparison of feedback versions on ratio of messages seen within 5 minutes

Chapter 6

Conclusions and Discussion

In this chapter, we present the conclusions derived from the user experiment performed. We start by discussing the original research questions, but also present relevant conclusions that were found in other areas. Afterwards, we relate these conclusions and the execution of the experimental process to the originally presented theoretical framework, as well as reflecting on the chosen methodology and summarizing the most significant practical issues encountered. Next, suggestions and recommendations regarding future work in this area are presented. We then end this text with a summary of the conclusions that were drawn.

6.1 Answers to Research Questions

We start by looking back at the research questions originally posed in Chapter 1.2, and attempting to provide answers to these questions based on the results discussed in Chapter 5. For clarity, we repeat the original research questions here. The main question was:

Does an ECA offer a valuable addition to mobile health coaching systems?

And we posed the following subquestions:

Does the addition of an ECA to a mobile health coaching application...

1. lead to an increase in users' physical activity levels?
2. lead to an increase in users' evaluations of the user experience?
3. lead to an increase in users' evaluations of the quality of coaching delivered by the system?
4. lead to users continuing to use the system for longer?
5. lead to an increase in users' perceived credibility of the system?

In the remainder of this section, we treat each of these subquestions individually, and end with a discussion of the main question.

6.1.1 Activity

In answer to subquestion 1, we can not say that the addition of an ECA to our health coaching application resulted in an increase in users' physical activity levels.

The interview process revealed only two participants who felt that using the system had made them more physically active in general, and that was mainly because the system caused them to be more aware of their physical activity, not because of coaching delivered by the system. Also, none of the participants reported a difference in activity levels between the two feedback versions.

The collected activity data also revealed no significant difference in activity between the two feedback versions in either of the activity measures used. Overall, the ECA version even showed less activity than the text version in both measures, but this difference was not statistically significant in either.

6.1.2 User Experience

In answer to subquestion 2, we can not say that the addition of an ECA to our health coaching application resulted in an increase in users' evaluations of the user experience.

The survey results showed no significant difference in overall user experience, nor in any of the AttrakDiff2 constructs that represent the different aspects of user experience. In fact, the only adjective pair that showed a significant difference between ECA and text versions revealed that participants found the ECA version to be more complicated (and thus less simple) than the text version.

The direct comparison questionnaire also showed that participants indicated a significantly stronger association with the text version on the aspects of fun and being pleasant to use, and a significantly stronger association with the ECA version on being cumbersome and irritating. While this questionnaire is not a verified tool for measuring user experience, these results do strengthen the indication that participants preferred the user experience of the text version to that of the ECA version.

6.1.3 Quality of Coaching

In answer to subquestion 3, we can not say that the addition of an ECA to our health coaching application resulted in an increase in users' evaluations of the quality of coaching delivered by the system.

The survey process revealed no significant differences between the two feedback versions in regards to the quality of coaching. There was some indication that the ECA version "helped recognize and celebrate activity goals" more than the text version, but this difference turned out be only marginally statistically significant.

Also related to the quality of coaching was the fact that the reference activity level proved to be too low for most participants, something that became obvious during the interview process, and that was confirmed by analysis of the activity data. This resulted in a very large portion of the coaching delivered to participants consisting of only compliments, and thus very few instances of the system providing participants with activity suggestions.

Also partially caused by the low reference level, the interview process revealed that participants had often ignored notifications and messages delivered by the system. The direct comparison questionnaire also showed that participants had a significant inclination towards the ECA version when asked which version they had ignored messages from more frequently. Analysis of the software's log data confirmed that both notifications and messages had been ignored more frequently during the use of the ECA version than during the use of the text version.

6.1.4 Duration of Use

In answer to subquestion 4, we can not say that the addition of an ECA to our health coaching application led to users intending to continue using the system for longer.

The UTAUT questionnaire did not show any significant results in regard to the difference between the two feedback versions, and neither did any of its constituent constructs or statements. During the interview process, most participants indicated a preference for the text version when asked about the direction they felt further development of the system should take. The explicit comparison questionnaire showed a significantly stronger association with the text version on which version participants "would rather use for a longer period of time".

6.1.5 Credibility

In answer to subquestion 5, we can not say that the addition of an ECA to our health coaching application led to an increase in users' perceived credibility of the system. In fact, we can say that the addition of an ECA to our health coaching application led to a decrease in users' perceived credibility of the system.

The interview process showed mixed opinions on the subject of credibility and believability. The survey process however revealed a significant difference on the overall scores of the SCS, in favor of the text version. Looking at the individual adjective pairs, participants judged the text version to be significantly more nice (and thus less awful) and more pleasant (and thus less unpleasant) than the ECA version. While the direct comparison questionnaire also showed a stronger association with the text version on the item "thought was more credible", this difference was not statistically significant.

6.1.6 Main Question

Looking at the answers to each of the subquestions, as well as the results in general, we can not say that the ECA offered a valuable addition to our mobile health coaching system.

None of the data collected during our experiment shows a significant advantage for the ECA version. A large majority of the data actually pointed to the participants preferring the text version over the ECA version. Nevertheless, there were small indications that an ECA could offer advantages over plain text feedback in a health coaching setting. For example, some of the participants did indicate during the interview process that the ECA allowed them to connect to the system on a more personal level.

6.1.7 Discussion of Results

We have not found evidence allowing us to give an affirmative answer to any of our original research questions. However, we must keep in mind that we performed a single experiment with one specific ECA on one specific health coaching system. Therefore, our results do not necessarily mean that ECAs do not provide the benefits we have looked for. It only tells us that they are not present in the specific situation that was used during our experiment. It also does not mean that nothing was learned from the experiment we performed. Here we discuss possible explanations for the results we found, and present any additional findings that do not directly relate to the original research questions.

It became obvious from the interview process that the reference activity level was set too low for almost every participant, which was later confirmed by the activity data. It should be clear that this was a major interfering factor in the experiment. Because the goal level was low, the system hardly ever attempted to persuade the participants to be more active. This made the system extra predictable, since feedback messages almost always contained compliments. According to the interview process, this predictability led to decreased interest from participants, as well as the system being taken less seriously.

The second important finding was the large influence of *glanceability* (or lack thereof). Most participants pointed to the fact that the ECA version takes more time and effort to view as the main reason for preferring the text version. This leads us to the conclusion that in the situation presented in our experiment, *glanceability* is a major factor to the success of the system. Unfortunately, none of our survey questions targeted the concept of *glanceability*, although participants did indicate a significantly stronger association with the ECA version on the item of irritation in the direct comparison questionnaire, which could well be related.

Next, the *glanceability* of an ECA is inherently less than that of a line of printed text, but why is it that this had such a large influence in our experiment? We believe that the influence of the *glanceability* factor was exacerbated by a combination of two aspects of our experiment: the frequency and predictability of the feedback messages. The fact that the reference level was low caused the contents of feedback messages to become very predictable, since the majority contained compliments. The fact that the collection of possible compliments delivered by the system was relatively limited surely did not help either. Combined with the fact that feedback was generated every hour, this more than likely resulted in participants simply wanting to check the message very quickly (if at all), and get back to what they were doing. Not actually being interested in the message (since its contents were already known) almost certainly led to the ECA becoming far less attractive because it took more time to view the message.

