How Do Community-Dwelling Older Adults Perceive the Usability of an Embodied Conversational Agent?

Student: Sander Groot

Study programme: MSc Management, Economics and Consumer studies

Registration number: 960714285050

Chair group: COM

Course code: CPT-81336

Supervisors: Bob Mulder (CPT) & Lean Kramer (CHL)

Wageningen University

March, 2020

Abstract

Background

Community-dwelling older adults are susceptible to chronic disease and social disengagement. Embodied conversational agents (ECAs) may have huge health benefits for this group, since users can establish an empathic relationship with ECAs, which is proven successful in lifestyle coaching. Furthermore, they are available for coaching at any moment in time. However, literature on evaluating the usability of ECA applications is lacking.

Aims

This study aims to provide insights into the usability of an ECA technology that is developed using a stakeholder-centered design approach. Moreover, we shed a light on several points of attention for application developers of ECA technologies to improve the fit between application and end user.

Methods and procedures

A usability evaluation of the PACO (Designing Persuasive E-Health Agents for Coaching Older adults) application was conducted. The System Usability Scale (SUS), Think Aloud, post-task interviews, task completion (time) and task satisfaction were used as benchmarking instruments to measure usability and identify usability issues.

Outcomes and results

In total 10 participants, of which six females and four males, participated in this study. Ages ranged between 71 and 93 (M=78.0, SD=8.2). On a scale of 0 to 100, participants evaluated the overall system usability with a score of M=72.0 (SD=25.7, median 80.0). Task satisfaction scores, ranging from 1 (very low) to 7 (very high) were M=6.6 (SD=.5) for ‘Account creation’ and M=6.4 (SD=.7) for ‘Recipe book’. In total 16 usability issues were identified, of which 5 were minor (31.25%), 6 were serious (37.50%) and 5 were critical (31.25%).

Conclusions and implications

The overall SUS score of the ECA application was good and task satisfaction scores were high. Various usability issues have been identified on a User Interface and User Experience level. This study provides insights into the usability of an ECA technology. The outcomes may be useful for ECA developers in creating a better fit between application and user.

Contents

1. Introduction

Ageing populations across the globe are associated with a growing burden of chronic disease (Liu et al., 2019). This makes the promotion of healthy ageing a valid point of attention. One of the key contributors to healthy ageing is diet (Reedy et al., 2014; Yin et al., 2017). However, dietary behavior of older adults is influenced by physiological changes that impact their quality of life through negative health outcomes and a lower general functioning as they get older (Hughes, Bennett & Hetherington, 2004). Common physiological changes are chewing problems and changes in taste and smell (Robinson, 2018). These alterations result in a loss of appetite and energy intake in older adults, as well as a negative shift in food choices and eating patterns (Hetherington, 1998). In general, this group tends to consume too much fat and salt, and too few wholemeal products, fruit and fish than dietary guidelines recommend (Ocké et al., 2013). Adhering to an unhealthy diet is a risk factor for cardiovascular diseases (WHO, 2017), anxiety and depression (Bonnet et al., 2005), and results in an increased mortality risk (Ford et al., 2014). Thus, the abovementioned negative health outcomes stress the importance of altering the dietary behavior of older adults to improve their quality of life.

In addition to changes in dietary behavior of older adults, this group must deal with environmental changes, such as a decrease in social engagement (Luo, Hawkley, Waite & Cacioppo, 2012). Previous research has identified the negative health effects of social disengagement among individuals. Experiencing a lack of social relationships is a cause and risk factor of mortality, as well as morbidity (House, Landis & Umberson, 1988; Brummett et al., 2001). Interventions aimed at improving social engagement are beneficial for health and well-being, because research demonstrated that it results in reduced stress (Heinrichs, Baumgartner, Kirschbaum & Ehlert, 2003), lower anxiety and depression (Thompson, 1989),

and increased treatment adherence (DiMatteo, 2004). Thus, stimulating older adults to establish and maintain social connections is beneficial for their quality of life.

eHealth has shown to be an effective medium to engender diet-related behavior change (West, Belvedere, Andreasen, Frandsen, Hall & Crookston, 2017) as well as reducing feelings of loneliness (Ring, Shi, Totzke & Bickmore, 2014). Furthermore, internet-based interventions can be more effective in treating mental disorders compared to regular face-to-face treatments (Andrews, Cuijpers, Craske, McEvoy & Titov, 2010). Numerous different eHealth tools exist, such as applications, websites, devices, video consults and webinars (Kampmeijer, Pavlova, Tambor, Golinowska & Groot, 2016). Such tools contain techniques or content to stimulate certain behavior changes (Ahern, Kreslake & Phalen, 2006). eHealth enables patients and health care professionals to use technology for health communication (Shaw et al., 2017).

An emerging tool in internet-based interventions in clinical psychology are embodied conversational agents (ECAs). Such agents are a promising tool to excite behavior change (Kramer, Ter Stal, Mulder, De Vet & Van Velsen, 2020). ECAs are virtual embodied representations of humans designed to interact with the user face-to-face to support or stimulate healthy behavior (Hartmann, Mancini, Buisine & Pelachaud, 2005). The relationship between an ECA and the user can contribute to trust, rapport and therapeutic alliance, which increases the chances of adherence to self-care treatments (Bickmore, 2010). ECAs use several relational behaviors, such as social dialogue, feedback, humor, facial expressions and body language (Kramer et al., 2020). Through these behaviors, an empathic relationship is established between the user and the ECA. This social bond motivates the user to continue the interactions with the ECA (Bickmore, Caruso, Clough-Gorr & Heeren, 2005). However, it should be kept in mind that an ECA would serve as a social support booster, rather than a substitution for real social connections (Loveys, Fricchione, Kolappa, Sagar &

Broadbent, 2019). ECA designers should keep this in mind, because older adults should still be stimulated to maintain their social network.

The role of ECAs in eHealth is promising, but limited understanding of the users, their needs and the context in which the technology is used cause such technologies to have a higher chance of failure (Kayser, Kushniruk, Osborne, Norgaard & Turner, 2015). Testing the usability of a system is therefore desirable. Performing a usability evaluation during the design process is crucial, since being unable to use the technology correctly prevents users from reaching a desired outcome (Bickmore et al., 2010; Corrao, Robinson, Swiernik & Naeim, 2010; Niranjanamurthy, Nagaraj, Gattu & Shetty, 2014). Involving users in the evaluation has positive effects on the success of a system and user satisfaction, and developers get more accurate user requirements for their system (Kujala, 2003). Technologies contain less usability issues when developed using a stakeholder-centered approach (Van Gemert-Pijnen et al., 2011). However, literature on the usability of ECA applications designed using a stakeholder-centered approach is lacking (Kramer et al., 2020).

Various methods to evaluate the usability of an ECA exist, such as questionnaires, think aloud and heuristic evaluation. The most frequently used evaluation methods will be discussed in more detail later. Before performing the usability evaluation, multiple factors should be carefully considered, such as the device to use, type of interaction method, data collection method, data analysis method, and the number of participants and their characteristics (Dumas & Redish, 1999).

The aim of this study is to assess the usability of an ECA aimed at community-dwelling older adults and provide practical suggestions that ECA developers can use in practice. The following research question will be answered: *"To what extent do community-dwelling older adults perceive an ECA application as usable and satisfactory, and what*

modifications do they suggest to make it more so?". In addition, we assess the maturity of the ECA application and compare this with similar technologies in relation to usability.

This study provides insights into the usability of an ECA technology that is developed using a stakeholder-centered design approach. Moreover, we provide practical points of attention for developers of ECA technologies that should be considered in order to improve the fit between application and end user.

2. Theoretical framework

This chapter describes an overview of methods frequently used in usability evaluations in the eHealth domain, and the theory of Technology Readiness Levels. At last, a selection of methods suitable for the evaluation of an ECA application is made, and its maturation stage at the time of the evaluation is described. The selected methods will be used in the assessment of the usability of an ECA technology. The degree of maturity will be used to compare the assessed ECA technology to similar technologies in terms of usability.

2.1 Usability benchmarking instruments in eHealth

Development and usage of technologies increased in the 20th century. The accessibility of personal computers in the 1980s increased, but this technology appeared to be complex for regular users. Therefore, identifying issues that could hinder the usability of a certain technology became more relevant during this time period. Usability evaluations existed long before eHealth applications were introduced.

The definition of usability differs between research fields. In the field of Human-Computer Interaction, usability is defined as the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use (Bevan, 2009). The user interacts with a User Interface (UI) to communicate with the application, e.g. by clicking on buttons. This view implies the interaction of the user with the software to see to what extent the software meets the user's expectation (Fernandez, Insfran & Abrahão, 2011). User Experience (UX), is a broader term than usability and refers to the thoughts, feelings and perceptions that users have when interacting with a system (Heo, Ham, Park, Song & Yoon, 2009).

Usability evaluations are conducted to investigate user interaction with a technology and to see whether the properties of the technology contribute to the overall usability. These

evaluations require a set of well-defined methods to gather information. Existing literature classifies evaluation methods into two types. Empirical methods involve representative end users and inspection methods that involve experts or designers (Fernandez et al., 2011). Usability evaluations are widely recognized as essential for the success of interactive health care applications (Jaspers, 2009). The generated insights can support the development of eHealth technologies by identifying system deficiencies. For instance, in a study on an ECA as virtual coach for physical activity, participants stated that the timing and content of the feedback needed improvement (Op Den Akker, Klaassen & Nijholt, 2016). End users and usability experts are both useful to identify usability issues. Experts are more suitable to evaluate earlier prototypes, while end users should evaluate a refined version to simulate a real-life setting (Tory & Moller, 2005).

Nowadays, there are numerous methods to evaluate the usability of an eHealth system, with questionnaires, task completion, thinking-aloud, interviews, heuristic evaluation and focus groups being the most widely used (Maramba, Chatterjee & Newman, 2019). These methods will be elaborated in this chapter. Each usability evaluation method aims to gather information about a system's deficiencies, e.g. usability issues or low user satisfaction, in order to be able to improve the system afterwards.

2.1.1 Questionnaires

Questionnaires, either as primary or secondary data collection method, are one of the most widely used methods to assess the usability of a system (Klaassen, Van Beijnum & Hermens, 2016). The System Usability Scale (SUS) is the predominant survey scale in the field of usability testing of products or services. It was developed by Brooke (1996) as a 'quick and dirty' usability scale with 10 statements that are scored on a five-point Likert scale ranging from strongly disagree to strongly agree. The outcome score ranges between 0 and 100. Scores above 70 are considered 'good', while better systems score in the high 70s to upper

80s. Systems with a score above 90 are considered truly superior (Bangor, Kortum & Miller, 2008). The SUS has been widely used by usability practitioners to assess the usability of a large range of interface technologies (Bangor et al., 2008).

The popularity of questionnaires as a method to assess usability can be explained by the advantages of easy distribution, automated analysis of results and cost-friendliness. In addition, they allow for receiving quick feedback on how the system is perceived by the user (Kushniruk & Patel, 2004).

However, the usefulness of questionnaires as a primary method for evaluating the usability of a system is limited. They collect opinions of users about the interface, instead of studying the actual user interface, which makes them an indirect method to assess usability (Holzinger, 2005). In addition, questionnaires contain items that are predetermined by the investigators, which makes them limited in terms of identification of usability issues that the investigators have not thought of when creating the questionnaire (Kushniruk & Patel, 2004). In particular, the SUS is less effective in identifying usability issues than task metrics, which makes using this scale insufficient as a standalone usability benchmark for eHealth technologies. Furthermore, it does not take groups with physical or cognitive impairments into account (Harrati, Bouchrika, Tari & Ladjailia, 2016; Broekhuis, Van Velsen & Hermens, 2019). Thus, questionnaires can generate valuable insights in the subjective values of users about a system but may be insufficient when no additional usability benchmarks are used.

2.1.2 Task metrics

Maramba et al. (2019) found that task completion is one of the most frequently used methods to assess a system's usability. The degree of completed tasks is an indicator for the usability of the system. There are various additional metrics that indicate the usability of a system, such

as the task duration, number of clicks required to complete the task, cursor distance (Harrati et al., 2016), task satisfaction, errors on task and steps per task (Broekhuis et al., 2019).

Task metrics can be applied in a usability evaluation to assess how efficiently and satisfactory certain tasks are performed by participants (Sauro & Lewis, 2012). Performance metrics such as cursor distance, mouse clicks and task duration can be computed automatically (Harrati et al., 2016). Task satisfaction, errors on task and steps per task require the data to be handled manually, by using e.g. questionnaires or screen recording. For instance, task satisfaction can be measured using the After-Scenario Questionnaire, which consists of three items measured on a seven-point Likert scale, ranging from strongly disagree to strongly agree (Lewis, 1991).

2.1.3 Think Aloud

The think aloud method was introduced to overcome the problem of knowing what users are doing, but not knowing why do they it. In contrast to the indirect nature of questionnaires, the think aloud method is a usability evaluation method that generates a direct understanding of which part of the system causes problems for the user (Holzinger, 2005). It allows for direct identification of shortcomings of the system, and generates knowledge of the way humans solve problems, since ongoing thought processes are exposed by the participant (Jaspers, Steen, Van Den Bos & Geenen, 2004).

There are two variants of the think aloud method: concurrent and retrospective. Retrospective thinking aloud involves participants silently carrying out predefined tasks, after which they are asked to verbalize their thoughts in retrospect, either with or without the support of screen recording (Van Den Haak & De Jong, 2003). Using retrospective rather than concurrent think aloud has several benefits. First, participants in a retrospective think aloud condition may perform more successful in terms of task completion and encountered

problems compared to participants in a concurrent think aloud condition, due to the increased workload of the latter (Van Den Haak, De Jong & Schellens, 2003). The support of screen recording is important when utilizing a retrospective think aloud protocol, due to participants forgetting what actions they performed during task completion (Branch, 2000). Second, measuring task completion time in a concurrent think aloud setting may result in incorrect times, since thinking out loud is likely to slow down the task execution (Van Den Haak et al., 2003).

The other variant, concurrent thinking aloud, requires participants to continuously verbalize their thoughts while interacting with a system by solving a problem or performing a task (Jaspers et al., 2004). Using concurrent instead of retrospective thinking aloud has a few advantages. During a retrospective think aloud protocol, participants may be biased in recalling the thoughts they had during task performance, e.g. by forgetting certain things that occurred during the evaluation. That way, vital information may be lost when using a retrospective think aloud protocol, which is not the case when using concurrent thinking akoud (Russo, Johnson & Stephens, 1989; Teague, De Jesus & Ueno, 2001). Furthermore, the duration of the evaluation session differs between both think aloud methods. Conducting a usability evaluation when using concurrent think aloud usually takes shorter than a session in which the retrospective thinking aloud is used (Van Den Haak et al., 2003).

Overall, Van Den Haak et al. (2003) found that both methods are equal in terms of quantitative output, but different in the way this output is established. The concurrent think aloud method represents a strict task-oriented usability test, while the retrospective think aloud method allows for a broader range of user responses. Both alternatives of the think aloud methods have their advantages and disadvantages, and it depends on the goals of the usability evaluation which method is preferred.

The test moderator's role is essential in making sure the think aloud protocol results in valuable insights. The moderator should disturb the cognitive processes of the participant as little as possible but, in the case of concurrent thinking aloud, should remind or probe participants to keep talking when they stop verbalizing their thoughts, e.g. by saying "Keep talking" or "Um-humm". Conversations between moderator and respondent are avoided and assistance is kept to a minimum (Bruun & Stage, 2015).

2.1.4 Interviews

Another qualitative method to explore the usability of a system is by interviewing participants. Interviews are widely used in multiple fields of research. In usability evaluations, they are often used as an addition to methods that are applied in task performances, such as the Think Aloud method, instead of being used as a standalone benchmark to test the usability of a system.

One purpose of interviews is to find out what first impressions and attitudes people have towards a product, before the actual usability evaluation is conducted. This knowledge is useful, since a person's attitudes may set the stage for the actual usage. For instance, if someone thinks a website looks easy to use, the threshold to use it is low and this person may spend more time on it than someone who had thinks the website appears difficult to use (Rubin & Chisnell, 2008).

Another purpose of interviews is to gain further insights into usability issues that the tasks did not unveil during the evaluation, and enable participants to provide explanations of their opinions, thoughts and actions (Page, 2014). A structured debriefing session with the participants may result in more valuable insights (Horsky et al., 2010).

2.1.5 Heuristic evaluation

Heuristic evaluation is a usability evaluation method for eHealth technologies that uses a set of principles of product design (called heuristics), e.g. appropriate functionality, that can be investigated by the evaluator before empirical tests are conducted (Baumel & Muench, 2016). This evaluation method was introduced by Nielsen and Molich (1990) that enabled expert evaluators to assess computer technologies using a small set of guidelines (heuristics) and indicate violations of the heuristics. These violations are indicators of usability issues of a system.

More recently, Kientz, Choe, Birch, Maharaj, Fonville, Glasson and Mundt (2010) developed a list of 10 heuristics to evaluate persuasive health technologies. These heuristics did not only relate to the usability of user interfaces, but also considered aspects of the design that were related to motivational strategies (Baumel & Muench, 2016). Their guidelines were validated and turned out to find more important usability issues than the original heuristics proposed by Nielsen (2005). The knowledge about heuristic evaluations was extended by including the ability to determine the level of eHealth literacy of lay people. Heuristics were used to not only investigate the usability, but also determine the degree of health literacy of people (Monkman and Kushniruk, 2013).

Evaluating a system's usability using heuristics is attractive for evaluators due to its cost-efficient nature, not requiring extensive planning and the possibility to be used early in the development process. In contrast, this method may identify usability issues without providing suggestions on how to resolve them, and it relies heavily on the ideas and mindset of the evaluators (Nielsen & Molich, 1990).

2.1.6 Focus groups

Compared to all previously explained methods, focus groups are used less frequently (Maramba et al., 2019). This is reflected in the amount of available literature with regards to focus group as usability evaluation method. Focus groups generate ideas and data through communication between research participants. It is a form of group interview, but participants are encouraged to talk to each other about a specific topic, instead of receiving questions from the researcher one by one (Kitzinger, 1995). Rubin and Chisnell (2008) argue that focus groups are not suitable to identify shortcomings of a system. They are, however, suitable for gathering general, qualitative information. Conducting a focus group is appropriate to find out what representative users think about preliminary concepts of a project. It is more about acceptability and usefulness of these concepts, rather than usability. This differentiates focus groups from techniques such as the think aloud method and heuristic evaluation.

2.2 Technology Readiness Level

Technology Readiness Levels (TRLs), as depicted in Table 1, are used to assess the maturity of a certain technology and they allow for comparisons of multiple technologies in terms of their maturity (Mankins, 1995). This concept was first introduced by the National Aeronautics and Space Administration (NASA) during the 1970s to categorize the maturity of new technologies in a more effective way (Mankins, 2009). Even though the concept of TRLs was developed for technologies in astronautics, it is usual to use them in different contexts, e.g. in eHealth.

The lowest level of technology maturation (TRL 1) is characterized by the observation and reporting of basic research principles. These principles “begin to be translated into more applied research and development (Mankins, 2009, p. 1217)”. At TRL 2, practical applications are identified based on the basic principles of the first level. TRL 3 is characterized by the initiation of active research and development. In this stage, an application

or concept, as identified in TRL 2, is proved by both analytical and experimental approaches. Some inventions (usually those that include unknown concepts) specifically require an experimental approach, while an analytical approach on more common, straightforward inventions will suffice. When an application or concept is successfully proven, the next level of maturity (TRL 4) integrates all elements of the invention so they work together. This is still on relatively low level in terms of appliance, which changes significantly at TRL 5. During this stage, the basic components are integrated with realistic supporting components, which creates a total application that can be tested in a, to some extent, realistic environment. At TRL 6, a prototype of the system is demonstrated in a relevant environment, followed by TRL 7, in which the prototype is nearly finished and demonstrated in the expected operational environment. TRL 8 represents the end of the system development where the system is completed and successfully tested. It can now be deployed and used in practice (TRL 9). The difference between levels eight and nine is in the actual operation (Mankins, 2009). For example, creating an ECA application is TRL 8, and launching that application is TRL 9.

Table 1. Overview of the technology readiness level scale.

TRL	Definition
Level 9	Actual system "proven" through successful system and/or mission operations
Level 8	Actual system completed and "qualified" through test and demonstration (in the operational environment)
Level 7	System prototype demonstration in the planned operational environment
Level 6	System/subsystem model or prototype demonstration in a relevant environment
Level 5	Component and/or breadboard validation in relevant environment
Level 4	Component and/or breadboard validation in laboratory environment

Level 3	Analytical and experimental critical function and/or characteristic proof-of-concept
Level 2	Technology concept and/or application formulated
Level 1	Basic principles observed and reported

2.3 When and how to evaluate an ECA application

Chapter 2.1 described methods to evaluate the usability of eHealth technologies. To determine which methods will be used in the usability evaluation of an ECA technology, multiple factors will be considered. In this study, community-dwelling older adults will test a specific eHealth application. The age of the group that is going to participate in the usability evaluation is important to consider in order to make a correct decision of the evaluation methods. The full methodology is explained in chapter 3.

Despite the absence of substitution for the SUS and its lack of perfection, it is still a valid option when it is combined with different methods. Additionally, due to possible concentration issues of the participants because of their ages, a concurrent think aloud will be performed, since it less time intensive than its counterpart. Moreover, a post-task interview will be conducted to generate more insights in usability issues and additional suggestions by the participants. In terms of task metrics, task satisfaction will be measured to test whether participants enjoy communicating with the ECAs. Completion rate and completion time will be monitored to discover whether participants are having trouble with the application.

Based on the TRLs as described in chapter 2.2, I will discuss the maturity of the specific ECA application that is going to be evaluated. At the time of the usability evaluation, the application can be subdivided into the sixth level of technology readiness. The environment that the prototype is tested in represents a living room, which is the environment

the final version of the application will be used in. After the prototype evaluation, the next step is to demonstrate the application in the planned operational environment, which are the residences of the end users.

3. Methodology

3.1 PACO Web Application

This study assessed the usability of an initial version of the PACO web application. PACO is an acronym of "Designing Persuasive E-Health Agents for Coaching Older adults towards dietary behavior change". The aim of the PACO application is to stimulate dietary change and social connections in community-dwelling older adults using ECAs. The embodied conversational agents of PACO are called Ellen and Herman and are pictured in figure 1. The final version of the PACO web application will consist of five modules: eating diary, goal book, recipe book, stories, and chat. The purpose of the eating diary is to enable users to keep track of their food intake during the day to make them more aware of what they consume. The goal book allows users to set certain goals for themselves (e.g. trying a new recipe every Friday) and keep track of whether they reached their goals or not. The recipe book contains a range of various meals that users can cook, including ingredient and nutrient list, and preparation method. The stories module is aimed at reducing the feeling of loneliness and consists of recorded stories of older adults in which they share their experience about fun activities. Users then have the possibility to sign up for these activities. The chat enables users to connect with and talk to other users and share their thoughts, problems, experiences, etc.

Figure 1. The embodied conversational agents of PACO, Ellen (left) and Herman (right). Ellen: "Hello, what would you like to do today?".

In order to further improve the PACO application, an analysis of the application's usability is required. Therefore, the aim of this usability evaluation was to see to what extent the PACO application was perceived as usable and satisfactory, to investigate which usability issues came up, and to receive additional suggestions to improve the application. It was set up as an individual evaluation with a concurrent think aloud design, so participants interacted with the system by performing a series of predefined tasks while verbalizing their thoughts and actions. The chat module was left out of the usability evaluation, since it was not yet available at the time of the evaluation. Therefore, the modules that were tested in this evaluation were: eating diary, goal book, recipe book, stories, and, additionally, creating an account was included. The recipe book could be evaluated within the application itself, but the other modules required an interactive PowerPoint with mockups, because these modules were not built in at the time of the evaluation. Because of this, during the evaluation of the modules 'eating diary', 'goal book' and 'stories', the participant was guided through the task by the moderator, while they still shared their thoughts and actions out loud. A more detailed description of the study procedure can be found in chapter 3.3.

3.2 Participants

In total, 10 participants took part in the usability evaluation. They were recruited by Roessingh Research and Development and Het Nationaal Ouderenfonds from a panel consisting of older adults. An e-mail containing information about the evaluation and a consent form were sent to the participants two weeks before the evaluation (appendix A).

Participants were included when they were retired, community-dwelling, native speaker or C2 level based on the European framework, willing to provide informed consent, and had some knowledge and skills regarding the use of tablets. All participants lived in the Netherlands, in the regions of Enschede and Amersfoort. It should be noted that they participate more often in similar studies.

3.3 Study Procedure

Prior to the actual evaluation, each participant practiced thinking out loud by performing a task that was irrelevant to the research while verbalizing his/her thoughts and actions. They were asked to plan a journey from their hometown to Amsterdam using the NS (Dutch Railways) website. This was done to make them more comfortable with the think aloud method.

During the evaluation session, each participant was asked to carry out three of the five tasks while verbalizing all thoughts and actions. Table 2 provides an overview of the scenarios that each participant was assigned to. Participants had five minutes to complete each task. If they failed to complete the task within that time frame, or when they did not want to finish it, they proceeded to the next task.

For task A, participants were asked to create an account in the system. Task B required participants to find a certain recipe, task C to put a meal in the system, task D to create a new goal for themselves, and task E to sign up for a certain activity. The full descriptions of tasks can be found in appendix C. Appendix B shows screenshots of when a task starts and ends.

Table 2. Allocation of scenarios to be tested by the participants. A = Account creation, B = Recipe book, C = Eating Diary, D = Goal book, E = Stories.

Participant	**Scenarios**		**Participant**	**Scenarios**
Participant 1	A | B | C		Participant 6	B | C | E
Participant 2	B | D | E		Participant 7	B | D | E
Participant 3	B | C | D		Participant 8	A | C | D
Participant 4	A | B | E		Participant 9	B | C | E
Participant 5	A | C | D		Participant 10	A | D | E

After completing a task, or when the time ran out, the participants were given the After-Scenario Questionnaire (appendix D) to measure task satisfaction. This questionnaire consisted of three items measured on a seven-point Likert scale ranging from completely disagree to completely agree. This was repeated until the three tasks were carried out. Each participant then filled out the SUS (appendix E), a system usability questionnaire consisting of 10 statements measured on a five-point Likert scale. Next, a short interview was conducted in which each participant received the same questions: a general opinion of the application, what people would like to see differently, what people liked, and their opinion about the tone of voice of the ECAs. Afterwards, participants' demographic and other information (gender, age, education, household composition, and information concerning technology usage) was acquired (appendix F). A complete and detailed overview of the usability protocol can be found in appendix G. The usability tests had an average length of 35 minutes and 25 seconds. Audio and screen capture recordings were made during the evaluation.