In summary, all this leads us to the following finding: in a system where messages are being delivered frequently and the content of these messages is predictable, the *glanceability* of the delivery method has a strong influence on the users' overall evaluations of the system. Since this was not the original focus of our experiment, we do not have the evidence needed to truly call this a conclusion, but the results of our experiment certainly point firmly in this direction.

6.2 Reflection on Theory

Here we look back at the theoretical background discussed in Chapter 2 and try to relate our findings to the theories discussed there.

6.2.1 Transtheoretical Model of Behavior Change

The most remarkable observation in regards to the *transtheoretical model* is that out of the five stages of change, the only one that was not represented in our group of participants was the *action* stage, which normally should be the main target audience for a health coaching system. While this area was not at the forefront of our experiment, the interview process certainly suggested that it is important to consider the stage of change a user is in. For example, some participants expressed not being that interested in what the system had to say because they already felt comfortable with their own level of physical activity, which is consistent with these participants being in the *maintenance* stage. Also interesting was the suggestion by several participants that the system would be more useful to someone with an intrinsic motivation to be more physically active, which would correspond to being in the *action* stage (or transitioning into it by starting to use the system). Additionally, a suggestion that was raised multiple times was that the system may be useful in a sort of "periodical tune-up"-capacity. These participants suggested that they might be interested in using the system for relatively infrequent and short periods at a time, to see if their physical activity was at the level they expected (and desired), and to make small adjustments if needed. This kind of behavior would certainly fit a user that is in the *maintenance* stage.

As for the processes of change included in the *transtheoretical model*, we can not claim to have truly exploited any, since there was no significant raise in activity to be found. However, the interview process did hint at some of these processes at work. For instance, there were participants that reported a higher awareness of their choice of transportation methods while using the system, for example choosing the bicycle over the car more often. While the system did not actively promote this kind of behavior change, this is an example of *counterconditioning*. In terms of *contingency management*, participants have certainly been provided with a large amount of positive reinforcement. What we may have seen here though is that positive reinforcement loses much of its effectiveness when it is delivered too frequently and/or at times where the user feels it has not been earned.

6.2.2 Persuasive Technology

Next, we look at the concepts of persuasive technology. Looking at the tactics available to persuasive tools, we can assess some of the successes and shortcomings in the ways these tactics were used in our system. In terms of *tailoring*, the most prominent use of this tactic was in the system's main screen. Presenting the participants with a simple graph of their activity level certainly made this information easy to absorb very quickly. However we did not entirely manage to meet the participants' demands in this area, as several indicated a desire for seeing the data expressed in a form they can relate to, such as calories burned or steps taken. Next, *suggestion* clearly should have been the main persuasive tactic used by the system. Unfortunately, due to the low reference level, *suggestion* was not employed nearly as often as we would have liked. Also, not every participant was equally pleased by the suitability of the suggestions in their working environment. Participants were however predominantly pleased with the *self-monitoring* capabilities offered by the system, with many finding it interesting to be able to see their own activity patterns displayed graphically. While there was no (intended) use of *surveillance* in our system, some participants did report that the visible presence of the system occasionally led to social interactions.

Looking at our ECA, we assess the ways in which it utilizes the social cues discussed in the theory. The drawn upper torso and head certainly presents a *physical presence* to the user, although it is clearly a drawing, and does not truly resemble a "real person". Our ECA does present some *psychological cues* in the facial expressions that accompany messages, but these are quite simple. This was also identified by one of the participants, who indicated that while the ECA did convey some form of personality, it was in the end "as flat as a character can be". The use of TTS to present the feedback messages in the form of spoken utterances was clearly a form of *language* cue. The problem here was that a large portion of the participants disabled the sound because it was not appreciated in the workplace, or in some cases because they simply found it annoying. Because our ECA really only delivers a message to the user and is not capable of more complex interactions, it does not really exhibit any use of *social dynamics*. In terms of *social role*, our ECA was originally designed as a physician. While we had attempted to make this role less obvious by removing the stethoscope, several participants still identified our ECA as a medical professional, although they were divided on whether this was an appropriate role for use in our system.

6.2.3 Ethics

In terms of ethics, we feel that in the feedback delivered there were no ethical violations in regards to the persuasion attempts delivered to the participants. Every participant was informed that they were free to ignore any suggestions they felt were inappropriate or unwanted. In terms of the ECA delivering spoken text through TTS, we do not feel that the participants' privacy was harmed by bystanders possibly overhearing these personal messages because playback was never started unless triggered by the participant, and a mute option was also available. We do believe that an ethical violation may have occurred in the presentation of the reference level. Several participants indicated that they were reassured by the fact that their measured activity exceeded the reference level, and that they felt better about their own level of physical activity after the experiment. Because the reference level was not based on any recognized standard for a healthy amount of physical activity, it is quite possible that the reassurance felt by these participants was in fact not justified, which could certainly be construed as an unethical result. In an attempt to remedy this, we did inform participants of the fact that the reference level did not necessarily represent a healthy level of physical activity. In hindsight however, it would have been prudent to inform participants of this fact beforehand instead of afterwards.

6.2.4 E-Health and ECAs

Now we take a look at the properties and advantages of E-Health systems in general, and how these apply to the system used in our experiment. Since our system is a good example of an E-Health system, most of the benefits associated with E-Health, such as anonymity, apply. One area which still poses an obstacle to aspects such as ease of access and scalability is the requirement to use an external sensor node, which is both expensive and not widely available. While our system is generally very respectful towards the user's privacy, some participants did feel that it was violated when the system used the TTS to utter feedback when other people were within earshot. In terms of the additional tools available to mobile E-Health systems, we have attempted to incorporate some form of *context awareness* by using different sets of possible activity suggestions based on the user being at work or at home. Of course, the "awareness" was based only on the time of day and was thus far from perfect. Also, only a small number of participants had noticed the difference in feedback message content.

The (possible) advantages of ECAs were obviously the focus of our experiment, and our conclusions on this subject were discussed earlier. While most other research in the field of ECAs has struggled to find significant effects on any kind of task performance, our experiment also failed to show the positive effect on aspects such as user experience that other studies did find. Although we have done our best to present an attractive ECA to the users in the limited time available, it is clear that there is still a lot of room for improvement in terms of the behaviors and emotion shown by the ECA.

6.3 Reflection on Methodology

In this section we reflect on the way the experiment was designed and executed, and identify areas for improvement.

6.3.1 Experimental Design

Looking at the final group of participants, we were fairly successful in finding a balanced group. One observation in this area is that because of the restricted search area, many of our participants were employed in research or related areas. While these are generally sedentary professions, they do include fairly frequent out-of-office activities such as meetings, lectures, presentations, etcetera. While such occurrences are not uncommon in most sedentary professions, it would be prudent to obtain a wider sample in terms of occupations if a larger-scale experiment were performed. One important issue with our group of participants was that the stage of change they were in was not considered. Clearly, selecting potential participants based on this criterion would have made it far more difficult to find suitable participants, and there simply was not enough time for this. The size of our group was acceptable even for statistical analysis, although a larger sample would be preferable.

The duration of the experimental period was too low to provide any information about long-term effects, but we believe it was appropriate for our study. We do not believe that simply employing a

longer testing period would have yielded different results. If there had been significantly more time however, something that would probably have had a significant impact on the results is the addition of a calibration week at the start, in order to determine a personalized reference level for each participant. This would have eliminated the problems caused by the low reference level. Given the small scale of the experiment, the within-subjects design was appropriate. Allowing participants to make a conscious comparison between the two feedback versions yielded valuable results.