3.4 Data analysis

The audio recording of each evaluation session was transcribed to identify usability issues. A free transcription web app, oTranscribe, was used to manually write verbatim transcriptions. Screen capture recordings were used alongside the audio recordings in the analysis. The issues that occurred were reported and categorized in Microsoft Excel by their severity, location within the application, whether it was a User Interface or User Experience issue and its frequency. Even though usability and user experience do not refer to the same concept, UX issues in this study are considered as issues related to thoughts and feelings of participants. In contrast, UI issues are related to the interface of the application. Additionally, a solution to each usability issue was provided by the usability moderator. The severity of an issue is classified as either critical, serious or minor. Critical issues prevent the user from completing tasks and/or recur across all participants. Serious issues do not prevent the user from completing the task successfully but increase the task completion time and/or recur frequently across participants. Minor (cosmetic) issues increase task completion time slightly and/or recur infrequently across participants. They do not hinder participants in completing a task at all (Duh, Tan & Chen, 2006).

4. Results

4.1 Demographics

In total 10 participants, of which six females and four males, participated in this study. Ages ranged between 71 and 93 (M=78.0, SD=8.2). Seven participants lived together with their significant other, and three lived on their own due to the passing away of their significant other. In terms of technological devices, most participants (seven out of 10) owned a smartphone, PC or laptop, and a tablet. In addition, most of the participants (seven out of 10) had a vocational or higher vocational education.

4.2 Usability benchmarks

On a scale of 0 to 100, participants evaluated the overall system usability with a score of M=72.0 (SD=25.7, median 80.0). The tasks to evaluate the usability of modules that were already included in the application during the time of the research, ‘Account creation’ and ‘Recipe book’, were both completed correctly and within the time limit of five minutes in each occasion. An overview of task completion rate and task completion time for tasks A and B can be found in Table 3. Task satisfaction scores, ranging from 1 (very low) to 7 (very high) were M=6.6 (SD=.5) for ‘Account creation’ and M=6.4 (SD=.7) for ‘Recipe book’.

Table 3. Task metrics results of tasks A and B.

	Task A: Account creation	**Task B: Recipe book**
Completion rate	6/6 (100%)	6/6 (100%)
Completion times Time (Participant ID)	02:04 (1)	02:07 (1)
	01:51 (2)	04:53 (3)
	01:37 (4)	02:22 (4)
	03:07 (5)	01:31 (6)

	04:38 (8)	02:36 (7)
	04:34 (10)	04:04 (9)
Average completion time	02:58	02:55

4.3 Usability issues

In total 16 usability issues were identified, of which 5 were minor (31.25%), 6 were serious (37.50%) and 5 were critical (31.25%). A full overview of usability issues including location and solution can be found in appendix H. The issues are divided into User Interface (UI) problems and User Experience (UX) problems. Six usability issues were attributed to UI and 10 issues were attributed to UX.

4.3.1 UI Issues

A critical UI problem that was identified during the usability evaluation was the malfunction of a button to finish creating an account. Four participants were not able to get to the home screen due to this issue. This issue was addressed between evaluation days and resolved after the fifth participant. Another critical UI issue concerned the interface on a Samsung tablet. The interface became small and hard to read when the keyboard was opened to type the username and password to create an account. Although it did not hinder them in successfully completing the task, it caused trouble for all participants attempting to create a PACO account on a Samsung tablet.

Serious UI issues included the participants' urge to click on unclickable buttons (two out of 10 participants felt this urge) and the recipe book being too small to read, encountered by two participants. *"I wanted to make it bigger, but it did not work (M, 73)."*

Minor UI problems included buttons appearing in the agent dialogue instead of the interaction area. All buttons that guide users through the application are supposed to appear at

the bottom of the screen, but some buttons were still located in the text screen of the agent. Furthermore, a minor UI problem that was found by one participant was a small typo.

4.3.2 UX Issues

Three out of 10 UX issues were critical. First, eight out of 10 participants did not know the purpose of certain modules. For instance, when being asked to put a failed fictional goal of eating dinner together in the system, some participants wanted to chat with the person they were supposed to have dinner with. Participant (M, 73): *"I would put in the chat that it did not succeed."* Moderator: *"With whom would you want to chat?"* Participant: *"With the person I was supposed to eat with"*. Arguably, this issue may be related to unclear use of words or confusing tasks. Comments by participants showed that assistance of the moderator made the goal and procedure of the modules clearer. Since it is not clear what the cause of the confusion exactly is, and to prevent confusion from happening with the final version of the application, this issue is addressed and will be tackled with the idea of 'Better safe than sorry' in mind. Second, five out of six people thought the words 'change a status' (een status veranderen) were confusing, which is underpinned by the following quotes. Participant (F, 73): *"Yes, change the status, because I want to indicate that I want to receive a notification every time."* Another participant (F, 86) mentions the following about the words 'change a status': *"Well, that sounds too complicated already*." Third, two out of six participants did not recognize the ability to scroll through the recipe list.

Five serious UX issues were identified. First, there was some confusion regarding the term 'Eating diary' (Eetboek). Three participants did not know what the eating diary comprised or thought it would be a recipe book. A participant (M, 72) said the following after receiving the task to look up a certain recipe: *"I have to prepare dinner and I think I have to press the button 'Eating diary'*." Second, five participants did not know what the 'Goal book' (Doelenboek) consisted of. *"After this task and with your help I understand what the goal*

book is about, but otherwise I would have had no clue what to do (F, 86).” Third, one participant thought the goal notification was sent too late as he would like to be notified the day prior to the day of the goal. Fourth, another serious UX issue found by one participant concerned the activities. He argued that some of the activities, that will be in the final version of the application, might be too far away for some people who are unable to travel. He suggested to build in a filter that lets people decide the maximum distance between them and the activity, and they only see relevant activities based on that.

Minor UX issues were the confusion regarding the word ‘Finish’ (Afronden). *“Afronden…? [Looks at moderators] I thought I had to choose (M, 71).”* In addition, recipes were not easy to find in the recipe list. This issue was addressed by one participant who had trouble finding the correct recipe due to the list not being in alphabetical order. Furthermore, a minor UX issue concerned the inconsistent use of informal (Jij, je, jouw) and formal salutation (U, uw). This issue was not specifically identified by participants, i.e. they were not obstructed in performing the tasks successfully and no comments about this issue were made. However, the post-task interviews revealed that one participant wanted to be formally saluted, which was the reason for the moderator to include this issue.

4.4 General impression

In general, people liked the interface of the PACO application and the underlying ideas of eating healthier and getting in touch with peers. Its simplicity and the fact that people are gradually directed through the app were positive remarks. Besides the tiny interface on a Samsung tablet, creating an account did not result in major issues and was straightforward. The recipe book did not result in any major remarks, other than the interest of one participant about nutrients and vitamins.

On the other hand, the eating diary triggered multiple comments. A distinction could be made between participants who do not see the value of keeping track of food intake and

those who do. Furthermore, some participants would like to receive visual information concerning macronutrients, while others did not. People like the idea of writing down goals and working on them. However, some participants did not see the value of keeping track of unsuccessful goals and are more focused on successfully fulfilling upcoming goals.

The stories module was welcomed with great satisfaction. The story about the Jordaanboot caused smiles on the faces of the participants. *"I like that, because when you hear an older voice, you saw it, I start to smile (F, 74)."* Another participant mentioned the following after listening to the story: *"Well, that makes you excited, doesn't it? (F, 72)"* Overall, participants believed hearing stories from peers and being able to sign up for activities positively contributed to the application.

5. Discussion

5.1 Principal findings

The aim of this research was to assess the usability of an ECA application for community-dwelling older adults and provide practical suggestions for ECA developers. PACO is perceived as usable and satisfactory, but still various usability issues have been identified. These issues may serve as a guideline for developers of ECA technologies.

As a first main finding, we found the overall usability of PACO to be high. The application was perceived as usable and the interface was clear and understandable. This finding is based on task completion rates and task completion times. These metrics showed that the tasks required little effort by the participants. Bickmore, Caruso & Clough-Gorr (2005) had a similar finding, since participants in their study also had very few problems using the system and communicating with the ECA. Literature on the usability of an ECA system in relation to task metrics is lacking, so a broader comparison is required to underpin the first main finding with literature. Harrati, Bouchrika, Tari and Ladjailia (2016) investigated e-learning systems using task metrics and found that completion rates and completion times play a role in expressing the degree of usability. Thus, the high completion rates and low completion times found in this study indicate a high degree of usability.

Furthermore, the SUS is used to assess the overall usability of PACO. With an average SUS score of 72.0, the ECA system is perceived as usable (Bangor et al., 2008). In comparison, systems in similar studies in this field scored lower than the PACO application on the SUS. A multi-device system called Kristina that monitored users' physical activity and medication intake scored 62.2 (Op Den Akker et al., 2016). A gamification platform (PERGAMON) that integrates educational gaming and coaching scored 50.2 (Klaassen, Bul, Op Den Akker, Van Der Burg, Kato & Di Bitonto, 2018), A web-based ECA scored 60.0

(Huff, Mack, Cummings, Womack, Gosha & Gilbert, 2019). Note that Huff et al. (2019) used the median to denote the SUS score.

In terms of technology maturation, PACO was in TRL 6 when conducting the usability evaluation. The ECA of Op Den Akker and colleagues (2016), Kristina, was in TRL 7, since their prototype was demonstrated in the expected operational environment: a field trial of seven weeks was conducted. The same holds for the gamified platform PERGAMON of Klaassen et al. (2018). They conducted a usability test in a hospital, which was the expected operational environment of the platform. Huff et al. (2019) did not specify in which environment their usability of an ECA took place, which makes it difficult to categorize their ECA application in terms of technology maturation. It is either in TRL 6 or 7, since a full prototype was tested. Thus, with regards to technological maturity at the time of the usability evaluations of the abovementioned technologies, PACO was one step behind in the process compared to two out of three previously mentioned ECAs, which is mainly due to the environment the evaluation took place in. The small difference in maturity does not explain why PACO scored higher on the SUS in comparison to similar ECA technologies. The higher score of PACO may be the result of the stakeholder-inclusive design approach of the ECAs, but further research is necessary to confirm this statement.

As a second main finding, participants perceived the PACO application as satisfactory to interact with. This finding is reflected in the task satisfaction scores, which were high for the tasks that measured satisfaction (task A and B). This finding is in line with a study of Bickmore et al. (2010), who reported high satisfaction among participants that communicated with an ECA. Yet, we can only mention the finding of high satisfaction levels of PACO, since is not clear what the exact cause of the satisfaction is. Follow-up studies would be needed to find out.

As a third main finding, various usability issues (and solutions) have been identified on a UI and UX level. These issues were identified based on the transcripts and screen recording. Solutions were provided to resolve the issues. Six user interface issues were found. First, the button to finish creating an account did not work. This issue was fixed during the evaluation days. Second, the interface was too small to read when typing a username and/or password on a Samsung tablet. This was resolved with a hack: if the screen is in landscape mode and the height decreases by 25%, it is assumed the keyboard is open. Third, participants sometimes feel the urge to click on buttons that are unclickable. This was solved by making buttons that were previously unclickable, clickable. Fourth, participants had trouble reading the recipe book. This issue was tackled by making the recipe book larger. Fifth, buttons that guide users through the application are supposed to be at the bottom of the screen, but some buttons were still located in the text screen of the agent. These buttons were moved to the right place. Sixth, there was a typo in the text, which required a minor fix.

In terms of user experience problems, 10 issues were found. First, participants did not seem to grasp the purpose of the modules, so providing a short description of each module after account creation was provided as solution. Second, participants did not recognize the ability to scroll through the recipe list. This was resolved by using a dialogue to let users know they can use the scrolling feature. Third, the issue of users receiving goal notifications too late was resolved by enabling users to decide when they want to be notified. Fourth, a user experience issue pointed out by a participant concerned the activities. He argued that some of the activities might be too far away for some people who are unable to travel. The solution was to enable users to sort activities in terms of postal code or radius of a certain number of kilometers. Fifth, participants perceived finding the correct recipe to be too hard. This was solved by making sure the recipe book is listed in alphabetical order. Sixth, a minor issue of inconsistent use of informal and formal salutation was solved by only using the formal

variant. Last, some words and phrases were confusing for the participant, such as ‘change a status’, ‘eating diary’, ‘goal book’ and ‘finish’ (in a specific context). The issue of confusing words was found by De Barros, Leitão and Ribeiro (2014) as well, who suggested that the wordings should correspond to the vocabulary of the target group. The confusing words in the PACO application were changed to better fit the vocabulary of the target group. All modifications were directly communicated to the development team and some alterations were implemented straight away.

5.2 Practical recommendations

Due to the relatively young field of research on ECAs, especially those created using a stakeholder-inclusive design approach, literature on the usability of such applications is lacking (Kramer et al., 2020). Moreover, research shows that involving end users and stakeholders in the development process has positive effects on usability (Van Velsen, Wentzel & Van Gemert-Pijnen, 2013). Therefore, this section provides practical insights for ECA application developers and designers, particularly for applications aimed at (community-dwelling) older adults.

This section contains practical suggestions for ECA developers based on the findings of this study. Most modifications suggested by participants or encountered usability issues are bound to a specific application, e.g. the malfunction of a certain button and suggestions to add a specific feature. However, some suggestions or issue fixes are universal, such as an easy to read font size, a clear understanding of the goal of the application, and words that are not confusing. I suggest avoiding technical language and jargon to avoid confusion of your target group. Furthermore, it is wise to investigate locations on the screen that users want to interact with, since they may feel the urge to click on areas that the application designers made unclickable, but make perfect sense to be made clickable. Also, keep minor issues such as

typos and inconsistencies in mind. These issues probably do not make-or-break an application, but it is better to prevent users from getting annoyed by small mistakes.