6.3.2 Data Collected

The way in which the system collected data was certainly relevant and provided for an interesting analysis. The one problem we had in this area was with the way the software assumes that activity continues to increase at the same pace as the reference level for any period over which no data is received. While this is an acceptable solution in case the connection to the sensor node is temporarily unavailable, it is an incorrect assumption in cases where the user is still asleep when the measuring period starts. For some of our participants, this compounded the problem of the reference level being too low.

In terms of the surveys, we feel that the questionnaires used were certainly applicable to the system, even though there were few significant differences found between the two feedback versions. In hindsight, it would have been valuable to include some sort of assessment of *glanceability* and related concepts, but that was a factor that we had not considered beforehand.

The interview process turned out to be a very valuable addition to the other collected data, as it not only provided valuable suggestions and insights about the system that could not have been revealed by a rigid questionnaire, but also provided us with an indication of the importance of *glanceability*, a significant element to the explanation of other results. Looking at the way the interview process was conducted, the analysis could have benefited from the interviews being more structured. At the same time however, we appreciate the fact that the informal nature allowed participants to discuss the things that stood out to them instead of just answering specific questions.

6.3.3 Procedure

In terms of the entire experimental process and the procedure followed with each participant, we are very happy with the way things unfolded. Most participants did not encounter any significant problems during the testing period, nor were there many questions after the introductory meetings. Participants also completed each of the surveys in time. This allowed us to have minimal interaction with the participants during the testing period (in most cases none at all), minimizing the risk of interfering with the outcome of the experiment.

6.4 Recommendations

In this section we provide recommendations for future work in the areas covered by this project. Each of the areas will be discussed separately. We start by identifying some interesting areas for further research studies, and follow this up with suggestions for the further development of both ECAs and activity coaching systems. These suggestions will be geared towards the systems we have used, but should prove valuable for other applications as well.

6.4.1 Further Research

We start by pointing out the importance of finding an appropriate activity goal for participants, assuming that the system being tested uses such a goal. As we have seen, attempts at persuading a user to reach a certain goal are futile when the user is not the least bit challenged by that goal (and thus does not need persuading). It is clearly difficult to find an appropriate goal when time and resources are limited. An interesting experiment would be to investigate whether an activity goal that is set too high (i.e. to a level that is difficult to achieve for most participants) results in more relevant data than one that is set too low.

In our work, we have found indications that *glanceability* is very important in settings where message content is predictable and/or messages are unwanted. While it is obviously better to avoid such situations,

any system will grow predictable over time, and a system can never predict with 100% accuracy when the user will be receptive to feedback. Therefore we believe it would be worthwhile to investigate this effect further, in order to first confirm the effect, and then investigate the best ways to alleviate it.

Taking a broader view, a lot of research into the field of ECAs focuses on finding some sort of general beneficial effect of using an ECA in a user interface. Most studies struggle to show such an effect. We believe it would be wise to focus more on the positive effects that have been found, and on the reasons these effects occur. Instead of trying to find general effects of the ECA as a concept, try to find out which aspects of ECAs cause positive user reactions, and how these aspects can be increased to help improve the ECAs that we have.

One method for testing ECAs that could help in both of these situations is allowing users to make their own choice between an ECA (or even more than one) and standard text interaction. In the end, especially when talking about commercialization, it is the user who decides what is succesfull and what is not. So by observing what ECAs are more popular it would be possible to find the properties that can make ECAs succesfull. Also, when performing this type of experiment over a longer period of time, seeing participants switch to standard feedback over time would be confirmation of the idea that once a system becomes predictable a user simply desires the most efficient interaction possible.

6.4.2 Activity Coaching Systems

During the experiment, we have found several areas of the C3PO system that could benefit from improvements. Due to the fact that our target user group (office workers) was different from the traditional C3PO user group (rehabilitation patients), some of these suggestions may only be interesting if this new user group is targeted by further development.

First of all, the hardware situation must improve. Obviously though, the systems used were intended for development and testing of the software, so most of this will be obvious. The sensor node is rather unwieldy and smaller hardware exists that can deliver the same functionality. The use of a second smartphone is very impractical and, since most people have their own nowadays, unnecessary. This does have implications for the software setup. Currently the smartphone application is designed to run at all times and replace the phone's standard software. Clearly, for it to be usable on a user's own phone, it should be converted to a regular app, which runs in the background and delivers notifications when attention is desired.

The second area we identified is the presentation of data. While the activity graph is popular and seems to serve its purpose well, participants indicated a desire for two things: the activity level expressed in a measure they can relate to, and the ability to look at and compare data for previous days. The first point may prove difficult due to the current unavailability of such information within the system, but it should be possible to either determine some sort of reasonable conversion rate, or even switch to different sensor hardware for which such data is available. In regards to the second point, it should be noted that the existing web platform associated with C3PO is capable of delivering summaries and more extensive overviews of collected data. However, in the interest of ease-of-access to this data, providing some additional views within the smartphone application may be a good idea.

The next opportunity for improvement is in the frequency with which feedback is generated. While a very structured and regular schedule may be preferable for rehabilitation patients, our findings suggest that a high frequency is unnecessary when the user is doing well, and can even lead to disinterest and lowered system credibility. A possible improvement would be to replace the standard schedule with a dynamic one which delivers feedback only occasionally when the user is doing well, and more frequently when the user is not doing so well.

One additional aspect to be considered when targeting office workers instead of rehabilitation patients is the pattern of activity. Where rehabilitation patients benefit from a gradual activity pattern throughout the day, office workers are unlikely to follow such a pattern. Factors such as commutes and sports practice provide not only frequent exceptions to the gradual pattern, but are also usually scheduled, meaning the user already knows that a big activity spike is coming later in the day. Any work in this area would start with deciding what exactly the system is meant to achieve: is reaching a certain daily total the only goal, or is it also important to stay active throughout the day even if the overall amount of activity is more than sufficient?

6.4.3 Mobile ECAs

The field of mobile ECAs clearly has room for improvement in virtually every direction, considering this field is still very much in its infancy. Instead of attempting to list all the things that could be improved upon, we focus on a small number of suggestions that we feel are realistic and relevant to the mobile Elckerlyc system.

By far the most deserving of an upgrade is the TTS system. The technical aspects of the Android TTS facility are utterly insufficient to the requirements posed by the Elckerlyc system. The lack of viseme/phoneme and timing information is unfortunate, but having to synthesize an utterance to a file in order to simply find its length is not only a disaster in the way it works, but also adds a significant delay to the entire process. Furthermore, the TTS does not sound as natural as other PC-based applications, even with a commercial voice pack.

Turning to the ECA itself, there is certainly room for improvements within the constraints of the PictureEngine. Presenting a dynamic and lifelike overall appearance is inherently difficult when working with static images. Replacing more images with animations and adding more behaviors should help in making the ECA more "lively". Additionally, it may be worth exploring the options for developing a mobile 3D embodiment. Smarthphones are increasingly capable of rendering 3D graphics, and while a scene and model like in the PC version would still be impossible, rendering a head and partial torso in real-time with reasonable detail should be an attainable goal.

One more way to improve the ECA and increase its impact would be to make it more interactive. In our experiment the ECA only presented short messages to the user, which is a rather one-sided interaction. In order to make the ECA more human-like, it is probably a good idea to make it more responsive to the user. This could be accomplished using technologies such as voice recognition, gaze tracking, or simply a smartphone's touchscreen.