In addition, Zapata, Fernández-Alemán, Idri and Toval (2015) argue that research in this field is still growing and they emphasize the importance of using automated evaluation methods since 73% of the papers they reviewed used only interviews or questionnaires. This study used multiple evaluation methods that were commonly used in this field of research. Having a variety of different methods allowed us to gather information in different ways. The questionnaires, SUS and ASQ, and task metrics provided insights in the degree of usability and satisfaction. The think aloud method and post-test interview were useful to identify specific usability issues. Most studies use the SUS as a benchmark for usability evaluation despite being relatively outdated. Some authors argue that this scale is not suitable as a standalone tool to measure usability, since it lacks predictive power and does not take groups with physical or cognitive impairments into account, and should be combined with task metrics such as task completion (Harrati et al., 2016; Broekhuis, van Velsen & Hermens, 2019). Lewis (2018) argues that this scale is still relevant for researchers and practitioners to measure usability. Therefore, if there is no compelling reason to not use the SUS, it is suitable as a usability benchmark provided that it is combined with other methods.

When it comes to maturity of the application to conduct a usability evaluation, I suggest either level five, six or seven of the technology readiness level scale. These levels are categorized by demonstrating a total application that can be tested in a realistic environment (TRL 5), testing a full prototype of the system in a relevant environment (TRL 6) or testing a full prototype in the planned operational environment (TRL 7). Conducing the usability evaluation in an earlier stage results in components of the system being tested separately, which does not represent the integrated application that is being used after release. Moreover, when conducing the evaluation during one of the stages mentioned above, modifications can

be made before completing and launching the application. The system is close to finished and tested in an environment that represents the actual environment the system is going to be used in. This allows for conducting a reliable evaluation with deficiencies that may actually occur during the actual use of the system.

5.3 Development of eHealth applications

This paragraph will emphasize the difficulties of developing an eHealth technology. The PACO application is used as a real-life example to reflect on these challenges. It illustrates the importance of using a stakeholder-centered design approach for applications in the eHealth domain. Further, a reflection on the usability research field is made.

eHealth can create huge opportunities for patients, health care services and other actors, but developing an eHealth application entails multiple challenges. Such challenges include the creation of a new healthcare culture, achieving compatibility with legacy systems, and taking expectations and needs of different stakeholders that may have conflicting interests and objectives into account (Troshani & Wickramasinghe, 2014). This shows that developing and implementing eHealth technologies is a complex operation, both socially and technologically.

The complexity of the health care environment and the fact that current eHealth development often does not connect human characteristics, socioeconomic and cultural environments and technology to each other, results in a relatively low impact of eHealth technologies (Van Gemert-Pijnen et al., 2011). To improve the successfulness of eHealth technologies, Van Gemert-Pijnen and colleagues (2011) proposed a holistic eHealth development approach. Their interdisciplinary framework, the CeHRes Roadmap, was derived from principles of existing eHealth frameworks, insights from eHealth research and theories from other disciplines, such as psychology, communication and human-computer

interaction design. It contains five iterative phases that can be used as a guideline in planning, coordination and execution of the development process of eHealth technologies.

Despite the in-depth development strategy, Van Gemert-Pijnen et al. (2011) do not underestimate possible barriers that disturb the use of the full framework. The authors mention time, policy and financial considerations as possible barriers. The PACO project team experienced time to be a limiting factor in implementing certain parts of the application. For instance, the chat module was planned to be implemented, but due to limited time this turned out to be an issue and an alternative solution had to be found. Every project team has limited time and financial resources due to deadlines and limited funding. Besides, even stakeholders within a project team may have different interests. On the one hand, due to project management and resource planning, application developers have limited time and financial resources, while, on the other hand, behavioral scientists would like to implement and investigate certain theories and models as thoroughly as possible. The needs and wishes of both parties cannot always be fully satisfied, so an optimal division of resources and implemented content is required.

5.3.1 Strengths of PACO

When comparing the CeHRes Roadmap to the development process of PACO, multiple principles that laid a foundation for the CeHRes Roadmap can be distinguished. First, PACO recognizes the importance of stakeholder participation. The usability evaluation, which is an important part of the technology development process, involved potential end users. Furthermore, the principle of technology development being an iterative, dynamic process is respected. Insights generated during the usability evaluation are directly used to implement changes. A third principle is the inclusion of persuasive design techniques that PACO uses: embodied conversational agents. The focal point of PACO is to gain insights in the working mechanisms and persuasiveness of ECAs, not necessarily implementing and

launching the application. Therefore, implementation issues are not taken into account during the development process.

5.3.2 Reflection on usability research field

In my opinion, the usability field contains thoughtful methods of measurement. Usability questionnaires can be combined with, for example, the Think Aloud method and post-test interviews. However, in some cases the role of the moderator may be too prevalent. I can imagine that the outcomes of two identical series of usability evaluations are slightly different when changing moderators. In some usability evaluations the moderator is not present in the same room as the participant, while the participant's actions are recorded. I like the idea of ruling out as many influencing factors as possible to conduct research in the most optimal way. In usability testing, this is achieved by keeping the surroundings and script the same for each participant.

However, when researching humans, there are too many factors that may influence the outcomes and it is impossible to keep the circumstances constant all the time. The most important to keep in mind is the goal of the study, since, e.g. in the case of PACO, research is done to eventually improve the quality of life of older adults. One should not be fixated on conducting research in the most optimal way while disregarding why the research is conducted in the first place, for example, to improve an application that may have substantial health benefits for certain population groups.

In terms of choosing the right time to evaluate a system's usability, in my opinion, carrying out a usability evaluation is only useful when the application you want to test is close to finished. There is no point in testing the usability at an early phase of the development process, after which many changes must be made. On the other hand, testing the usability of a

finished application may result in certain modules needing an overhaul, which is a waste of time and money.

5.4 Limitations

Although the moderator tried to create an atmosphere in which the participants felt comfortable to criticize the system by explaining that constructive criticism is useful for the development of the application and that there were no right or wrong answers, a limitation of this study may have been the socially desirable answers given by the participants. While conducting the research I noticed some participants had trouble criticizing. One participant (F, 86) almost justified giving negative comments by saying "*But if I can be a bit critical… Of course I can say oh fantastic etcetera*…", but eventually proceeded by explaining a negative part of the application. Critique was still given, but almost with hesitation. Other participants may have refrained from giving honest opinions about the application, simply to seem polite. Possible socially desirable answers of participants may have resulted in a higher SUS score, higher satisfaction scores, and less usability issues.

A second limitation might be related to the research panel the participants were recruited from. People who like technology might be more willing to participate in a study like this, and they are arguably more experienced with applications like PACO. Some of them had participated in similar studies before, including testing the usability of a different ECA. These people may not correctly represent the target group, which may have resulted in a distortion of the analysis.

5.5 Conclusion, relevance and future work

In conclusion, PACO is perceived to be a usable ECA application that is satisfactory to interact with. The overall SUS score of the ECA application was good and task satisfaction

scores were high. Various usability issues have been identified on a User Interface and User Experience level that may prove to be useful for ECA designers.

Due to the lack of human-centered, stakeholder-inclusive design approaches of ECA applications, usability evaluations of agents designed in this specific way are scarce. This study provided insights into the usability of an ECA technology aimed at community-dwelling older adults that was designed with the input of end users. Besides, it sheds a light on several points of attention for application developers of ECA technologies that can be implemented to create a product that improves the alignment of the application characteristics and users' wishes.

Future work should investigate the usability of similar ECA applications to increase the knowledge on this topic, since embodied conversational agents are an effective modality in health communication and health behavior change interventions in community-dwelling older adults. This study had a one-sided approach with regards to measuring task satisfaction, since scores were measured without investigating which part of the system contributed to that score. Further research should investigate this to elicit more helpful insights. This allows for more in-depth recommendations for ECA technology developers.

References

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction*, 24(6), 574-594.

Baumel, A., & Muench, F. (2016). Heuristic evaluation of eHealth interventions: establishing standards that relate to the therapeutic process perspective. *JMIR mental health*, 3(1), e5.

Bickmore, T. W., Caruso, L., Clough-Gorr, K., & Heeren, T. (2005). 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers*, 17(6), 711-735.

Bickmore, T. W., Caruso, L., & Clough-Gorr, K. (2005). Acceptance and usability of a relational agent interface by urban older adults. In *CHI'05 extended abstracts on Human factors in computing system*s (pp. 1212-1215).

Bickmore, T. W. (2010). Relational agents for chronic disease self management. *Health Informatics: A Patient-Centered Approach to Diabetes*, 181-204.

Bickmore, T. W., Pfeifer, L. M., Byron, D., Forsythe, S., Henault, L. E., Jack, B. W., Silliman, R., & Paasche-Orlow, M. K. (2010). Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. *Journal of health communication*, 15(S2), 197-210.

Bonnet, F., Irving, K., Terra, J. L., Nony, P., Berthezène, F., & Moulin, P. (2005). Anxiety and depression are associated with unhealthy lifestyle in patients at risk of cardiovascular disease. *Atherosclerosis*, 178(2), 339-344.

Branch, J. L. (2000). Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters. *Library & Information Science Research*, 22(4), 371-392.

Broekhuis, M., van Velsen, L., & Hermens, H. (2019). Assessing usability of eHealth technology: A comparison of usability benchmarking instruments. *International journal of medical informatics*, 128, 24-31.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.

Brummett, B. H., Barefoot, J. C., Siegler, I. C., Clapp-Channing, N. E., Lytle, B. L., Bosworth, H. B., Williams, R. B., & Mark, D. B. (2001). Characteristics of socially isolated patients with coronary artery disease who are at elevated risk for mortality. *Psychosomatic medicine*, 63(2), 267-272.

Bruun, A., & Stage, J. (2015). An empirical study of the effects of three think-aloud protocols on identification of usability problems. In *IFIP Conference on Human-Computer Interaction* (pp. 159-176). Springer, Cham.

Corrao, N. J., Robinson, A. G., Swiernik, M. A., & Naeim, A. (2010). Importance of testing for usability when selecting and implementing an electronic health or medical record system. *Journal of Oncology Practice*, 6(3), 120-124.

De Barros, A. C., Leitão, R., & Ribeiro, J. (2014). Design and evaluation of a mobile user interface for older adults: navigation, interaction and visual design recommendations. *Procedia Computer Science*, 27(27), 369-378.

DiMatteo, M. R. (2004). Social support and patient adherence to medical treatment: a meta-analysis. *Health psychology*, 23(2), 207.

Duh, H. B. L., Tan, G. C., & Chen, V. H. H. (2006). Usability evaluation for mobile device: a comparison of laboratory and field tests. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services* (pp. 181-186).

Dumas, J. S., & Redish, J. (1999). A practical guide to usability testing. Intellect books.

Ford, D. W., Hartman, T. J., Still, C., Wood, C., Mitchell, D. C., Erickson, P., Bailey, R., Smiciklas-Wright, H., Coffman, D. L., & Jensen, G. L. (2014). Body mass index, poor diet quality, and health-related quality of life are associated with mortality in rural older adults. *Journal of nutrition in gerontology and geriatrics*, 33(1), 23-34.

Harrati, N., Bouchrika, I., Tari, A., & Ladjailia, A. (2016). Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis. *Computers in Human Behavior*, 61, 463-471.

Hartmann, B., Mancini, M., Buisine, S., & Pelachaud, C. (2005, July). Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems* (pp. 1095-1096). ACM.

Heinrichs, M., Baumgartner, T., Kirschbaum, C., & Ehlert, U. (2003). Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress. *Biological psychiatry*, 54(12), 1389-1398.

Heo, J., Ham, D. H., Park, S., Song, C., & Yoon, W. C. (2009). A framework for evaluating the usability of mobile phones based on multi-level, hierarchical model of usability factors. *Interacting with Computers*, 21(4), 263-275.

Hetherington, M. M. (1998). Taste and appetite regulation in the elderly. *Proceedings of the Nutrition Society*, 57(4), 625-631.

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM*, 48(1), 71-74.

Horsky, J., McColgan, K., Pang, J. E., Melnikas, A. J., Linder, J. A., Schnipper, J. L., & Middleton, B. (2010). Complementary methods of system usability evaluation: surveys and observations during software design and development cycles. *Journal of biomedical informatics*, 43(5), 782-790.

House, J. S., Landis, K. R., & Umberson, D. (1988). Social relationships and health. *Science*, 241(4865), 540-545.

Huff, E. W., Mack, N. A., Cummings, R., Womack, K., Gosha, K., & Gilbert, J. E. (2019). Evaluating the Usability of Pervasive Conversational User Interfaces for Virtual Mentoring. In *International Conference on Human-Computer Interaction* (pp. 80-98). Springer, Cham.

Hughes, G., Bennett, K. M., & Hetherington, M. M. (2004). Old and alone: barriers to healthy eating in older men living on their own. *Appetite*, 43(3), 269-276.

Jaspers, M. W. (2009). A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. *International journal of medical informatics*, 78(5), 340-353.

Jaspers, M. W., Steen, T., Van Den Bos, C., & Geenen, M. (2004). The think aloud method: a guide to user interface design. *International journal of medical informatics*, 73(11-12), 781-795.

Kampmeijer, R., Pavlova, M., Tambor, M., Golinowska, S., & Groot, W. (2016). The use of e-health and m-health tools in health promotion and primary prevention among older adults: a systematic literature review. *BMC health services research*, 16(5), 290.

Kayser, L., Kushniruk, A., Osborne, R. H., Norgaard, O., & Turner, P. (2015). Enhancing the effectiveness of consumer-focused health information technology systems through eHealth literacy: a framework for understanding users' needs. *JMIR human factors*, 2(1), e9.

Kientz, J. A., Choe, E. K., Birch, B., Maharaj, R., Fonville, A., Glasson, C., & Mundt, J. (2010). Heuristic evaluation of persuasive health technologies. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 555-564).

Kitzinger, J. (1995). Qualitative research: introducing focus groups. *Bmj*, *311*(7000), 299-302.

Klaassen, B., van Beijnum, B. J., & Hermens, H. J. (2016). Usability in telemedicine systems—A literature survey. *International journal of medical informatics*, 93, 57-69.

Klaassen, R., Bul, K., Op den Akker, R., Van der Burg, G. J., Kato, P. M., & Di Bitonto, P. (2018). Design and evaluation of a pervasive coaching and gamification platform for young diabetes patients. *Sensors*, 18(2), 402.

Kramer, L., Ter Stal, S., Mulder, B., De Vet, E., Van Velsen, L. (2020). Developing Embodied Conversational Agents for Coaching People in a Healthy Lifestyle: Scoping Review. *Journal of Medical Internet Research*, 22(2), e14058.

Kujala, S. (2003). User involvement: a review of the benefits and challenges. *Behaviour & information technology*, 22(1), 1-16.

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. *ACM Sigchi Bulletin*, 23(1), 78-81.

Lewis, J. R. (2018). The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, 34(7), 577-590.

Liu, Y. H., Gao, X., Mitchell, D. C., Wood, G. C., Still, C. D., & Jensen, G. L. (2019). Diet Quality Is Associated With Mortality in Adults Aged 80 Years and Older: A Prospective Study. *Journal of the American Geriatrics Society*.

Loveys, K., Fricchione, G., Kolappa, K., Sagar, M., & Broadbent, E. (2019). Reducing Patient Loneliness With Artificial Agents: Design Insights From Evolutionary Neuropsychiatry. *Journal of medical Internet research*, 21(7), e13664.

Luo, Y., Hawkley, L. C., Waite, L. J., & Cacioppo, J. T. (2012). Loneliness, health, and mortality in old age: A national longitudinal study. *Social science & medicine*, 74(6), 907-914.

Mankins, J. C. (1995). Technology readiness levels. *White Paper*.

Mankins, J. C. (2009). Technology readiness assessments: A retrospective. *Acta Astronautica*, 65(9-10), 1216-1223.

Maramba, I., Chatterjee, A., & Newman, C. (2019). Methods of usability testing in the development of eHealth applications: A scoping review. *International journal of medical informatics*. 126(March), 95–104.

Monkman, H., & Kushniruk, A. (2013a). Applying usability methods to identify health literacy issues: An example using a personal health record. *Studies in Health Technology and Informatics*, 183, 179–185.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249-256).

Nielsen, J. (2005). Ten usability heuristics.

Niranjanamurthy, M., Nagaraj, A., Gattu, H., & Shetty, P. K. (2014). Research Study on Importance of Usability Testing/User Experience (UX) Testing. *International Journal of Computer Science and Mobile Computing*, IJCSMC, 3(10), 78-85.

Ocké, M. C., Buurma-Rethans, E. J. M., De Boer, E. J., Wilson-Van den Hooven, C., Etemad-Ghameslou, Z., Drijvers, J. J. M. M., & Van Rossum, C. T. M. (2013). Diet of community-dwelling older adults: Dutch national food consumption survey older adults 2010-2012.

Op den Akker, H. J., Klaassen, R., & Nijholt, A. (2016). Virtual coaches for healthy lifestyle. In *Toward Robotic Socially Believable Behaving Systems*-Volume II (pp. 121-149). Springer, Cham.

Page, T. (2014). Touchscreen mobile devices and older adults: a usability study. International *Journal of Human Factors and Ergonomics*, 3(1), 65-85.

Reedy, J., Krebs-Smith, S. M., Miller, P. E., Liese, A. D., Kahle, L. L., Park, Y., & Subar, A. F. (2014). Higher diet quality is associated with decreased risk of all-cause, cardiovascular disease, and cancer mortality among older adults. *The Journal of nutrition*, 144(6), 881-889.

Ring, L., Shi, L., Totzke, K., & Bickmore, T. (2015). Social support agents for older adults: longitudinal affective computing in the home. *Journal on Multimodal User Interfaces*, 9(1), 79-88.

Robinson, S. M. (2018). Improving nutrition to support healthy ageing: what are the opportunities for intervention?. In *Proceedings of the Nutrition Society*, 77(3), 257-264.

Rubin, J., & Chisnell, D. (2008). How to plan, design, and conduct effective tests. *Handbook of usability testing*, 348.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & cognition*, 17(6), 759-769.

Sauro, J., Lewis, J.R. (2012). Quantifying The User Experience. *Practical Statistics for User Research*.

Shaw, T., McGregor, D., Brunner, M., Keep, M., Janssen, A., & Barnet, S. (2017). What is eHealth (6)? Development of a conceptual model for eHealth: qualitative study with key informants. *Journal of medical Internet research*, 19(10), e324.

Teague, R., De Jesus, K., & Ueno, M. N. (2001). Concurrent vs. post-task usability test ratings. *CHI'01 extended abstracts on Human factors in computing systems*, 289-290.

Thompson, D. R. (1989). A randomized controlled trial of in-hospital nursing support for first time myocardial infarction patients and their partners: effects on anxiety and depression. *Journal of Advanced Nursing*, 14(4), 291-297.

Tory, M., & Moller, T. (2005). Evaluating visualizations: do expert reviews work?. *IEEE computer graphics and applications*, 25(5), 8-11.

Troshani, I., & Wickramasinghe, N. (2014). Tackling complexity in e-health with actor-network theory. In *2014 47th Hawaii International Conference on System Sciences* (pp. 2994-3003). IEEE.

Van Den Haak, M. J., & De Jong, M. D. (2003). Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols. In IEEE International Professional Communication Conference, 2003. IPCC 2003. *Proceedings*. (pp. 3-pp). IEEE.

Van Den Haak, M., De Jong, M., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology*, 22(5), 339-351.

Van Gemert-Pijnen, J. E., Nijland, N., van Limburg, M., Ossebaard, H. C., Kelders, S. M., Eysenbach, G., & Seydel, E. R. (2011). A holistic framework to improve the uptake and impact of eHealth technologies. *Journal of medical Internet research*, 13(4), e111.

Van Velsen, L., Wentzel, J., & Van Gemert-Pijnen, J. E. (2013). Designing eHealth that matters via a multidisciplinary requirements development approach. *JMIR research protocols*, 2(1), e21.

West, J. H., Belvedere, L. M., Andreasen, R., Frandsen, C., Hall, P. C., & Crookston, B. T. (2017). Controlling your "App" etite: how diet and nutrition-related mobile apps lead to behavior change. *JMIR mHealth and uHealth*, 5(7), e95.

World Health Organization, 2017. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed September 11, 2019.

Yin, Z., Fei, Z., Qiu, C., Brasher, M. S., Kraus, V. B., Zhao, W., Shi, X., & Zeng, Y. (2017). Dietary diversity and cognitive function among elderly people: A population-based study. *The journal of nutrition, health & aging*, 21(10), 1089-1094.

Zapata, B. C., Fernández-Alemán, J. L., Idri, A., & Toval, A. (2015). Empirical studies on usability of mHealth apps: a systematic literature review. *Journal of medical systems*, 39(2), 1.

Appendix A: Information letter and consent form

Beste heer/mevrouw,

Leuk dat u interesse heeft om mee te werken aan het onderzoek PACO. Voordat u de keuze maakt om wel of niet deel te nemen, geven wij u graag eerst wat meer informatie. Mocht u daarna besluiten om mee te doen, dan vragen wij u om het toestemmingsformulier op de laatste pagina te ondertekenen.

Wat is het doel van het onderzoek?

Dit onderzoek is onderdeel van het project PACO. In dit project ontwikkelt het Nationaal Ouderenfonds samen met anderen een nieuwe technologie om gezond eten te ondersteunen. In dit onderzoek wordt er gekeken hoe deze technologie verbeterd kan worden, en daar hebben we uw hulp bij nodig.

Wie wordt gevraagd om mee te doen?

Ouderen die thuis wonen en gepensioneerd zijn vragen wij om mee te doen aan dit onderzoek. Omdat het onderzoek op een tablet uitgevoerd gaat worden, is het fijn als u een vrij algemeen idee heeft hoe een tablet werkt.

Wat houdt het onderzoek in?

In het onderzoek krijgt u 5 taken die uitgevoerd gaan worden op de tablet. Tijdens die taken vragen wij u om uw gedachtes en acties hardop uit te spreken. Na elke taak krijgt u 3 korte vragen op papier. Aan het einde krijgt u een langere vragenlijst bestaande uit X vragen. In totaal duurt het onderzoek ongeveer 45 minuten.

Wat zijn de voor- en nadelen als u deelneemt?

Met dit onderzoek draagt u bij aan de ontwikkeling van een nieuwe gebruiksvriendelijke technologie voor senioren en aan onderzoek op een belangrijk thema: gezond leven.

Wat gebeurt er als u niet deel wilt nemen?

Deelname is geheel vrijwillig. U beslist zelf of u mee wilt doen aan het onderzoek. Als u besluit om niet mee te doen, zal er niks gebeuren. U hoeft ook niet te zeggen waarom u niet mee wilt doen. Als u wel mee wilt doen, kunt u zich op ieder moment bedenken en toch stoppen.

Wat gebeurt er met uw gegevens?

Het Nationaal Ouderenfonds en Wageningen Universiteit vinden het heel belangrijk om zorgvuldig om te gaan met uw persoonlijke gegevens. We zullen uw gegevens dan ook vertrouwelijk behandelen en deze anoniem maken; uw naam zal dus niet in een verslag komen. Uw gegevens zullen uitsluitend worden gebruikt binnen het project PACO en publicaties die hieruit voortkomen. U heeft tevens de mogelijkheid om deze publicaties per post of e-mail te ontvangen.

Wilt u verder nog iets weten?

Voor vragen over het onderzoek kunt u contact opnemen met Lean Kramer, onderzoeker bij Wageningen Universiteit, email lean.kramer@wur.nl.

Toestemmingsformulier PACO

- Ik bevestig dat ik de informatiebrief heb gelezen. Ik begrijp de informatie. Ik heb de gelegenheid gehad om aanvullende vragen te stellen. Ik heb voldoende tijd gehad om over deelname na te denken.

- Ik weet dat mijn deelname geheel vrijwillig is en dat ik mijn toestemming op ieder moment kan intrekken zonder dat ik daarvoor een reden hoef te geven.

- Ik geef toestemming voor het gebruik van de gegevens die ik verstrek tijdens het onderzoek.

- Na afloop van het onderzoek wil ik **wel/geen** (doorhalen wat niet van toepassing is) kopie ontvangen van de publicatie per **e-mail/post** (doorhalen wat niet van toepassing is).

Ik stem in met mijn deelname aan bovengenoemd onderzoek

Naam: _____

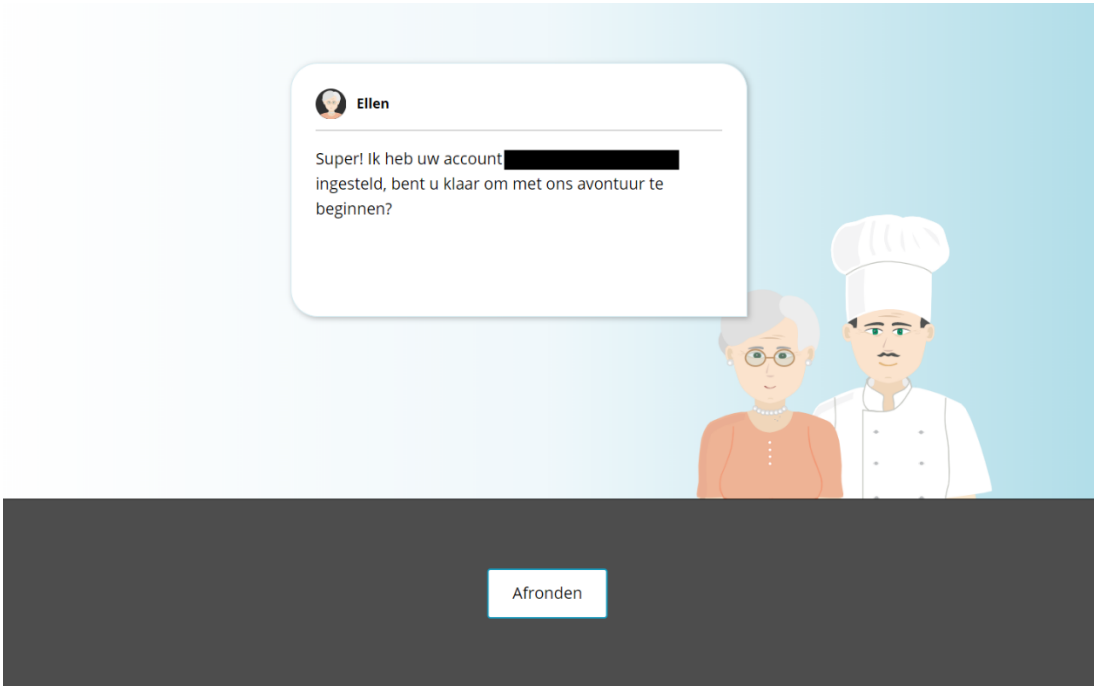
Handtekening: _____Datum: _____ /_____ /_____