6.5 Closing Summary

In closing, we revisit the original research question:

Does an ECA offer a valuable addition to mobile health coaching systems?

In the data collected during the user experiment we performed, we have not found any evidence suggesting that an ECA offers a valuable addition to the mobile health coaching system that was used. We did discover that in our situation, on of the main drawbacks of the ECA was the fact that feedback delivered by ECA is less *glanceable* than feedback delivered by plain text. We also noted that it is very important to provide users with a suitable goal level in terms of physical activity in order for any persuasive tactics to be effective. We feel these results can be used in further research and development to provide users with an ECA that is better suited to the task domain.

Bibliography

[1] World Health Organization, "Obesity and overweight." `http://www.who.int/mediacentre/factsheets/fs311/en/index.html`, may 2012. visited: 18-04-2013.

[2] World Health Organization, "Physical inactivity: A global public health problem." `http://www.who.int/dietphysicalactivity/factsheet_inactivity/en/index.html`. visited: 18-04-2013.

[3] C. LaPlante and W. Peng, "A systematic review of e-health interventions for physical activity: an analysis of study design, intervention characteristics, and outcomes.," *Telemedicine and e-Health*, vol. 17, no. 7, pp. 509–523, 2011.

[4] R. Klaassen, H. J. A. op den Akker, and A. Nijholt, "Digital lifestyle coaches on the move," in *Proceedings Thirteenth International Symposium on Social Communication* (L. R. Miyares, M. A. Rosa, S. M. Alvarado, and A. M. Alvarado, eds.), Actualizaciones en Comunicacin Social, (Santiago de Cuba), pp. 338–344, Centro de Linguistica Aplicada, 2013. ISBN=978-959-7174-22-6.

[5] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction*, vol. 12, pp. 293–327, June 2005.

[6] J. K. Hendrix, "Elckerlyc on android: A lightweight embodiment." July 2012.

[7] R. Klaassen, J. Hendrix, D. Reidsma, *et al.*, "Elckerlyc goes mobile: enabling technology for ECAs in mobile applications," in *UBICOMM 2012, The Sixth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, pp. 41–47, 2012.

[8] H. Oinas-Kukkonen, "Behavior change support systems: A research model and agenda," in *Persuasive Technology* (T. Ploug, P. Hasle, and H. Oinas-Kukkonen, eds.), vol. 6137 of *Lecture Notes in Computer Science*, pp. 4–14, Springer Berlin / Heidelberg, 2010.

[9] Great Britain Dept. of Health, Physical Activity, Health Improvement and Prevention, *At least five a week: evidence on the impact of physical activity and its relationship to health: a report from the Chief Medical Officer*. London: Dept. of Health, 2004.

[10] J. O. Prochaska and W. F. Velicer, "The transtheoretical model of health behavior change," *American Journal of Health Promotion*, vol. 12, pp. 38–48, Sept. 1997.

[11] B. Reeves and C. Nass, *The media equation: how people treat computers, television, and new media like real people and places*, ch. 2, pp. 19–36. New York, NY, USA: Cambridge University Press, 1996.

[12] A. Bandura, "Self-efficacy: Toward a unifying theory of behavioral change.," *Psychological Review*, vol. 84, no. 2, pp. 191–215, 1977.

[13] R. Gass and J. Seiter, *Persuasion, social influence, and compliance gaining*. Boston: Allyn and Bacon, 2003.

[14] B. J. Fogg, *Persuasive technology: Using computers to change what we think and do*. Morgan Kaufmann, 1 ed., 2003.

[15] A. Bandura, *Social learning theory*. Englewood Cliffs: Prentice-Hall, 1976.

[16] C. Nass, B. J. Fogg, and Y. Moon, "Can computers be teammates?," *International Journal of Human-Computer Studies*, vol. 45, no. 6, pp. 669–678, 1996.

[17] B. J. Fogg and C. Nass, "How users reciprocate to computers: an experiment that demonstrates behavior change," in *CHI '97 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '97, (New York, NY, USA), pp. 331–332, ACM, 1997.

[18] D. Berdichevsky and E. Neuenschwander, "Toward an ethics of persuasive technology," *Commun. ACM*, vol. 42, pp. 51–58, May 1999.

[19] S. Consolvo, K. Everitt, I. Smith, and J. A. Landay, "Design requirements for technologies that encourage physical activity," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, (New York, NY, USA), pp. 457–466, ACM, 2006.

[20] J. Maitland, S. Sherwood, L. Barkhuus, I. Anderson, M. Hall, B. Brown, M. Chalmers, and H. Muller, "Increasing the awareness of daily activity levels with pervasive computing," in *Pervasive Health Conference and Workshops, 2006*, pp. 1–9, 29 2006-dec. 1 2006.

[21] N. Oliver and F. Flores-Mangas, "MPTrain: a mobile, music and physiology-based personal trainer," in *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, MobileHCI '06, (New York, NY, USA), pp. 21–28, ACM, 2006.

[22] A. Andrew, G. Borriello, and J. Fogarty, "Toward a systematic understanding of suggestion tactics in persuasive technologies," in *Persuasive Technology* (Y. de Kort, W. IJsselsteijn, C. Midden, B. Eggen, and B. Fogg, eds.), vol. 4744 of *Lecture Notes in Computer Science*, pp. 259–270, Springer Berlin / Heidelberg, 2007.

[23] J. Beskov and S. McGlashan, "Olga – a conversational agent with gestures," in *Proceedings of the IJCAI'97 workshop on Animated Interface Agents-Making them Intelligent*, 1997.

[24] K. R. Thórisson, "Gandalf: an embodied humanoid capable of real-time multimodal dialogue with people," in *Proceedings of the first international conference on Autonomous agents*, AGENTS '97, (New York, NY, USA), pp. 536–537, ACM, 1997.

[25] J. Cassell, T. Bickmore, H. Vilhjálmsson, and H. Yan, "More than just a pretty face: affordances of embodiment," in *Proceedings of the 5th international conference on Intelligent user interfaces*, IUI '00, (New York, NY, USA), pp. 52–59, ACM, 2000.

[26] M. W. Kadous and C. Sammut, "Mobile conversational characters," in *Proceedings of the Virtual Conversational Characters: Applications, Methods, and Research Challenges Workshop*, 2002.

[27] L. Chittaro, F. Buttussi, and D. Nadalutti, "MAge-AniM: a system for visual modeling of embodied agent animations and their replay on mobile devices," in *Proceedings of the working conference on Advanced visual interfaces*, AVI '06, (New York, NY, USA), pp. 344–351, ACM, 2006.

[28] B. Tomlinson, M. L. Yau, and E. Baumer, "Embodied mobile agents," in *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, AAMAS '06, (New York, NY, USA), pp. 969–976, ACM, 2006.

[29] T. Bickmore, "Towards the design of multimodal interfaces for handheld conversational characters," in *CHI '02 extended abstracts on Human factors in computing systems*, CHI EA '02, (New York, NY, USA), pp. 788–789, ACM, 2002.

[30] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal, "The persona effect: affective impact of animated pedagogical agents," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '97, (New York, NY, USA), pp. 359–366, ACM, 1997.

[31] D. M. Dehn and S. van Mulken, "The impact of animated interface agents: a review of empirical research," *International Journal of Human-Computer Studies*, vol. 52, no. 1, pp. 1–22, 2000.