Appendix B: Determination of task start and -end

Appendix B.1: Account creation

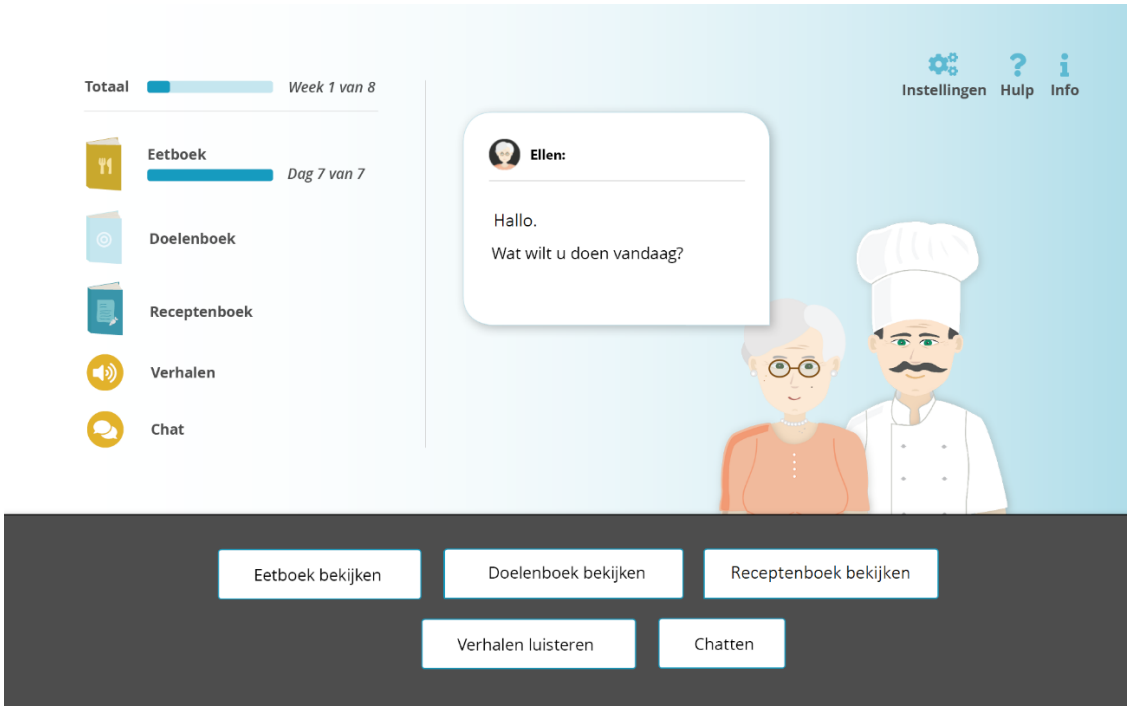
Print screen at start task A: Task starts on the home page (https://portals.rrdweb.nl/paco/).

Print screen at end task A: Task ends when the participant clicks on the finish (Afronden) button.

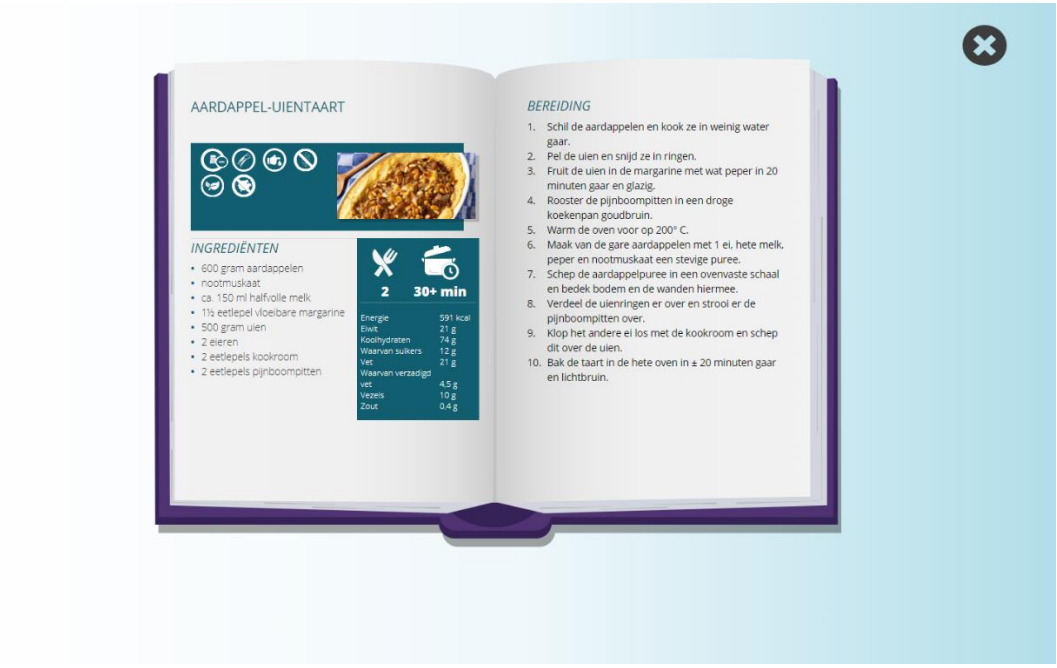


Appendix B.2: Recipe book

Print screen at start task B: Task starts on the home page (logged in).

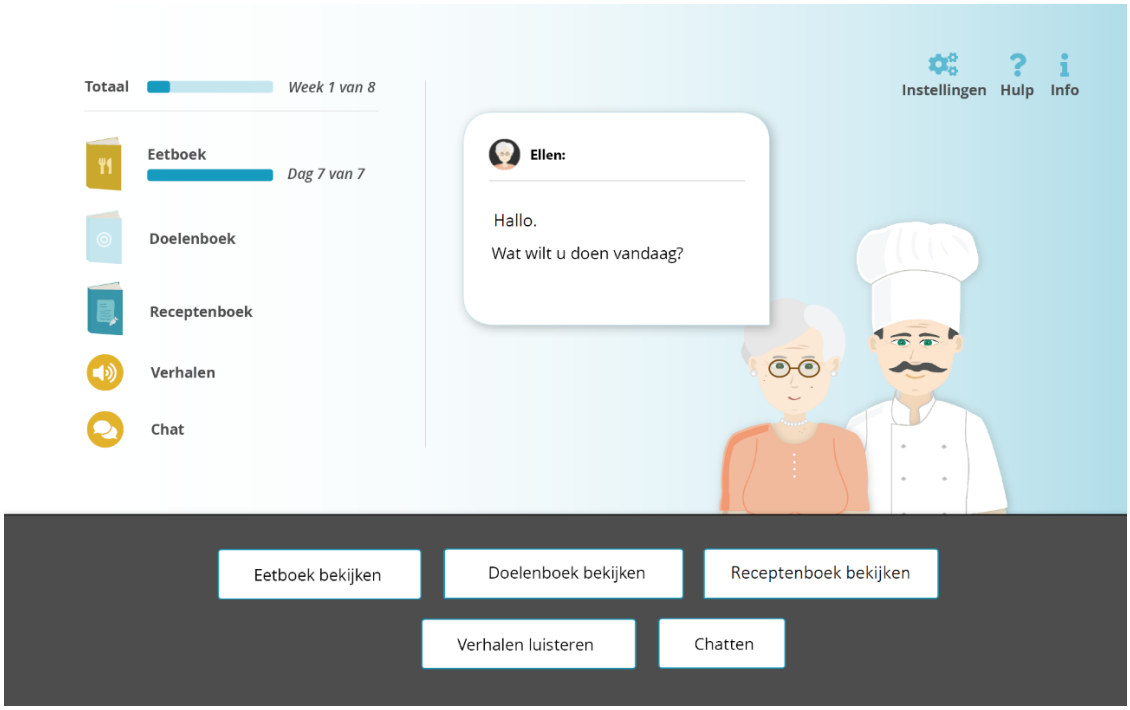


Print screen at end task B: Task ends when the participant indicates a certain number of grams required.

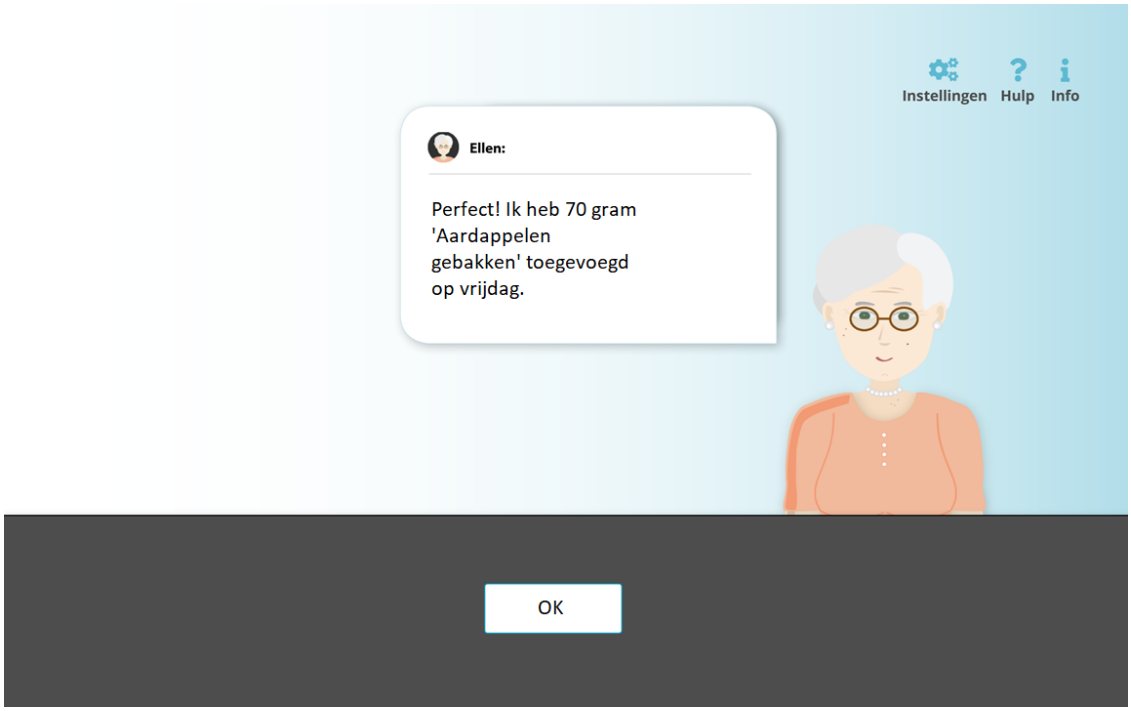


Appendix B.3: Eating diary

Print screen at start task C: Task starts on the home page (logged in).

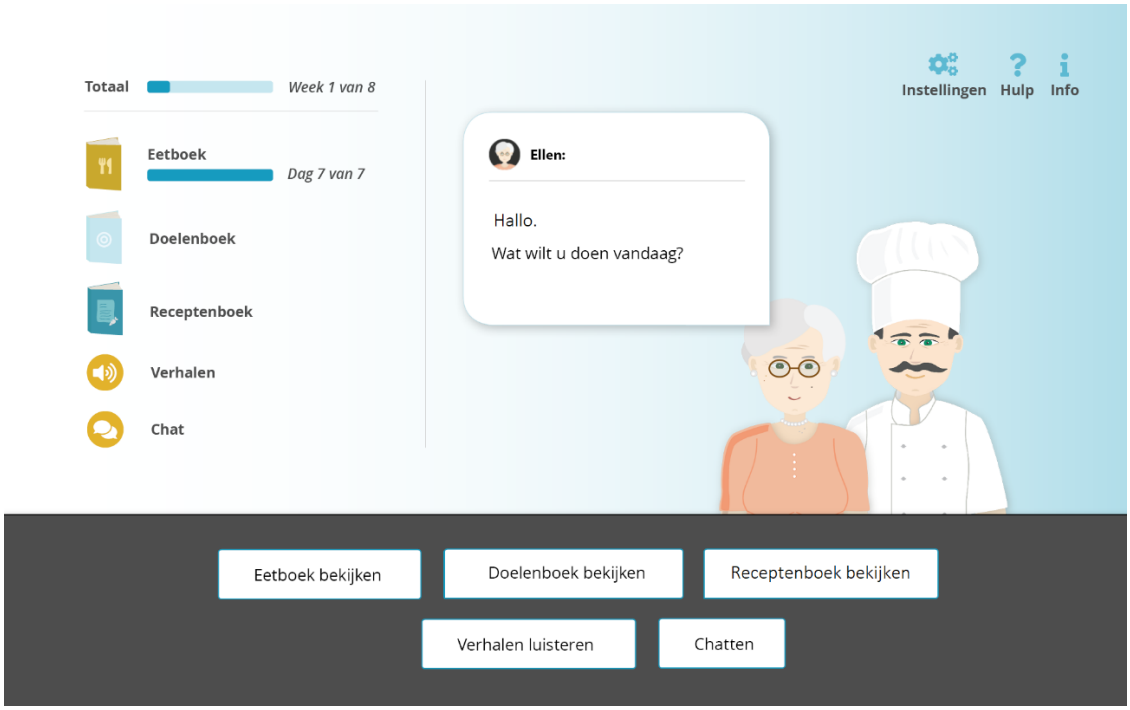


Print screen at end task C: Task ends when the participant clicks on the OK button.

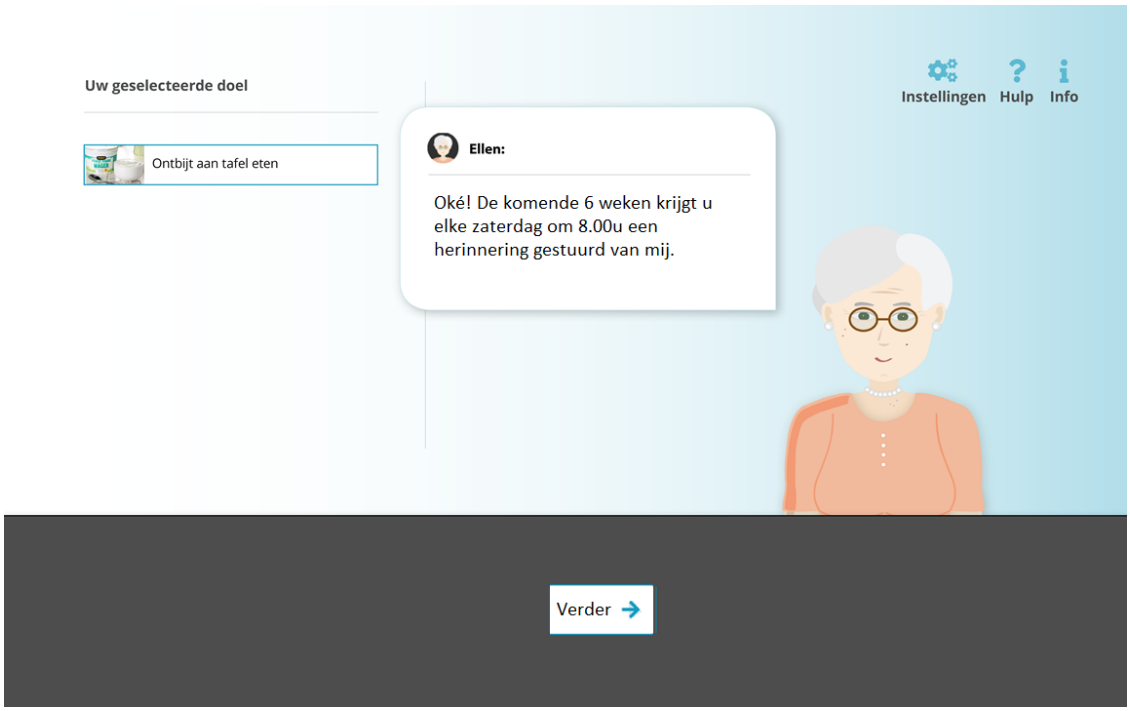


Appendix B.4: Goal book

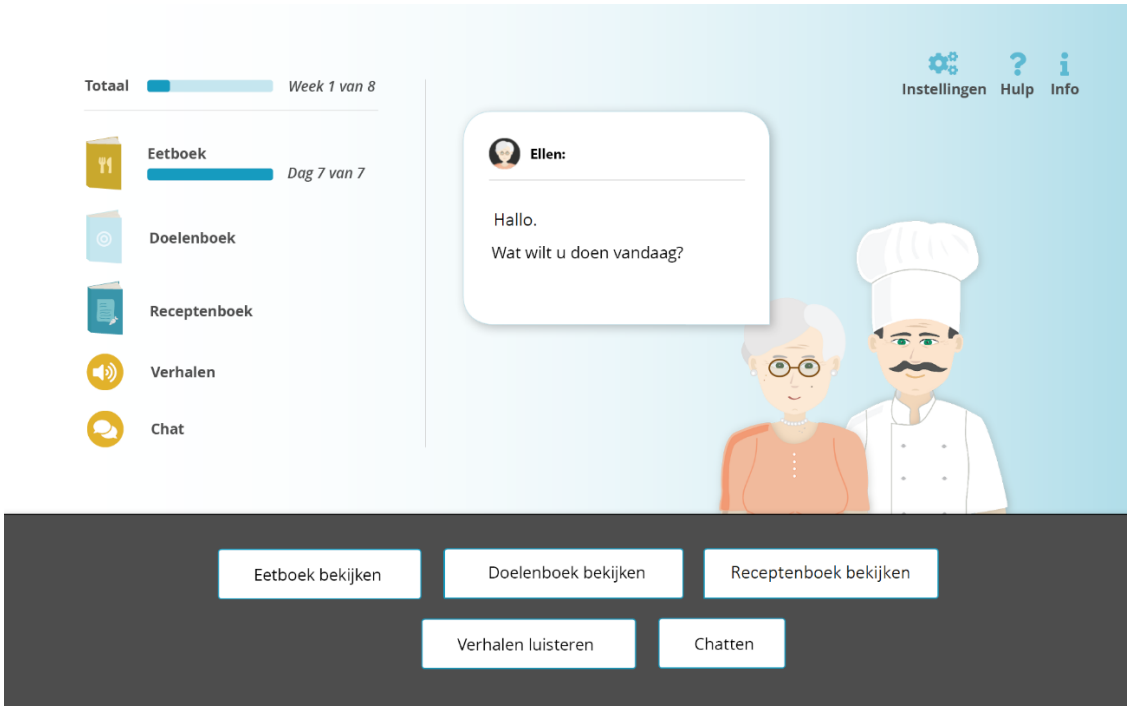
Print screen at start task D.1: Task starts on the home page (logged in).



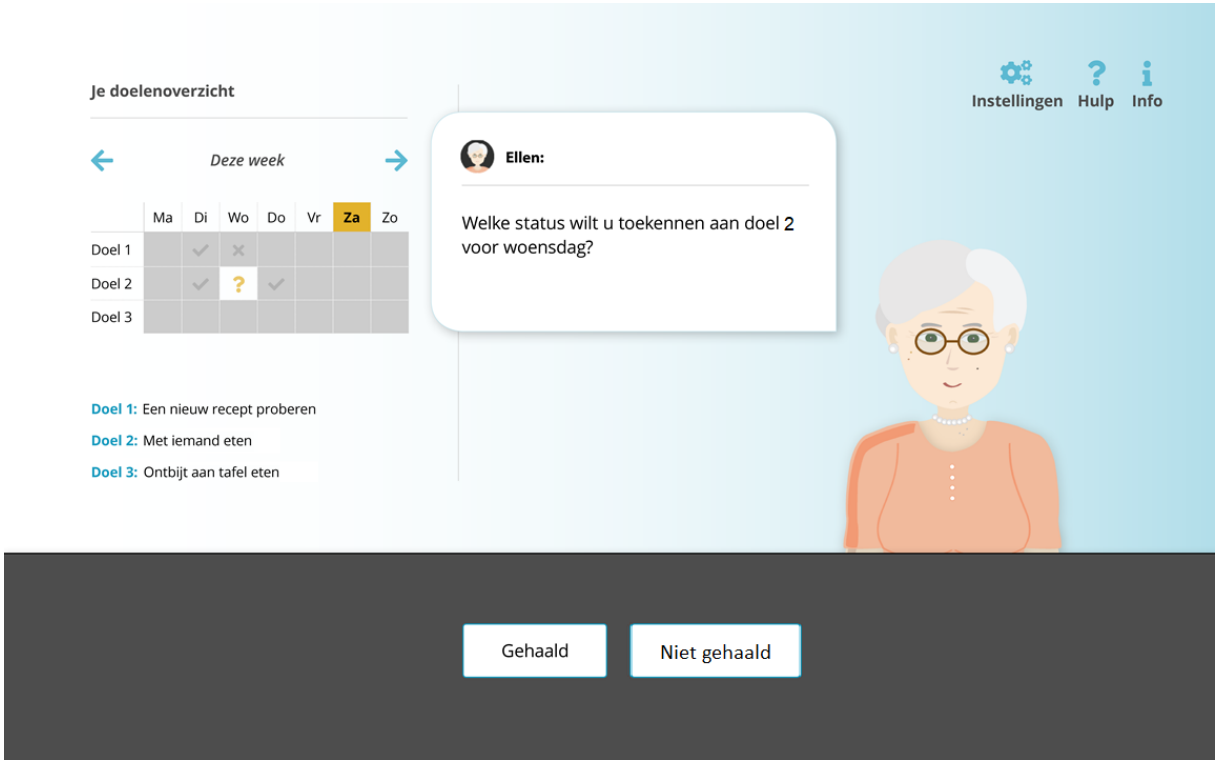
Print screen at end task D.1: Task ends when the participant clicks on the continue (Verder) button.



Print screen at start task D.2: Task starts on the home page (logged in).

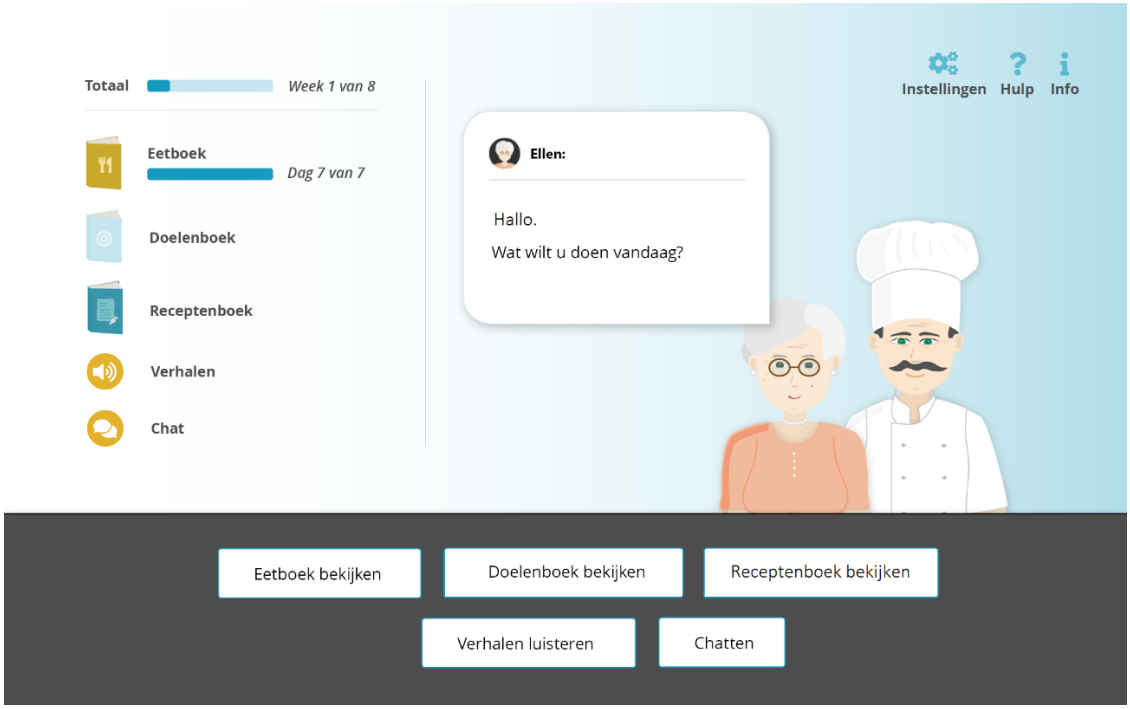


Print screen at end task D.1: Task ends when the participant clicks on the 'Not achieved' (Niet gehaald) button.

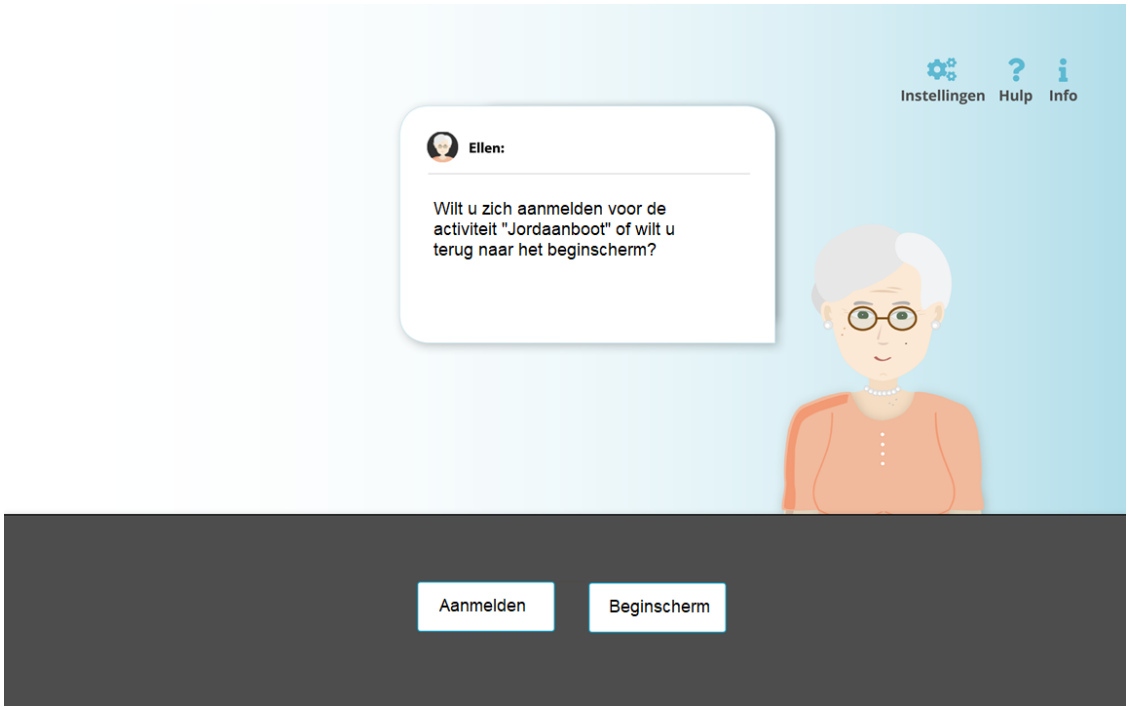


Appendix B.5: Stories

Print screen at start task E: Task starts on the home page (logged in).



Print screen at end task E: Task ends when the participant clicks on the ‘Sign up’ (Aanmelden) button.



Appendix C: Task- descriptions and metrics

Task A – Account creation

Task description: ***Create an account****. (****Maak een account aan****).*

Task completed? (y/n): ……..

Time of completion: …….. min …….. sec

Task B – Recipe book

Task description: *You want to prepare dinner. You still need a tasty recipe for the main course. Today, you choose a vegetarian meal with potatoes.* ***Look up how many grams of potatoes you need to make a potato-onion pie. You may mention it out loud.***
(U wilt het avondeten voorbereiden. U heeft nog een lekker recept nodig voor het hoofdgerecht. Vandaag kiest u voor een vegetarisch recept met aardappelen. ***Bekijk hoeveel gram aardappelen u nodig heeft voor het maken van aardappel-uientaart. U mag het hardop zeggen.****)*

Task completed? (y/n): ……..

Time of completion: …….. min …….. sec

Task C – Eating diary

Task description: *On Friday at 18.45h, you have eaten 70 grams of 'baked potatoes'.* ***Put this in the system.*** *(U heeft op vrijdag om 18.45u 's avonds 70 gram 'aardappelen gebakken gegeten.* ***Zet dit in het systeem.****)*

Task D – Goal diary

Task description D.1: *You want to create a new goal to eat breakfast at the table the next six weeks on Saturdays. You also want to receive a notification about this every week.* ***Put this in***

the system. *(U wilt een nieuw doel aanmaken, namelijk de komende zes weken op zaterdag uw ontbijt aan tafel eten. Hier wilt u ook elke week een herinnering over ontvangen.* ***Zet dit in het systeem.****)*

Task description D.2: *Last Wednesday, your goal was to eat together with somebody, but this failed.* ***Put this in the system.*** *(U had afgelopen woensdag als doel om met iemand te eten, maar dit is niet gelukt.* ***Zet dit in het systeem.****)*

Task E – Stories

Task description: *You feel like doing a fun activity.* ***Sign up for the Jordaanboot, after listening to the story about this activity.*** *(U heeft zin in een gezellige activiteit.* ***Meld u aan voor de Jordaanboot, nadat u het verhaal over deze activiteit heeft geluisterd.****)*

Appendix D: After-Scenario Questionnaire

	Helemaal niet mee eens						**Helemaal mee eens**
1. Deze taak uitvoeren was makkelijk voor mij.	☐	☐	☐	☐	☐	☐	☐
	1	2	3	4	5	6	7
2. Het kostte me weinig tijd om deze taak uit te voeren.	☐	☐	☐	☐	☐	☐	☐
	1	2	3	4	5	6	7
3. De PACO applicatie hielp me genoeg om de taak te voltooien.	☐	☐	☐	☐	☐	☐	☐
	1	2	3	4	5	6	7

Appendix E: System Usability Scale (SUS)

Participant ID: ……

1. Ik denk dat ik deze applicatie vaak zal gebruiken.

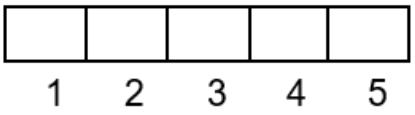
2. Ik vind de applicatie onnodig complex.

3. Ik vind de applicatie makkelijk te gebruiken.

4. Ik denk dat ik hulp van een technisch persoon nodig heb om deze applicatie te kunnen gebruiken.

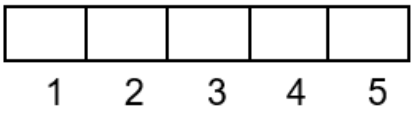
5. Ik vind de verschillende functies in deze applicatie goed geïntegreerd.

6. Ik vind dat er te veel inconsistentie in de applicatie zit.

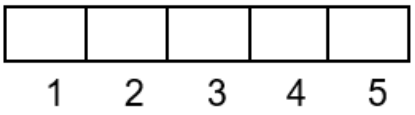


7. Ik kan me voorstellen dat de meeste mensen snel door hebben hoe ze de applicatie moeten gebruiken.

8. Ik vond de applicatie erg omslachtig te gebruiken.



9. Ik voelde me zelfverzekerd toen ik deze applicatie gebruikte.



10. Ik moest veel leren over deze applicatie voordat ik het goed kon gebruiken.

Appendix F: Demographics Questionnaire

Vragenlijst: Demografieken

In dit onderdeel stellen we een aantal vragen over uzelf. Kunt u de volgende vragen invullen?

1. Wat is uw geslacht?

□ Man

□ Vrouw

2. Wat is uw geboortedatum?

DD / MM / JJJJ

..... / ..... / .....

3. Wat is uw hoogst afgeronde opleiding?

□ Basisonderwijs

□ Middelbaar beroepsonderwijs (MBO)

□ Hoger beroepsonderwijs (HBO)

□ Wetenschappelijk onderwijs (WO)

4. Wat is uw woonsituatie?

□ Ik woon samen met mijn echtgenoot/echtgenote

□ Ik woon samen met een vriend/familielid/anders

□ Ik woon alleen

5. Kunt u alle apparaten aankruisen die u thuis gebruikt? U kunt meerdere antwoorden aankruisen.

□ Smartphone

□ PC / laptop

□ Tablet

□ Smartwatch

□ Game computer

□ Anders, namelijk: .....

Appendix G: Usability Evaluation Protocol

Nr	Fase	Activiteit	Uitleg	Materialen	Tijd
1	Introductie	Welkom	Welkom en uitleg studie. “Heel fijn dat u mee wilt doen. Heeft u al meegedaan aan een onderzoek van PACO?” “PACO is een project waarin we onderzoek doen naar het ontwikkelen van virtuele gezondheidscoaches. En om de applicatie waarin deze coaches verwerkt zitten zo goed mogelijk te ontwikkelen, leggen we een eerste versie voor aan mogelijke gebruikers. Op die manier krijgen we inzicht in hoe de applicatie verbeterd kan worden. En dat gaan we vandaag doen.”	Koffie / thee	5
		Krijg toestemming voor gebruik van gegevens en audio-opname	Deelnemer vult toestemmingsformulier in en geeft toestemming voor de audio-opname.	Informatiebrief & toestemmings-formulier	3
2	Interactie met PACO Concurrent Think Aloud &	Introductie en uitleg/oefenen Concurrent Think Aloud	“De volgende 20 minuten zult u 3 taken krijgen die u gaat uitvoeren. Tijdens deze taken spreekt u uit wat u denkt en doet, en dat gaan we zo even oefenen zodat u hier	Tablet Schermopname software Audio recorder	5

	vragen-lijsten		comfortabel mee wordt. Bovendien mag u tijdens de taken alles zeggen wat u wilt. Dan gaan we nu even hardop denken oefenen door de route te zoeken van Enschede naar Amsterdam Centraal, terwijl u dus hardop nadenkt." Einde introductie: moderator zet video- en audio-opname aan.		
		Taak A: Account aanmaken	"**Maak een account aan**." Taak begint op de hoofdpagina (https://portals.rrdweb.nl/paco/) (niet ingelogd). De deelnemer registreert met zijn/haar eigen e-mailadres en wachtwoord. Taak eindigt als de deelnemer op 'Afronden' klikt en terecht komt op het hoofdscherm. Moderator: houdt tijd bij. Moderator vult in of deelnemer de taak heeft afgerond binnen de tijdslimiet of niet.	Stopwatch Formulier start en eind voor taak A (appendix B.1) Taak-beschrijving en metrics voor taak A (appendix C)	5

			Als de taak niet is afgerond binnen 5 minuten wordt de taak stopgezet door de moderator.		
			Deelnemer vult after-scenario questionnaire in voor taak A.	After-Scenario Questionnaire (appendix D)	2
		Taak B: Receptenboek	"U wilt het avondeten voorbereiden. U heeft nog een lekker recept nodig voor het hoofdgerecht. Vandaag kiest u voor een vegetarisch recept met aardappelen. **Bekijk hoeveel gram aardappelen u nodig heeft voor het maken van aardappel-uientaart. U mag het hardop zeggen**." Taak begint op hoofdscherm (ingelogd). Moderator vraagt hoe de deelnemer deze taak zou aanpakken. Taak eindigt wanneer de deelnemer een aantal gram noemt. Moderator: houdt tijd bij. Als de taak niet is afgerond binnen 5 minuten wordt de taak stopgezet door de moderator.	Stopwatch Formulier start en eind voor taak B (appendix B.2) Taak-beschrijving en metrics voor taak B (appendix C)	5

			Deelnemer vult after-scenario questionnaire in voor taak B.	After-Scenario Questionnaire (appendix D)	2
		Uitleg ‘PowerPoint taken’	Voor de volgende taak gaan we het iets anders aanpakken. Hierbij werken de knoppen nog niet, maar toch willen we van u weten hoe u de taak zou uitvoeren. Dan gaan we samen stapsgewijs door het systeem heen.		
		Taak C: Eetboek	“**U heeft op vrijdag om 18.45u ‘s avonds 70 gram ‘Aardappelen gebakken’ gegeten. Zet dit in het systeem.”** Taak begint op hoofdscherm (ingelogd). Moderator vraagt hoe de deelnemer deze taak zou aanpakken. Taak eindigt wanneer de deelnemer op slide 9 op “OK” drukt en op het hoofdscherm belandt. Moderator: houdt tijd bij.	Stopwatch Formulier start en eind voor taak C (appendix B.3) Taak-beschrijving voor taak C (appendix C)	5

			Als de taak niet is afgerond binnen 5 minuten wordt de taak stopgezet door de moderator.		
			Deelnemer vult after-scenario questionnaire in voor taak C.	After-Scenario Questionnaire (appendix D)	2
		Taak D: Doelenboek	Deze taak bestaat uit 2 subtaken. “Als eerste: U wilt een nieuw doel aanmaken, namelijk de komende 6 weken op zaterdag uw ontbijt aan tafel eten. Hier wilt u ook elke week een herinnering over ontvangen. Zet dit in het systeem.” “Als tweede: U had afgelopen woensdag als doel om met iemand te eten, maar dit is niet gelukt. Zet dit in het systeem.” Taak begint op hoofdscherm (ingelogd). Moderator vraagt hoe de deelnemer deze taak zou aanpakken. Taak eindigt wanneer de deelnemer op slide 16 op “Niet gehaald” drukt en op het hoofdscherm belandt. Moderator: houdt tijd bij	Stopwatch Formulier start en eind voor taak D (appendix B.4) Taak-beschrijving voor taak D (appendix C)	5

			Als de taak niet is afgerond binnen 5 minuten wordt de taak stopgezet door de moderator.		
			Deelnemer vult after-scenario questionnaire in voor taak D.	After-Scenario Questionnaire (appendix D)	2
		Taak E: Verhalen	"**U heeft zin in een gezellige activiteit. Meld u aan voor de Jordaanboot."** Taak begint op hoofdscherm (ingelogd). Moderator vraagt hoe de deelnemer deze taak zou aanpakken. Taak eindigt wanneer de deelnemer op slide 8 op "Aanmelden" drukt. Moderator: houdt tijd bij. Als de taak niet is afgerond binnen 5 minuten wordt de taak stopgezet door de moderator.	Stopwatch Formulier start en eind voor taak E (appendix B.5) Taak-beschrijving voor taak E (appendix C)	
			Deelnemer vult after-scenario questionnaire in voor taak D.	After-Scenario Questionnaire (appendix D)	2
		SUS	Deelnemer vult SUS in.	SUS Vragenlijst (appendix E)	5

		Verkrijg demografieken	Deelnemer vult demografieken vragenlijst in.	Demografieken Vragenlijst (Appendix F)	2
3	Afsluiting	Algemene indruk PACO	Moderator zet video- en audio recorder uit. Moderator stelt de vragen: Wat vond u van de applicatie? Wat zou u liever anders zien? Wat vond u fijn aan het systeem? Wat vond u van de manier waarop de coach met u communiceerde?		5
			Tijdsduur		48 min p.p.

Appendix H: Usability issues

Type	Severity	Description	Location	Solution
UI	Critical	Button '*afronden'* does not work	Account creation	Fix the button
UI	Critical	Interface is too small to read when typing username and password on a Samsung tablet	Account creation	Solve with hack: if screen is in landscape mode and the height suddenly decreases by 25%, it is assumed the keyboard is open
UI	Serious	Participants want to click on unclickable buttons	Goal book Stories	Make every answer option on the left side of the screen clickable
UI	Serious	Participants have trouble reading the recipe book	Recipe book	Make the recipe book larger or create the possibility to zoom in
UI	Minor	Buttons appear in dialogue instead of interaction area	Account creation Recipe book	Move buttons from dialogue to interaction area
UI	Minor	Typo: '*anderen maaltijden*'	Recipe book	Replace '*anderen maaltijden*' with '*andere maaltijden*'

UX	Critical	Participants do not know the purpose of the modules	Across app	Provide a short description of each module using dialogues after an account is created
UX	Critical	The words '*Change a status*' are confusing	Goal book	Change the answer options in the goal book to 1) Create a new goal (*Nieuw doel aanmaken)* and 2) Goal successful / unsuccessful (*Doel gehaald / niet gehaald*)
UX	Critical	Participants do not recognize the ability to scroll through the recipe list	Recipe book	Let users know they can scroll through the recipe list in Herman's dialogue
UX	Serious	The term '*Eating diary*' is confusing	Home screen	Change '*Eating diary*' (*Eetboek*) to '*My nutrition*' (*Mijn voeding)*
UX	Serious	The term '*Goal book*' is confusing	Home screen	Change '*Goal book*' (Doelenboek) to '*Goals*' (*Doelen*)
UX	Serious	Participants think they get notified about a certain goal too late	Goal book	Create the possibility to adjust date and time of notifications by participants themselves

UX	Serious	Some activities are not within reach of participants	Stories	Create the possibility to sort activities in terms of postal code area or a radius of X kilometers
UX	Minor	The word '*Finish*' (*Afronden*) is confusing	Recipe book	Change '*Finish*' to '*Continue*' (*Verder*)
UX	Minor	Finding the correct recipe is hard	Recipe book	Make sure the recipes are listed in alphabetical order
UX	Minor	Inconsistent use of informal and formal salutation	Across app	Solely use formal salutation