[32] O. A. Blanson Henkemans, P. J. van der Boog, J. Lindenberg, C. A. van der Mast, M. A. Neerincx, and B. J. Zwetsloot-Schonk, "An online lifestyle diary with a persuasive computer assistant providing feedback on self-management," *Technology and Health Care*, vol. 17, pp. 253–267, Jan. 2009.

[33] T. Bickmore and D. Mauer, "Modalities for building relationships with handheld computer agents," in *CHI '06 extended abstracts on Human factors in computing systems*, CHI EA '06, (New York, NY, USA), pp. 544–549, ACM, 2006.

[34] F. Buttussi and L. Chittaro, "MOPET: A context-aware and user-adaptive wearable system for fitness training," *Artificial Intelligence in Medicine*, vol. 42, no. 2, pp. 153–163, 2008.

[35] T. W. Bickmore, D. Mauer, and T. Brown, "Context awareness in a handheld exercise agent," *Pervasive and Mobile Computing*, vol. 5, no. 3, pp. 226–235, 2009.

[36] M. Turunen, J. Hakulinen, O. Ståhl, B. Gambäck, P. Hansen, M. C. R. Gancedo, R. S. de la Cámara, C. Smith, D. Charlton, and M. Cavazza, "Multimodal and mobile conversational health and fitness companions," *Computer Speech & Language*, vol. 25, no. 2, pp. 192–209, 2011.

[37] H. Op den Akker, M. Tabak, M. Marin-Perianu, M. H. A. Huis in t Veld, V. M. Jones, D. Hofs, T. M. Tönis, B. W. van Schooten, M. M. R. Vollenbroek-Hutten, and H. J. Hermens, "Development and evaluation of a sensor-based system for remote monitoring and treatment of chronic diseases - the continuous care & coaching platform," in *Proceedings of the sixth international symposium on E-Health Services and Technologies EHST 2012*, pp. 19–27, 2012.

[38] H. van Welbergen, D. Reidsma, Z. Ruttkay, and J. Zwiers, "Elckerlyc - a BML realizer for continuous, multimodal interaction with a virtual human," *Journal on Multimodal User Interfaces*, vol. 3, pp. 271–284, 2009.

[39] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. Thorisson, and H. Vilhjalmsson, "Towards a common framework for multimodal generation: The behavior markup language," in *Intelligent Virtual Agents*, vol. 4133 of *Lecture Notes in Computer Science*, pp. 205–217, Springer Berlin / Heidelberg, 2006.

[40] D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt, "Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents," in *Intelligent Virtual Agents* (H. Prendinger, J. Lester, and M. Ishizuka, eds.), vol. 5208 of *Lecture Notes in Computer Science*, pp. 117–130, Springer Berlin Heidelberg, 2008.

[41] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann, "SmartBody: behavior realization for embodied conversational agents," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems - Volume 1*, AAMAS '08, (Richland, SC), pp. 151–158, International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[42] A. Heloir and M. Kipp, "Real-time animation of interactive agents: Specification and realization," *Applied Artificial Intelligence*, vol. 24, no. 6, pp. 510–529, 2010.

[43] M. Mancini, R. Niewiadomski, E. Bevacqua, and C. Pelachaud, "Greta: a SAIBA compliant ECA system," *Agents Conversationnels Animes*, 2008.

[44] W. Wieringa, "User handover in a cross media device environment," Master's thesis, University of Twente, 2012.

[45] World Health Organization, "Physical activity and adults." `http://www.who.int/dietphysicalactivity/factsheet_adults/en/index.html`. visited: 18-04-2013.

[46] M. Hassenzahl, M. Burmester, and F. Koller, "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität," in *Mensch & Computer 2003: Interaktion in Bewegung* (G. Szwillus and J. Ziegler, eds.), (Stuttgart), pp. 187–196, B. G. Teubner, 2003.

[47] J. C. McCroskey, "Scales for the measurement of ethos," *Speech Monographs*, vol. 33, no. 1, pp. 65–72, 1966.

[48] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. pp. 425–478, 2003.

[49] J. Côté, J. Yardley, J. Hay, W. Sedgwick, and J. Baker, "An exploratory examination of the coaching behavior scale for sport," *Avante Research Note*, vol. 5, no. 2, 1999.

[50] T. L. Matthews, J. Forlizzi, and S. Rohrbach, "Designing glanceable peripheral displays," Tech. Rep. UCB/EECS-2006-113, EECS Department, University of California, Berkeley, Sep 2006.

Appendix A

Feedback Message Listing

Presented below is a listing of all the possible feedback messages the system has to choose from, sorted by category. Because the experiment was performed with Dutch-speaking participants, the messages are only available in Dutch.

Neutral - General

- "Ga zo door."
- "Gaat u vooral zo door."
- "Ga op deze manier verder."
- "Probeer zo door te gaan."
- "Blijf op deze manier doorgaan."
- "Hou dit niveau vol!"
- "Probeer dit niveau vol te houden."
- "Blijf dit niveau volhouden."
- "Zorg dat u dit niveau vol blijft houden."
- "Goed zo, hou dit vol!"
- "Zo gaat het goed."
- "Het gaat goed."
- "U bent goed bezig."
- "Goed bezig, maar blijf actief."

Encouraging - Minor - General

- "Misschien wordt het tijd om even de benen te strekken?"
- "Strekt u even de benen."
- "Probeer even uw benen te strekken."
- "Het zou goed zijn om even uw benen te strekken."
- "Tijd om even de benen te strekken."
- "Komt u even in beweging."
- "Tijd om even een beetje te bewegen."
- "U moet even in beweging komen."
- "Misschien kunt u even wat gaan bewegen?"
- "Het zou goed zijn om even in beweging te komen."

Encouraging - Minor - At Work Only

- "Gaat u even wat te drinken halen."
- "Is het misschien tijd om iets te drinken te halen?"
- "Loop even naar de koffieautomaat."
- "U kunt even wat te drinken gaan halen."
- "Misschien kunt u even naar de koffieautomaat lopen."
- "Loop even een rondje over de gang."
- "Misschien kunt u even een eindje door de gang wandelen?"
- "U kunt even een rondje lopen binnen het gebouw."
- "Even een rondje lopen is misschien een goed idee."
- "Tijd om even een stukje door de gang te wandelen."
- "Heeft u toevallig nog iets te doen buiten uw werkplek?"
- "Moet u misschien nog bij een collega zijn voor het een of ander?"
- "Zijn er nog werkzaken waarvoor u even uw werkplek moet verlaten?"

Encouraging - Minor - At Home Only

- "Misschien kunt u even wat klusjes gaan doen."
- "Heeft u nog klusjes te doen?"
- "Als u nog klusjes te doen heeft zou dit een goed moment zijn."
- "Moet u toevallig nog iets opruimen of schoonmaken?"
- "Doet u even wat rek en strekoefeningen."
- "Misschien kunt u even rekken en strekken."
- "Even wat rek en strekoefeningen doen zou een goed idee zijn."

Encouraging - Major - General

- "Tijd om even een frisse neus te halen."
- "U kunt even een frisse neus gaan halen."
- "Gaat u even een frisse neus halen."
- "Misschien moet u even een frisse neus gaan halen."

- “Misschien kunt u even een frisse neus gaan halen?”
- “Gaat u even een blokje om.”
- “U kunt even een blokje om lopen.”
- “Misschien kunt u even een blokje om lopen?”
- “Tijd om even een blokje om te lopen.”
- “Misschien moet u even een blokje om lopen.”
- “U kunt even een stukje buiten gaan wandelen.”
- “Misschien moet u buiten even een stukje gaan wandelen.”
- “Is een wandelingetje buiten misschien een optie?”
- “Tijd om even een stukje buiten te wandelen.”
- “Een flinke wandeling zou geen kwaad kunnen.”
- “Tijd om even een stuk actiever te worden.”
- “U moet even een stuk actiever worden.”

Encouraging - Major - At Work Only

- “Loopt u even een flink stuk over de gang.”
- “Misschien moet u even een flink stuk door de gang wandelen.”
- “Heeft u toevallig nog werk buiten het kantoor?”
- “Moet u misschien voor werk nog buiten kantoor zijn?”

Encouraging - Major - At Home Only

- “Gaat u even lekker een stuk fietsen.”
- “U kunt even een leuke fietstocht maken.”
- “Misschien moet u even een eind gaan fietsen.”
- “Een stukje fietsen zou een goed idee zijn.”
- “Heeft u nog boodschappen te doen?”
- “Heeft u nog iets nodig uit de winkel?”
- “Gaat u eens bij een bekende langs.”
- “Bezoek eens een vriend of vriendin.”
- “Misschien kunt u even een stuk gaan hardlopen.”
- “Gaat u eens een stuk hardlopen.”
- “Een stuk hardlopen zou goed voor u zijn.”
- “Gaat u eens lekker sporten!”
- “Misschien kunt u even lekker gaan sporten.”

Discouraging - Minor - General

- “Zeer goed!”
- “U bent zeer goed bezig!”
- “Het gaat zeer goed!”
- “Prima!”
- “U doet het prima!”
- “Het gaat prima!”
- “Mooi zo!”
- “Mooi dat u zo actief bent!”

Discouraging - Major - General

- “Uitstekend!”
- “U bent uitstekend bezig!”
- “Het gaat uitstekend!”
- “Geweldig!”
- “U bent geweldig bezig!”
- “Het gaat geweldig!”
- “Knap van u!”
- “Indrukwekkend!”
- “U verdient een schouderklopje!”

Appendix B

Surveys

Presented below are the surveys each participant completed during the experiment. Because the experiment was performed with Dutch-speaking participants, the surveys are only available in Dutch. The surveys were presented to the participants through an on-line service, but the appearance has been maintained as much as possible in this printed version.

B.1 Intake Survey

UNIVERSITEIT TWENTE.

Mobiele Coaching van Fysieke Activiteit voor Kantoormedewerkers

Algemene Achtergrond

Fijn dat u wilt deelnemen aan dit onderzoek! Hieronder staan een aantal vragen om relevante achtergrondinformatie over u te weten te komen. Vult u alstublieft alle vragen in.

1. Wat is het nummer van uw systeem?

Dit nummer is op vrijwel alle onderdelen van het systeem terug te vinden. Het ziet er bij voorbeeld uit als "RRDMT2001". Enkel de laatste twee cijfers zijn van belang.

2. Wat is uw geslacht?

- ○ Man
- ○ Vrouw

3. Wat is uw leeftijd?

In gehele jaren.

4. Wat is uw lengte?

In gehele centimeters. Mocht u dit niet exact weten, probeer dan een redelijke schatting te doen.

5. Wat is uw gewicht?

In gehele kilogrammen. Mocht u dit niet exact weten, probeer dan een redelijke schatting te doen.

6. Welke van de onderstaande vijf uitspraken past het beste bij u?

Met de term "regelmatig lichamelijk actief" wordt bedoeld: ten minste 5 dagen in de week minimaal 30 minuten per dag op een middelmatig of zwaarder niveau lichamelijk actief zijn (bijvoorbeeld zwemmen, stofzuigen, dansen, fietsen, wandelen, fitness). Dit hoeft niet 30 minuten aaneengesloten te zijn maar kan ook in kortere periodes van minstens 10 minuten plaatsvinden.

- Op dit moment ben ik **niet** regelmatig lichamelijk actief en ik ben **niet** van plan lichamelijk actiever te worden in de komende **6 maanden**.
- Op dit moment ben ik **niet** regelmatig lichamelijk actief maar ik denk er **wel** over lichamelijk actiever te worden in de komende **6 maanden**.
- Op dit moment ben ik **niet** regelmatig lichamelijk actief maar ik denk er **wel** over lichamelijk actiever te worden in de komende **30 dagen**.
- Op dit moment ben ik **wel** regelmatig lichamelijk actief maar ik ben daar pas in de afgelopen **6 maanden** mee begonnen.
- Op dit moment ben ik **wel** regelmatig lichamelijk actief en ben dat al **langer dan 6 maanden**.

7. Geef aan hoeveel ervaring u heeft met de onderstaande zaken.

Op een schaal van 1 tot 10, waarbij 1 staat voor helemaal geen ervaring, en 10 voor zeer veel ervaring.

	1	2	3	4	5	6	7	8	9	10
Smartphones en/of PDAs	○	○	○	○	○	○	○	○	○	○
Meetsystemen voor activiteit (bijvoorbeeld stappentellers)	○	○	○	○	○	○	○	○	○	○
Digitale coaching systemen	○	○	○	○	○	○	○	○	○	○
Niet-interactieve hulpmiddelen voor gedragsaanpassing (bijvoorbeeld zelfhulpboeken)	○	○	○	○	○	○	○	○	○	○
Embodied Conversational Agents (virtuele karakters)	○	○	○	○	○	○	○	○	○	○
Menselijke coaching m.b.t. fysieke activiteit (bijvoorbeeld therapie)	○	○	○	○	○	○	○	○	○	○

Dat was het. Bedankt voor het invullen van deze vragen. Succes bij het gebruik van het systeem de komende tijd!

B.2 Halfway Survey

UNIVERSITEIT TWENTE.

Mobiele Coaching van Fysieke Activiteit voor Kantoormedewerkers

Welkom

1 / 6

U heeft inmiddels de eerste week met het systeem erop zitten.

1. Wat is het nummer van uw systeem?

Dit nummer is op vrijwel alle onderdelen van het systeem terug te vinden. Het ziet er bij voorbeeld uit als "RRDMT2001". Enkel de laatste twee cijfers zijn van belang.

Op de volgende pagina's krijgt u een aantal vragen over uw ervaringen met het systeem in de afgelopen week. U heeft in deze week gewerkt met de eerste versie van de software. Deze vragen hebben **alleen betrekking op de software**. Probeert u dus alstublieft uw mening over de hardware zo min mogelijk door te laten wegen in uw antwoorden.

Coaching

2 / 6

Het systeem heeft u in de afgelopen week regelmatig coaching gegeven. Deze vraag gaat specifiek over de coaching die u van het systeem hebt gekregen in de afgelopen week.

2. Geef van de volgende woordparen aan wat U de beste beschrijving van het systeem als coach vindt.

Nummers 1 en 7 geven een heel sterk gevoel aan. Nummers 2 en 6 een redelijk sterk gevoel. Nummer 3 en 5 een zwak gevoel. Nummer 4 geeft een neutraal gevoel aan. Vul dit alstublieft vlot in en denk er niet te veel over na, er is geen goed of fout antwoord.

	1	2	3	4	5	6	7	
leek	○	○	○	○	○	○	○	deskundig
intelligent	○	○	○	○	○	○	○	dom
plezierig	○	○	○	○	○	○	○	onplezierig
egoïstisch	○	○	○	○	○	○	○	altruïstisch

waardevol						waardeloos
ongeschikt						geschikt
vervelend						aardig
betrouwbaar						onbetrouwbaar
eerlijk						oneerlijk
deugdzaam						zondig
onwetend						op de hoogte
onvriendelijk						vriendelijk

Coaching

3 / 6

Deze vraag gaat wederom specifiek over de coaching die u in de afgelopen week van het systeem hebt gekregen.

3. Hieronder staan een aantal stellingen over de coaching die door het systeem wordt gegeven. Geef aan in welke mate u vindt dat het systeem in de afgelopen week aan de volgende stellingen voldeed.

	Nooit				Altijd
Het systeem geeft advies over hoe ik gericht kan blijven (op mijn activiteitsdoelen).					
Het systeem houdt mijn voortgang richting mijn doel in de gaten.					
Het systeem heeft de juiste kennis en vaardigheden om goede coaching te geven.					
Het systeem laat zien dat het me begrijpt.					
Het systeem geeft advies over hoe ik vertrouwen kan blijven houden in mijn vaardigheden.					
Het systeem reikt me strategieën aan om mijn doelen te halen.					
Het systeem geeft advies over hoe ik positief over mezelf kan blijven.					
Het systeem geeft advies over hoe ik kan doorzetten.					
Het systeem helpt me te ontdekken welke dingen het best voor mij werken om actiever te worden.					
Het systeem helpt me om kortetermijndoelen te stellen en actieplannen te maken.					

Het systeem laat zien dat het om me geeft (dus ook andere aspecten van mijn leven en niet alleen lichamelijke activiteit).	◯	◯	◯	◯	◯
Het systeem geeft coaching van goede kwaliteit.	◯	◯	◯	◯	◯
Het systeem helpt me om mijn activiteitsprestaties te herkennen en vieren.	◯	◯	◯	◯	◯
Het systeem geeft me ondersteuning om mijn doelen te behalen.	◯	◯	◯	◯	◯

Gebruikservaring

4 / 6

Deze vraag gaat over de algehele ervaring van het gebruiken van de softwareversie die u de afgelopen week heeft gebruikt.

4. Geef aan in hoeverre u het met de volgende beweringen eens bent.

	Helemaal mee oneens			Neutraal			Helemaal mee eens
Het systeem gebruiken is een goed idee.	◯	◯	◯	◯	◯	◯	◯
Ik ben bang om iets fout te doen met het systeem en zo iets kwijt te raken of stuk te maken.	◯	◯	◯	◯	◯	◯	◯
Ik zou het systeem in de komende 3 maanden wel willen gebruiken.	◯	◯	◯	◯	◯	◯	◯
Het systeem maakt werk interessanter.	◯	◯	◯	◯	◯	◯	◯
Ik voel mij ongemakkelijk over het gebruiken van het systeem.	◯	◯	◯	◯	◯	◯	◯
Ik verwacht dat ik het systeem in de komende 3 maanden zou gebruiken als dat mogelijk was.	◯	◯	◯	◯	◯	◯	◯
Het systeem is voor mij enigszins intimiderend.	◯	◯	◯	◯	◯	◯	◯
Het zou makkelijk voor mij zijn om vaardig te worden in het gebruik van het systeem.	◯	◯	◯	◯	◯	◯	◯
Ik twijfel over het gebruiken van het systeem in de angst fouten te maken die ik niet kan herstellen.	◯	◯	◯	◯	◯	◯	◯
Ik zou het systeem makkelijk te gebruiken vinden.	◯	◯	◯	◯	◯	◯	◯
Het systeem leren te gebruiken is makkelijk voor mij.	◯	◯	◯	◯	◯	◯	◯
Werken met het systeem is plezierig.	◯	◯	◯	◯	◯	◯	◯
Mijn interactie met het systeem zou helder en begrijpelijk zijn.	◯	◯	◯	◯	◯	◯	◯
Ik vind het leuk om met het systeem te werken.	◯	◯	◯	◯	◯	◯	◯

Gebruikservaring

5 / 6

Deze vraag gaat ook om de algehele ervaring van het gebruik van de softwareversie die u de afgelopen week heeft gebruikt.

5. Geef van de volgende woordparen aan hoe u de versie van de software waar u de afgelopen week mee hebt gewerkt zou beschrijven.

Ieder paar bestaat uit twee extremen. De keuzeopties ertussenin stellen u in staat de intensiteit van de gekozen eigenschap uit te drukken. Besteed geen tijd aan het nadenken over de woordparen, en probeer spontaan antwoord de geven. Kies overal een antwoordmogelijkheid, ook als u van mening bent dat het woordpaar niet voldoende van toepassing is op de software.

	1	2	3	4	5	6	7	
innovatief	○	○	○	○	○	○	○	conservatief
onhandelbaar	○	○	○	○	○	○	○	handelbaar
niet toonbaar	○	○	○	○	○	○	○	toonbaar
vernieuwend	○	○	○	○	○	○	○	alledaags
motiverend	○	○	○	○	○	○	○	ontmoedigend
goedkoop	○	○	○	○	○	○	○	waardevol
praktisch	○	○	○	○	○	○	○	onpraktisch
omslachtig	○	○	○	○	○	○	○	direct
saai	○	○	○	○	○	○	○	boeiend
goed	○	○	○	○	○	○	○	slecht
buitensluitend / vervreemdend	○	○	○	○	○	○	○	integrerend / inbegrepen

lelijk						mooi
brengt mij dichter bij mensen						isoleert mij van mensen
afstotend						aantrekkelijk
menselijk						technisch
afwijzend						uitnodigend
aangenaam						onaangenaam
professioneel						amateuristisch
voorspelbaar						onvoorspelbaar
origineel						conventioneel
isolerend						verbindend
fantasieloos						creatief
eenvoudig						ingewikkeld
verwarrend						overzichtelijk
stijlvol						stijlloos
sympathiek						onsympathiek
moedig						voorzichtig
eenvoudig						uitdagend

Dank u!

6 / 6

U heeft het einde van deze vragenlijst bereikt. In de aankomende tweede week zal u de tweede versie van de software gebruiken. Het systeem schakelt vanzelf over naar deze versie als u het op dag 8 weer aanzet.

Mocht u in de loop van dag 8 helemaal geen verschil opmerken met de eerste versie, neemt u dan even contact op via jordihendrix@gmail.com, wellicht is er sprake van een technisch probleem.

B.3 Final Survey

Note that questions 2, 3, 4 and 5 of the final survey were identical to questions 2, 3, 4 and 5 of the halfway survey, and have therefore been omitted from this section. See Section B.2 for these questions.

UNIVERSITEIT TWENTE.

Mobiele Coaching van Fysieke Activiteit voor Kantoormedewerkers

Welkom

1 / 7

U heeft inmiddels de tweede week met het systeem erop zitten. Alvast bedankt voor uw deelname.

1. Wat is het nummer van uw systeem?

Dit nummer is op vrijwel alle onderdelen van het systeem terug te vinden. Het ziet er bij voorbeeld uit als "RRDMT2001". Enkel de laatste twee cijfers zijn van belang.

Op de volgende pagina's krijgt u een aantal vragen over uw ervaringen met het systeem in de afgelopen week. Deze vragen gaan dus **enkel over de tweede versie van de software**. Probeert u alstublieft uw ervaringen met de eerste versie zo veel mogelijk buiten beschouwing te laten bij het beantwoorden van deze vragen.

Wederom is het zo dat de vragen **alleen betrekking hebben op de software**. Probeert u dus alstublieft uw mening over de hardware zo min mogelijk door te laten wegen in uw antwoorden.

Vergelijking Softwareversies

6 / 7

Gedurende het onderzoek heeft u gewerkt met twee verschillende versies van de software: een waarbij de feedbackberichten werden gepresenteerd in de vorm van text (de "Textversie"), en een waarbij de feedbackberichten werden gepresenteerd door een geanimeerde agent (de "Agentversie"). In de volgende vraag moet u expliciet vergelijken tussen deze twee softwareversies. Hierbij mag u dus uiteraard wel uw ervaringen met beide versies laten meewegen.

6. Geef alstublieft aan welke van de twee versies van het systeem u...

	Textversie		Geen verschil		Agentversie
... irritanter vond.	○	○	○	○	○
... omslachtiger vond in het gebruik.	○	○	○	○	○
... geloofwaardiger vond.	○	○	○	○	○
... leuker vond.	○	○	○	○	○
... accurater vond in het meten vanuw activiteit.	○	○	○	○	○
... vaker boodschappen van genegeerd heeft.	○	○	○	○	○
... prettiger vond in het gebruik.	○	○	○	○	○
... mooier vond.	○	○	○	○	○
... meer aanzette tot activiteit.	○	○	○	○	○
... repetitiever vond.	○	○	○	○	○
... vaker adviezen van opgevolgd heeft.	○	○	○	○	○
... vaker geraadpleegd heeft.	○	○	○	○	○
... liever voor een langere periode zou gebruiken.	○	○	○	○	○
... beter advies vond geven.	○	○	○	○	○

Dank u!

7 / 7

U heeft het einde van de vragenlijst bereikt. Nogmaals hartelijk bedankt voor uw deelname, hopelijk heeft u het leuk gevonden.

Mocht u nog verdere vragen en/of opmerkingen hebben over het onderzoek, dan is daar uiteraard ruimte voor tijdens het afsluitende gesprekje. Ook kunt u natuurlijk altijd mailen naar jordihendrix@gmail.com.

Appendix C

Forms and Information

Presented here are the papers each participant was given during the introductory explanation. Because the experiment was performed with Dutch-speaking participants, these are only available in Dutch.

C.1 General Information Sheet

Informatiesheet Onderzoek

Mobiele Coaching voor Fysieke Activiteit van Kantoormedewerkers

Wat houdt het onderzoek in?

Dit onderzoek gaat om een mobiel coachingsysteem ter bevordering van fysieke activiteit, bedoeld voor kantoormedewerkers die interesse hebben in het onderhouden van de hoeveelheid fysieke activiteit die zij dagelijks ondernemen. Het onderzoek zal twee versies van de software voor dit systeem vergelijken.

Waaruit bestaat het testsysteem?

Het systeem bestaat uit een sensor die op de heup gedragen wordt, en een smartphone die op een willekeurige plek kan worden gedragen. De sensor meet de fysieke activiteit van de proefpersoon. De coachingsoftware op de smartphone zal deze activiteit weergeven en regelmatig de gebruiker proberen te coachen. Het is helaas niet mogelijk om het systeem te installeren op een eigen smartphone.

Welke gegevens worden er verzameld en wat gebeurt hiermee?

De sensor verzamelt gegevens over de mate van fysieke activiteit, welke wordt opgeslagen. Verder zullen er een aantal vragenlijsten moeten worden ingevuld (een korte vooraf, een tussendoor en een achteraf) en zal er ter afsluiting eventueel een kort interview worden afgenomen. De gegevens die hieruit worden verzameld zullen vertrouwelijk worden behandeld door mij (Jordi Hendrix, afstudeerder bij HMI aan de Universiteit Twente) en mijn begeleiders bij de Universiteit Twente en Roessingh Research & Development. Anonimiteit is uiteraard gegarandeerd.

Wat wordt er verwacht van proefpersonen?

Er wordt verwacht dat proefpersonen het systeem 14 dagen lang gebruiken (7 dagen per versie van de software). Dit houdt in van ’s ochtends vroeg tot ’s avonds laat (voor zover redelijkerwijs mogelijk) de sensor en smartphone bij zich dragen. Proefpersonen zijn uiteraard niet verplicht de coaching van het systeem op te volgen.

C.2 Journal

Dagboek

Mobiele Coaching van Fysieke Activiteit voor Kantoormedewerkers

Naam: _____

Systeemnummer: _____

Omdat het onderzoek zich richt op kantoormedewerkers is het belangrijk om te weten wanneer u heeft gewerkt. Vul daarom hieronder alstublieft voor iedere dag in hoeveel uren u op die dag gewerkt heeft. Mocht u nog relevante opmerkingen hebben over de betreffende dag (bijvoorbeeld als u het systeem tijdelijk niet heeft gedragen wegens omstandigheden), dan kunt u die hier ook kwijt.

Dag 1	**LET OP: Vult u alstublieft de intake-vragenlijst in.** **U vindt deze op https://nl.surveymonkey.com/s/activiteit-intake**
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 2	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 3	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 4	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 5	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 6	
Datum:	
Gewerkt:	uur
Opmerkingen:	

Dag 7	**LET OP: Vult u aan het eind van deze dag (en in ieder geval voordat u het systeem morgen weer aanzet) alstublieft de tussentijdse vragenlijst in. U vindt deze op https://nl.surveymonkey.com/s/activiteit-tussentijds**
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 8	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 9	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 10	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 11	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 12	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 13	
Datum:	
Gewerkt:	uur
Opmerkingen:	
Dag 14	**LET OP: Vult u in ieder geval voor het afsluitende gesprek alstublieft de laatste vragenlijst in. U vindt deze op https://nl.surveymonkey.com/s/activiteit-einde**
Datum:	
Gewerkt:	uur
Opmerkingen:	

C.3 Informed Consent Form

Toestemmingsverklaring (Informed Consent)

Ik heb de informatiesheet over het onderzoek gelezen. Ik heb de mogelijkheid gehad aanvullende vragen te stellen. Mijn vragen zijn naar tevredenheid beantwoord. Ik heb genoeg tijd gehad om te besluiten of ik wilde deelnemen.

Ik ben mij ervan bewust dat deelname geheel vrijwillig is. Ik ben mij ervan bewust dat ik op elk willekeurig moment kan besluiten (zonder opgaaf van redenen) alsnog niet deel te nemen.

Ik geef toestemming dat bevoegde personen van de Universiteit Twente en Roessingh Research & Development en bevoegde autoriteiten inzage kunnen krijgen in mijn onderzoeksgegevens.

Ik geef toestemming om mijn gegevens te gebruiken voor de doelen die in de informatiesheet genoemd zijn.

Ik verklaar de apparatuur behorende bij het onderzoek (plastic koffertje met daarin sensor, smartphone en toebehoren, zoals beschreven in de handleiding) te hebben ontvangen, en deze (in dezelfde staat) weer te zullen retourneren na afloop van het onderzoek.

Ik ben akkoord met deelname aan dit onderzoek.

Naam proefpersoon:

E-mailadres:

Datum: __ / __ / 2013

Handtekening:

Ik verklaar hierbij dat ik bovenstaande proefpersoon volledig heb geïnformeerd over het genoemde onderzoek.

Naam onderzoeker: Jordi Hendrix

Datum: __ / __ / 2013

Handtekening: