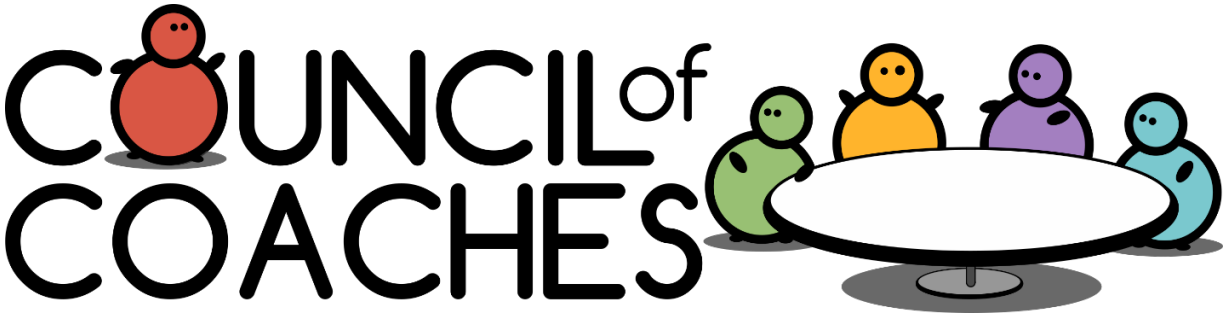

D7.7: Final Demonstration Results

Dissemination level: Public
Document type: Report
Version: 1.0.0
Date: August 26th, 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Document Details

Project Number	769553
Project title	Council of Coaches
Title of deliverable	Final Demonstration Results
Due date of deliverable	August 31st, 2020
Work package	7
Author(s)	Marian Hurmuz (RRD), Tessa Beinema (RRD), Stephanie Jansen-Kosterink (RRD), Dominic De Franco (UDun), Silke ter Stal (RRD), Harm op den Akker (RRD)
Reviewer(s)	Álvaro Fides Valero (UPV) and Sita Ramchandra Kotnis (DBT)
Approved by	Coordinator
Dissemination level	Public
Document type	Report
Total number of pages	81

Partners

- University of Twente – Centre for Monitoring and Coaching (CMC)
- Roessingh Research and Development (RRD)
- Danish Board of Technology Foundation (DBT)
- Sorbonne University (SU)
- University of Dundee (UDun)
- Universitat Politècnica de València, Grupo SABIEN (UPV)
- Innovation Sprint (iSPRINT)

Abstract

This deliverable reports the results of the final evaluation of the Council of Coaches web application as performed in the Netherlands and Scotland. This evaluation was conducted in two rounds in both countries. The aim of this evaluation study was to assess the user experience, the use and the potential health effects of a fully working Council of Coaches system implemented in a real-world setting among the target population, consisting of older adults or adults with diabetes mellitus type 2 or chronic pain.

Table of Contents

List of figures

List of tables

Symbols, abbreviations and acronyms

CMC	Centre for Monitoring and Coaching
COUCH	Council of Coaches
CP	Chronic Pain
D	Deliverable
DBT	Danish Board of Technology Foundation
DM2	Diabetes Mellitus Type 2
ISPRINT	Innovation Sprint
M	Month
MRT	Micro-Randomized Trial
RRD	Roessingh Research and Development
SMAS-s	Self-Management Ability Scale – short version
SU	Sorbonne University
SUS	System Usability Scale
TAM	Technology Acceptance Model
UDun	University of Dundee
UPV	Universitat Politècnica de València
UT	University of Twente
VAS	Visual Analogue Scale
WAI-ReD	Working Alliance Inventory questionnaire – Dutch version in rehabilitation setting
WMO	Wet medisch-wetenschappelijk onderzoek met mensen [Medical Research Involving Human Subjects Act]
WP	Work Package

1 Introduction

This document describes the final evaluation of the functional prototype in the Netherlands and Scotland. First, we shortly describe in Section 3 the methods which are described in detail in D7.6 (Hurmuz M. , Jansen-Kosterink, Op den Akker, & De Franco, 2020). In Section 4 we describe the results of this evaluation in the Netherlands, and in Section 5 the results of the evaluation as performed in Scotland. Table 1 shows a brief overview of the method, setting and participants of the evaluation in both countries. The results of the Micro-Randomized Trial (MRT) within this evaluation will be described in Section 6. In Section 7 we discuss the results, and in Section 8 we conclude this deliverable.

Table 1: Overview of the final demonstration studies performed in D7.7.

Study	**Method**	**Setting**	***N***	**Participants < 54**	**Participants > 55**	**Participants with health conditions (DM2, CP)**
Evaluation of the final functional prototype in the Netherlands	Observational cohort study with a pre-test/post-test design	Real-world setting	51	0	51	DM2 (7) CP (6) DM2+CP (1)
Evaluation of the final functional prototype in Scotland	Observational cohort study with a pre-test/post-test design	Real-world setting	41	4	37	DM2 (10) CP (0) DM2+CP (2)
		Totals	**92**			**DM2 (17)** **CP (6)** **DM2 + CP (3)**

2 Objectives

The objective of this deliverable is to report on the results of the final evaluation of the functional demonstrator of the Council of Coaches (COUCH) web application in the Netherlands and Scotland. In Council of Coaches, there are four official cycles of demonstrator releases followed by evaluations, as depicted in Table 2 below. Each of these cycles have resulted in a report on the evaluation results and updated requirements. The first, second and third rounds have been completed; see D2.4 (Beinema, et al., 2019), D2.5 (Beinema, et al., 2019), and D2.6 (Van der Kamp, et al., 2019). This deliverable (D7.7) describes the results of the fourth (and final) evaluation round.

Table 2: Council of Coaches evaluation cycles.

Council of Coaches Evaluation Cycles	
Cycle 1	
M9	Milestone 2: First Functional Prototype
M12	D2.4: Evaluation results of first functional prototype and updated requirements
Cycle 2	
M15	Milestone 3: Second Functional Prototype
M18	D2.5: Evaluation results of second functional prototype and updated requirements
Cycle 3	
M21	Milestone 4: Third Functional Prototype
M24	D2.6: Evaluation results of third functional prototype and updated requirements
Cycle 4	
M27	Milestone 5: Technical Prototype
M36	D7.7: Final Demonstration Results

3 Methods

The objective of this study was to evaluate the user experience with-, the use of-, and the potential health effects of a fully working Council of Coaches functional demonstrator implemented in a real-world setting among the target population.

The methods of this evaluation study have been described in detail in D7.6 (Hurmuz M. , Jansen-Kosterink, Op den Akker, & De Franco, 2020), and are published in a scientific journal (Hurmuz M. Z., Jansen-Kosterink, Op den Akker, & Hermens, 2020). This section shortly describes the methods and changes made in the protocol in relationship to the description in D7.6.

3.1 Study design and participants

This observational cohort study consisted of 5-9 weeks (see Figure 1). The first week is the baseline week, during this week participants only wore the Fitbit inspire activity tracker (https://www.fitbit.com/global/us/products/trackers/inspire). From the second week, the implementation phase started, and participants used the functional demonstrator of Council of Coaches. After the implementation phase (week 5), participants could choose whether they want to continue using this functional demonstrator for four extra weeks, this is the follow-up phase (week 6 – week 9).

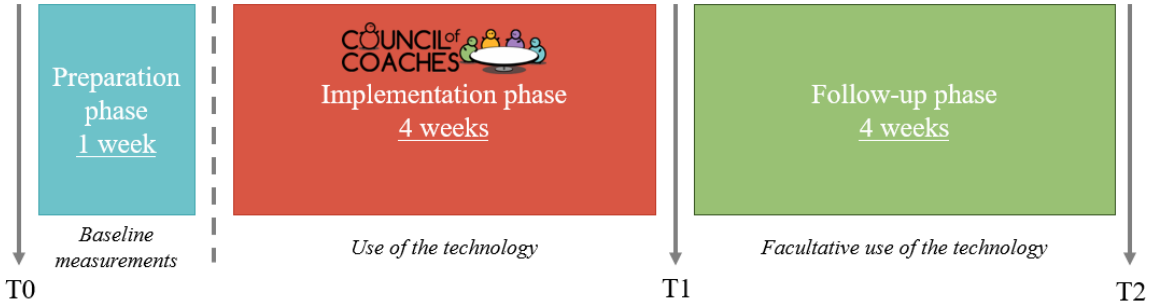

Figure 1: Study design and timeline for this evaluation study.

This study was conducted in two rounds in the Netherlands and in Scotland. The first round of this study started on January 31st, 2020 in the Netherlands and on February 20th, 2020 in Scotland. The second round started on May 29th, 2020 in the Netherlands and on June 11th, 2020 in Scotland.

Participants were included in this study if they met the following inclusion criteria:

- Older adults aging 55 years or older
- Able to read and speak Dutch or English
- Having Wi-Fi connection at home
- Able to give informed consent
- Able to see smartphone or tablet screen clearly

3.2 Study procedure

Due to COVID-19, some changes are made to the study procedure stated in the protocol in D7.6 (Hurmuz M. , Jansen-Kosterink, Op den Akker, & De Franco, 2020). During the first round of the evaluation, the Netherlands and Scotland went into (an "intelligent") lockdown to slow down the spread of the COVID-19 virus. Due to this, we could not and did not consider it responsible to have face-to-face contact with the participants. The exit-interviews which were planned for execution after the implementation phase were conducted by phone instead of by visiting the participants. The questionnaires were sent by e-mail and the participants could complete them on their phone/tablet/laptop/PC. Furthermore, the equipment we gave the participants was returned by mail after the implementation phase or after the facultative phase.

For the second round, everything was conducted online. Before the beginning of the study, the participants received their equipment by post. With this equipment a comprehensive guideline was given which explained step-by-step how to download the Fitbit app, how to connect the Fitbit activity tracker and how to use them, and how to create an account in the functional demonstrator and use it.

3.3 Study outcomes

The primary study outcomes were the use of the functional demonstrator, user experience with this demonstrator and potential health effects. First of all, use was determined by the log history of the functional demonstrator platform, defined as frequency (number of sessions in each week), as duration of use (number of minutes per session in each week), and as interaction steps (number of dialogue steps per session in each week).

For user experience, several questionnaires were used, and an interview was conducted at T1. The questionnaires used were: the Technology Acceptance Model (TAM) (Davis, 1989; Davis, Bagozzi, & Warshaw, 1989) which measures 7 user experience domains, the System Usability Scale (SUS) (Brooke, 1996), and the Willingness-to-Pay.

Potential health effects were measured using the EQ-5D-5L questionnaire (Van Reenen & Janssen, 2015), the six domains of the Positive Health tool (Huber, et al., 2016), and the Self-Management Ability Scale – short version (SMAS-s) (Schuurmans, et al., 2005).

The secondary study outcomes were applicability of the virtual coaches and user's interaction with the virtual coaches. The applicability of the virtual coaches was measured with the Working Alliance Inventory questionnaire – Dutch version in rehabilitation setting (WAI-ReD), with a rating scale (from 1 to 10), and by interviews. The user's interaction with the virtual coaches was conducted by means of a Micro-Randomized Trial, which is explained in detail in section 3.6.

Table 3 gives an overview of the questionnaires used and shows at which measurement time.

Table 3: Overview of which questionnaires were used at which point during the evaluation.

	T0	T1	T2
Demographics	X		
User experience			
Technology Acceptance Model		X	
System Usability Scale		X	
Willingness-to-Pay		X	
Potential health effect			
EQ-5D-5L	X	X	X
Positive Health dimensions	X	X	X
Self-Management Ability Scale – short version	X	X	X
Applicability of the virtual coaches			
Rating scale	X	X	
Working Alliance Inventory		X	

3.4 Sample size

Because of the explorative character of this study, no sample size calculation was conducted beforehand. To be able to answer the objectives of this study, the goal was to include 50 participants per country. So, in each round, we aimed at including 25 participants per country.

3.5 Data analyses

Statistical analyses were performed using SPSS, version 19 for Windows. For all the analyses, the confidence intervals were set at 95%. Descriptive statistics, such as frequency, mean, standard deviation and percentages, were used to describe demographics, user experience, use, and the applicability of the virtual coaches.

For analysing the log data some rules were specified:

- The duration of use is the number of minutes in which participants interacted with the coaches. If a participant was just listening to the radio, or browsing through the recipe book, it was not considered as a session.
- A session was deleted when it was shorter than 1 minute.
- If the time between two interactions was at least 1 minute or more, it was considered as a break. All breaks were listed, and the median break time was searched. All break times that were above this median, were deleted from the number of minutes of the corresponding session.
- A session was considered to become another session when the break between two interactions lasted for at least 20 minutes.

The outcome on the EQ-5D-5L, the Positive Health tool, and the SMAS-s were assessed on normality. With histograms we saw that the variables were not normally distributed. Therefore, we needed to use a non-parametric test to assess the potential health effects. For the results of the Netherlands, the Friedman test was used, post hoc comparisons were made with the Wilcoxon Signed-rank test with a Holm-Bonferroni correction. For the results of Scotland, only the Wilcoxon Signed-rank test was used because few participants completed the T2 questionnaire, so in this case we only looked at the potential health effects between T0 and T1.

For the analyses of the T1-questionnaire, T2-questionnaire and interviews we used a per protocol analysis. Participants that did not use the functional demonstrator were omitted in these results, because they could not give proper answers if they had not used it. Furthermore, for user experience and potential health effect results in the Netherlands, we divided the data in two use groups:

- Use group A: participants that used the functional demonstrator for less than four days during the implementation phase.
- Use group B: participants that used the functional demonstrator for at least four days during the implementation phase.

3.6 Micro-Randomized Trial

As described previously in Deliverable 7.6 (Hurmuz M. , Jansen-Kosterink, Op den Akker, & De Franco, 2020), the Micro-Randomized Trial (MRT) was included to assess the effectiveness of the interaction between the user and the virtual coaches.

The MRT was implemented for the interaction with the Council's physical activity coach. Every time the user starts a conversation with that coach, the initiative in directing the conversation is randomized. This randomization is done with the following two conditions:

- Condition 1: *'Coach initiative'*. The coach takes the initiative and suggests the topic of conversation.
- Condition 2: *'User initiative'*. The user gets the initiative and chooses the topic of conversation.

The hypothesis for the MRT was that when the coach takes the initiative (suggests something to talk about) this is cause for longer *interactions* between user and coach.

The outcome parameter that will be used to measure the length of an interaction with the coach is the number of *dialogue steps* that is completed from the start of the interaction. That is, every statement made by the coach counts as a dialogue step, and the same holds for replies that the user gives.

A more detailed description of the full Micro-Randomized Trial can be found in:

Beinema, T., op den Akker, H, Hurmuz, M., Jansen-Kosterink, S., Hermens, H. (2020). *Automatic topic selection for embodied conversational agents in health coaching: a micro-randomized trial.* To be published.

3.6.1 Design and implementation

The content of the physical activity coach for the MRT consisted of eight topics, following the structure described in Deliverable 3.4 (Beinema, Op den Akker, Kosterink, ter Stal, & van den Boer, 2019), namely:

- *Introduction*. A dialogue in which the coach globally introduces themselves to the user and provides some short information on their background and coaching content.
- *Background story.* Dialogues in which the coach shares a part of their background story with the user. E.g. a short story about her dog ('Brian') and how she likes to go running with him.
- *Discuss sensors*. Dialogues that cover subtopics such as 'Why should I use an activity tracker?' and 'How do I connect my tracker to the Council of Coaches?'.
- *Goal-setting*. Dialogues in which the user can set a new goal or change their current goal.
- *Feedback*. Dialogues that allow the user to view their measured activity data.
- *Gather information*. Dialogues in which the coach asks the user some questions to gather information that can be used tailor the coaching. E.g. 'Do you have a dog?' which can be used to suggest that the user walks with their dog more often.
- *Inform 'why'*. Dialogues in which the coach explains why it is good to be physically active. E.g. 'being active increases blood flow, which is healthy for your brain'.
- *Inform 'how'*. Dialogues in which the coach explains how to be more physically active. E.g. 'take the stairs instead of the elevator'.

Every time the user would click on the physical activity coach ('Olivia') the system would micro-randomly select one of the two conditions, both with a 50% chance. The difference between the two conditions was in the start of the dialogues while the dialogues that followed and the topics available in both conditions were the same (with the exception of the 'gather information' topic which was only included in the 'coach initiative' condition).

In the 'coach initiative' condition a dialogue would be started on a relevant topic, which was selected by a topic selection algorithm that took into account parameters such as e.g. available dialogues and completion dates. In this case, the dialogues for the different topics would be preceded by a short 1-step dialogue that would add a sentence such as 'Why don't we discuss *X*?' (see Figure 2). This allowed for the same dialogues to be used in both conditions, with just a change to the start of the dialogues. The user would then be able to reply with, e.g.:

- 'Go ahead!' which would start the dialogue about the suggested topic.
- 'I want to talk about something else.' which would lead to the menu-dialogue from which they could choose another topic.
- 'Goodbye.' which would end the conversation.

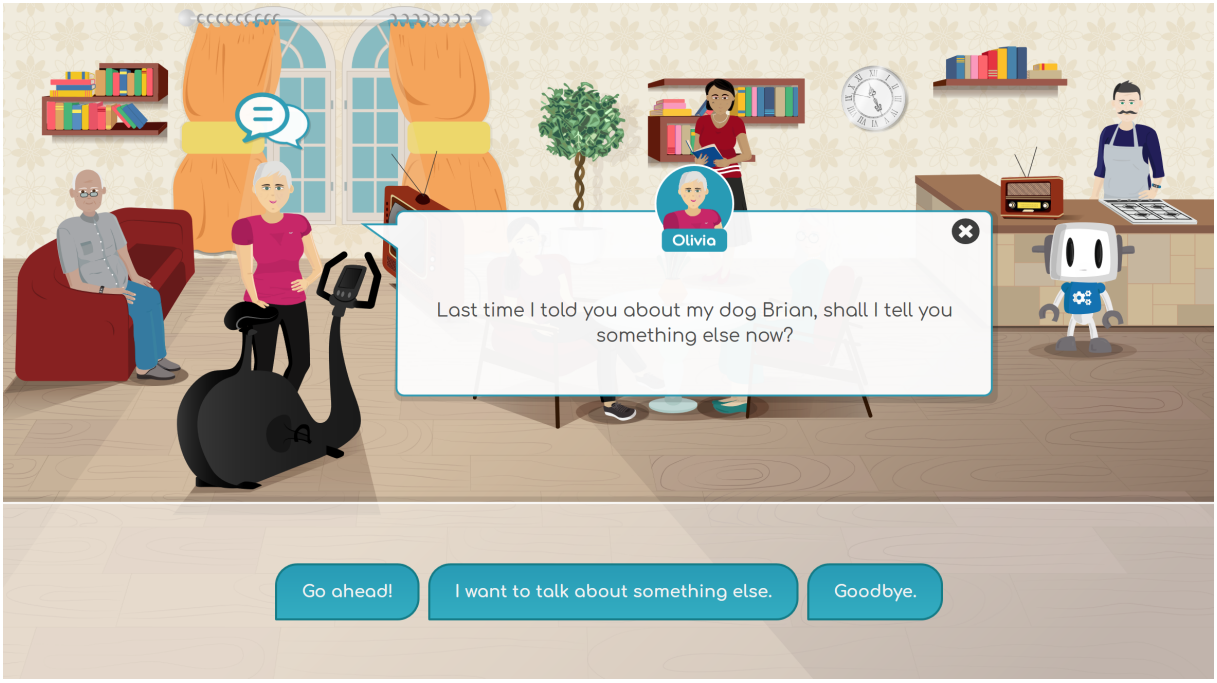

Figure 2: An example initiation of a 'background story' dialogue in Condition 2 ('coach initiative' / 'AI'-condition) for the MRT.

In the 'user initiative' condition, the user would be shown a menu-dialogue (see Figure 3). The coach would state 'Hey there, nice to see you again! How can I help you?' and the user could select:

- 'I just wanted to chat.' leading to the possibility to select a social topic, such as the background stories.
- 'Let's talk about coaching.' leading to the possibility to select a coaching topic, such as feedback or inform 'how'.
- 'Goodbye.' ending the conversation.

These responses could lead to another choice that allowed the user to further specify their desired topic of conversation until they would have selected a topic that involved a dialogue. In this manner, the menu-dialogue would allow the user to navigate towards the topic of their preference.

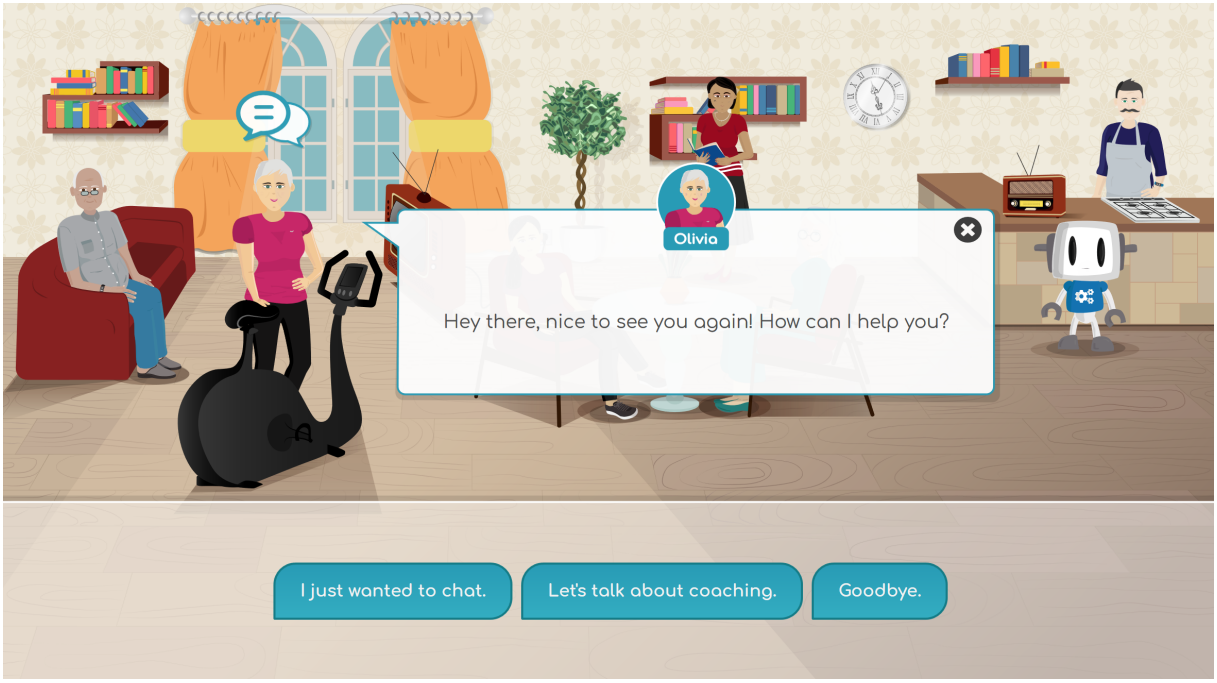

Figure 3: The start of a dialogue in Condition 1 ('user initiative' / 'menu'-condition) in the MRT.

3.6.2 Data and pre-processing

We will retrieve data from the system's dialogue logs for the physical activity coach. The system's dialogue logs will include a record for every dialogue that was started with a coach and has details for each dialogue such as:

- The statements by the coach and the user's replies to those statements. These will be used to compute the number of dialogue steps.
- Timestamps per statement and reply.
- Whether the dialogue was *cancelled*, i.e. the speech bubble was closed using the 'X'-button.
- Whether the dialogue was *completed*, i.e. the dialogue ended because it was finished, or the user ended the dialogue by e.g. selecting the 'Goodbye'-reply.
- The MRT condition for the dialogue ('coach initiative' or 'user initiative').
- If the dialogue was started because the user was referred to it by another dialogue and which one (e.g. if the user selected the 'I want to talk about something else' option in a dialogue they would be referred to the 'main menu' dialogue).

The outcome parameter that will be used for analysis will be the *number of dialogue steps per interaction*. We define an *interaction* as follows:

- It starts when the user clicks on the coach.
- It ends when a dialogue:
  - is completed. This occurs when the user chooses a reply that ends the dialogue, such as 'Goodbye'.
  - is cancelled. This occurs when the user closes the speech bubble using the 'X'-button.
  - is never completed or cancelled. This can occur when the browser is closed, or the user is not active for a while (which causes the browser to refresh).

In addition to our definition for an interaction in general, we set the following boundaries for the interactions in the two conditions:

- A 'coach initiative' condition interaction:
  - Starts with a dialogue ending in '-ai'.
- A 'user initiative' condition ('menu'-condition) interaction:
  - Starts with the 'olivia-menu' dialogue, and
  - That 'olivia-menu' dialogue has no *referring* dialogue.

Furthermore, we will apply the following pre-processing steps to the system's log data:

1) We exclude dialogue logs that are not for dialogues held with the physical activity coach.
2) We exclude logs for dialogues that were the result of a 'double click' error. That is, two dialogues are started within 1 second of each other, with the first log only including the agent's first statement.

3.6.3 Analysis

We analysed both evaluation rounds separately. A paired samples t-test was used to compare the mean number of dialogue steps per participant for the interactions that they had in the coach initiative condition with the mean number of dialogue steps they had in the user initiative condition. Analysis was performed using version 25 of the SPSS statistics program.

3.7 Ethical approval

This study was conducted according to the principles of the Declaration of Helsinki (64th WMA General Assembly, Fortaleza, Brazil, October 2013) and in accordance with the Medical Research Involving Human Subjects Act (Dutch law: *Wet medisch-wetenschappelijk onderzoek met mensen* (WMO)). According to the WMO, this study did not require formal medical ethical approval to carry this out in the Netherlands. This was checked by the MREC CMO Arnhem-Nijmegen (file number: 2019-5555). For Scotland, the ethical approval was given by the School of Science and Engineering Research Ethics Committee (SSEREC) at UDun. In both countries, each participant gave his/her informed consent on paper beforehand.

4 Results – the Netherlands

4.1 Demographics

During the first round, 26 participants were included, of which one participant dropped out directly at the beginning of the study. During the second round, 25 participants were included. The majority of both groups was female (round 1 = 69.2%, round 2 = 72.0%), and had a mean age of 66.5 (SD=7.5) in the first round and 64.0 (SD=7.3) in the second round. Table 4 shows all demographics of both rounds.

Table 4: Demographics of the study population in the Netherlands of the first (N=26) and second round (N=25).

Demographic	First round (N=26)	Second round (N=25)
Gender (%)		
Male	30.8	28.0
Female	69.2	72.0
Age (M (SD))	66.5 (7.4)	64.0 (7.3)
Level of education (%)		
Preparatory secondary vocational education	23.1	12.0
Higher general secondary education, pre-university education	34.6	28.0
Higher vocational education, university	42.3	60.0
Living situation (%)		
Alone	23.1	32.0
Married/living together	76.9	68.0
Work status (%)		
Employed	26.9	28.0
Volunteer/caregiver	19.2	8.0
Retired	50.0	40.0
Other	3.8	24.0
Health literacy (M (SD))	4.0 (0.6)	4.1 (0.6)
Self-reported level of physical activity (%)		
Not at all	3.8	-
Not at all, but thinking about beginning	3.8	4.0
Less than 2.5 hours a week	34.6	28.0
More than 2.5 hours a week in the last six months	11.5	24.0
More than 2.5 hours a week for more than six months	46.2	44.0
Attitude towards technology (M (SD))	4.6 (1.3)	4.4 (1.7)
Motivation type to live healthy (M (SD))		
Intrinsic motivated	5.2 (1.2)	5.0 (0.8)
External regulated motivation	2.6 (1.2)	2.9 (1.2)
A-motivated	2.1 (1.4)	2.1 (1.1)

4.2 Use of Council of Coaches functional demonstrator

Table 5,

Table 6 and Figure 4 show the use data of the functional demonstrator of total study duration, divided into the first and second round. The implementation phase is from week 2 until 5, the facultative phase is from week 6 until 9. Figure 5 shows an overview of the number of participants per phase and how many participants stopped using the functional demonstrator at which time point during the study period. Two participants that dropped out during the implementation phase of the first round gave reasons for this. One participant did not like using technology for coaching; (s)he rather wants face-to-face coaching. The other one did not like the way the coaches were talking to him/her. One participant in the first round did not use the functional demonstrator during the whole study period. The questionnaires data and interview data of this participant will be excluded in the results. One participant in the second round did only use it during the facultative phase. The questionnaires data and interview data of this participant will also be excluded in the results. During the implementation phase of both rounds, 24 participants used the functional demonstrator. The mean number of sessions participants had in this phase was highest in week 2; first round: M=3.6 (SD=2.1), and second round: M=3.1 (SD=2.3), and lowest in week 4; first round: M=1.6 (SD=0.9), and second round: M=1.4 (SD=0.7). During the facultative phase, 12 participants of the first round and 10 participants of the second round continued using the functional demonstrator. In this phase, the mean number of sessions was highest in week 8; first round: M=2.0 (SD=1.5), and second round: M=2.0 (SD= -), and lowest in week 7 in the first round: M=1.6 (SD=0.8), and week 9 in the second round: M=1.5 (SD=0.7).

Table 5: Use data of total system of both rounds: number of participants used the functional demonstrator per phase, mean (SD), min and max number of days used the functional demonstrator and of sessions within the functional demonstrator.

First round					
Phase	*N*	*Mean (SD) number of days used COUCH*	*Range min-max number of days used COUCH*	*Mean (SD) number of sessions within COUCH*	*Range min-max number of sessions within COUCH*
Implementation phase	24	5.9 (4.0)	1 – 15	6.9 (5.4)	1 – 23
Facultative phase	12	3.7 (3.6)	1 – 11	4.1 (4.6)	1 – 15
Second round					
Phase	*N*	*Mean (SD) number of days used COUCH*	*Range min-max number of days used COUCH*	*Mean (SD) number of sessions within COUCH*	*Range min-max number of sessions within COUCH*
Implementation phase	24	4.7 (3.4)	1 – 12	5.3 (4.0)	1 – 15
Facultative phase	10	1.9 (1.6)	1 – 6	2.4 (2.0)	1 – 7

Table 6: Use data of total system of both rounds: number of participants used the functional demonstrator per week, mean (SD), min and max number of sessions per week, mean (SD), min and max duration in minutes per session per week, and mean (SD), min and max number of interactions per session per week

First round							
Week	*N*	*Mean (SD) number of sessions*	*Range min-max number of sessions*	*Mean (SD) duration in minutes per session*	*Range min-max duration in minutes per session*	*Mean (SD) number of interactions per session*	*Range min-max number of interactions per session*
1	-	-	-	-	-	-	-
2	22	3.6 (2.1)	1 – 9	6.4 (4.7)	1.0 – 21.0	109.2 (90.0)	9 – 471
3	17	2.1 (1.5)	1 – 6	5.7 (4.0)	1.2 – 19.2	78.3 (43.5)	20 – 192
4	14	1.6 (0.9)	1 – 4	5.8 (3.2)	1.9 – 13.6	92.2 (44.7)	19 – 204
5	14	2.1 (1.7)	1 – 7	4.3 (3.5)	1.1 – 18.0	66.0 (38.3)	18 – 142
6	9	1.9 (1.4)	1 – 5	4.0 (3.1)	1.3 – 14.6	71.5 (43.6)	16 – 168
7	7	1.6 (0.8)	1 – 3	5.6 (3.1)	1.8 – 11.2	83.5 (41.2)	36 – 173
8	6	2.0 (1.5)	1 – 4	4.7 (2.4)	1.3 – 8.4	67.3 (28.3)	13 – 115
9	5	1.8 (1.3)	1 – 4	3.5 (2.0)	1.2 – 7.7	53.2 (25.6)	21 – 104
Second round							
Week	*N*	*Mean (SD) number of sessions*	*Range min-max number of sessions*	*Mean (SD) duration in minutes per session*	*Range min-max duration in minutes per session*	*Mean (SD) number of interactions per session*	*Range min-max number of interactions per session*
1	-	-	-	-	-	-	-
2	22	3.1 (2.3)	1 – 11	8.1 (5.4)	1.4 – 23.1	129.2 (100.6)	14 – 697
3	16	1.9 (1.3)	1 – 5	9.3 (6.6)	1.1 – 23.1	154.3 (114.8)	6 – 448
4	11	1.4 (0.7)	1 – 3	11.0 (8.9)	1.1 – 29.5	166.6 (92.4)	31 – 339
5	8	1.6 (0.7)	1 – 3	9.4 (9.0)	1.4 – 28.6	147.5 (107.5)	24 – 388
6	9	1.6 (0.9)	1 – 3	6.5 (4.8)	1.2 – 15.4	112.8 (82.1)	19 – 282
7	3	1.7 (1.2)	1 – 3	3.8 (2.0)	2.3 – 7.2	71.6 (24.6)	45 – 100
8	1	2.0 (-)	2 – 2	6.1 (3.2)	3.8 – 8.3	163.0 (97.6)	94 – 232
9	2	1.5 (0.7)	1 – 2	9.5 (7.2)	1.2 – 13.9	161.3 (106.9)	38 - 227

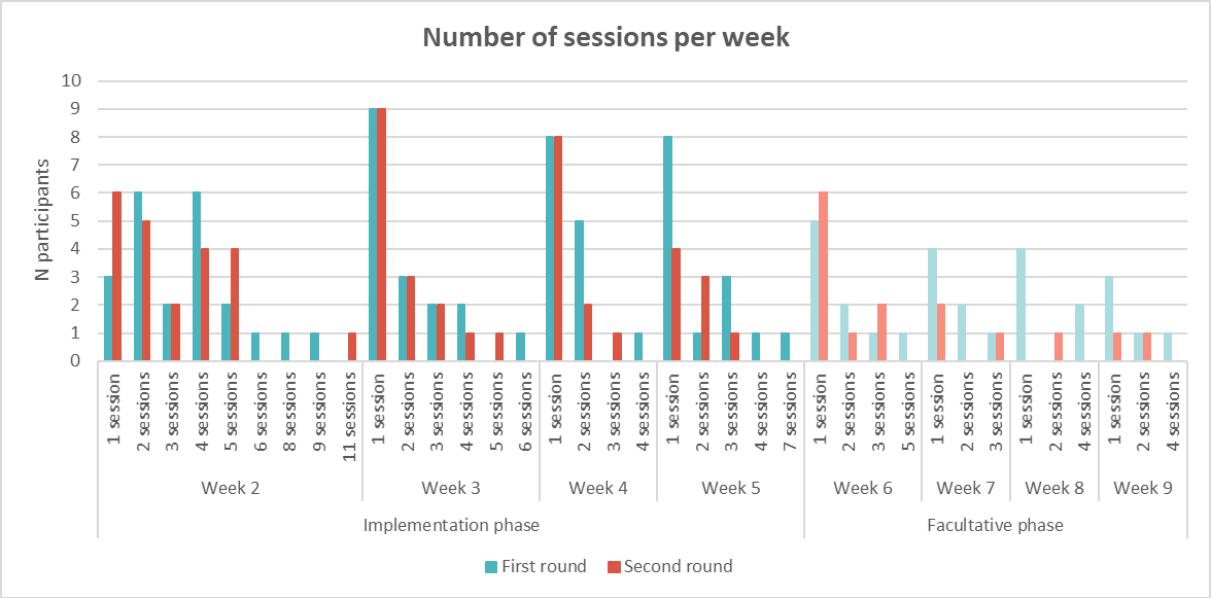

Figure 4: Frequency graph; number of participants vs. number of sessions within the functional demonstrator per week.

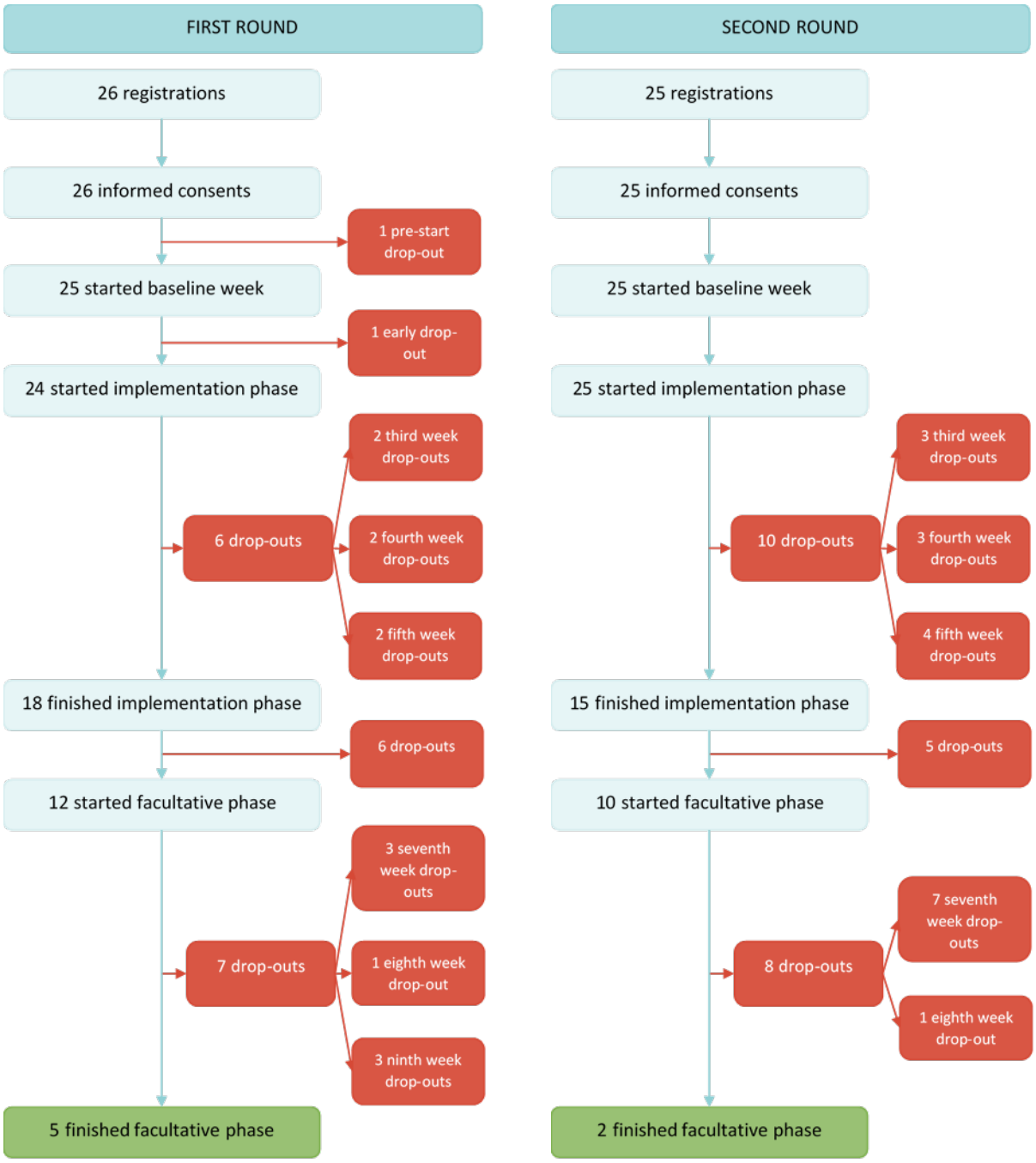

Figure 5: Flowchart number of participants per phase and drop-outs.

4.3 User experience

The usability of the functional demonstrator was scored with a mean of 35.3 (SD=19.6, N=8) by participants in the first round and with a mean of 45.9 (SD=20.2, N=11) by participants in the second round of use group A. Participants in use group B, scored the usability with a mean of 58.8 (SD=14.5, N=15) during the first round, and 57.7 (SD=20.4, N=12) during the second round. This means that, according to participants that used the functional demonstrator less than four times, the usability of it was not acceptable, and according to participants that used the functional demonstrator for at least four times, the acceptability of the usability was low.

One participant (12.5%) in the first round and two participants (18.2%) in the second round of use group A, indicated they are willing to pay for the COUCH functional application. The participant of the first round is willing to pay 5 euros per month (N=1), and the participants of the second round are willing to pay an average of 6 euros per month (N=5). During both rounds, three participants (20% in the first

round, 25% in the second round) of use group B, indicated they are willing to pay for the COUCH application. The average amount of Euros (€) the participants in the first round of this group (N=4) were willing to pay was €7,50 per month, and the participants in the second round (N=3) were willing to pay an average of €5,00 per month.

Regarding the user experience measured with the TAM, participants were in general neutral about the functional demonstrator. During the first round, use group A participants, were most positive about the aesthetics of the COUCH functional demonstrator (M=4.3, SD=1.2), and most negative about control (M=3.2, SD=1.7). Participants in use group B, were most positive about the trust in the technology (M=4.8, SD=1.0), and most negative about the intention to use it (M=2.6, SD=1.6).

During the second round, participants in use group A, were most positive about the trust in technology (M=4.7, SD=1.0), and most negative about the intention to use it (M=2.6, SD=1.8). Participants in use group B, were most positive about the perceived ease of use (M=4.9, SD=0.8), and most negative about the intention to use it (M=3.3, SD=1.9).Table 7 shows the means of the measured user experience domains in both use groups and it shows the percentages of participants that were positive, neutral and negative towards each domain. Figure 6 shows the boxplots of each domain of both use groups.

Table 7: User experience assessed on 7 domains in the first (N=23) and second round (N=23), divided into participants that used the functional demonstrator for less than 4 times (use group A) and that used the functional demonstrator for at least 4 times in the implementation phase (use group B).

	First round (N=23)							
	Use group A (N=8)				*Use group B (N=15)*			
	M (SD)	*% positive*	*% neutral*	*% negative*	*M (SD)*	*% positive*	*% neutral*	*% negative*
Enjoyment	4.3 (1.5)	37.5	50.0	12.5	3.6 (1.1)	6.7	80.0	13.3
Aesthetics	4.3 (1.2)	25.0	75.0	-	4.0 (0.9)	6.7	86.7	6.7
Control	3.2 (1.7)	12.5	50.0	37.5	4.0 (1.6)	26.7	53.3	20.0
Trust in technology	4.2 (1.3)	12.5	62.5	25.0	4.8 (1.0)	33.3	66.7	-
Perceived usefulness	3.6 (1.8)	12.5	62.5	25.0	3.2 (1.5)	20.0	40.0	40.0
Perceived ease of use	3.2 (1.5)	-	75.0	25.0	4.4 (1.5)	40.0	46.7	13.3
Intention to use	3.3 (1.4)	12.5	50.0	37.5	2.6 (1.6)	13.3	33.3	53.3
	Second round (N=23)							
	Use group A (N=11)				*Use group B (N=12)*			
	M (SD)	*% positive*	*% neutral*	*% negative*	*M (SD)*	*% positive*	*% neutral*	*% negative*
Enjoyment	3.8 (0.8)	-	90.9	9.1	3.9 (1.5)	25.0	58.3	16.7
Aesthetics	4.1 (1.3)	18.2	72.7	9.1	4.8 (1.0)	33.3	66.7	-
Control	3.6 (1.4)	9.1	72.7	18.2	4.6 (1.5)	41.7	50.0	8.3

Trust in technology	4.7 (1.0)	36.4	63.6	-	4.6 (0.9)	25.0	75.0	-
Perceived usefulness	3.2 (1.2)	9.1	63.6	27.3	3.8 (1.8)	25.0	50.0	25.0
Perceived ease of use	3.5 (1.7)	27.3	45.5	27.3	4.9 (0.8)	41.7	58.3	-
Intention to use	2.6 (1.8)	9.1	36.4	54.5	3.3 (1.9)	16.7	41.7	41.7

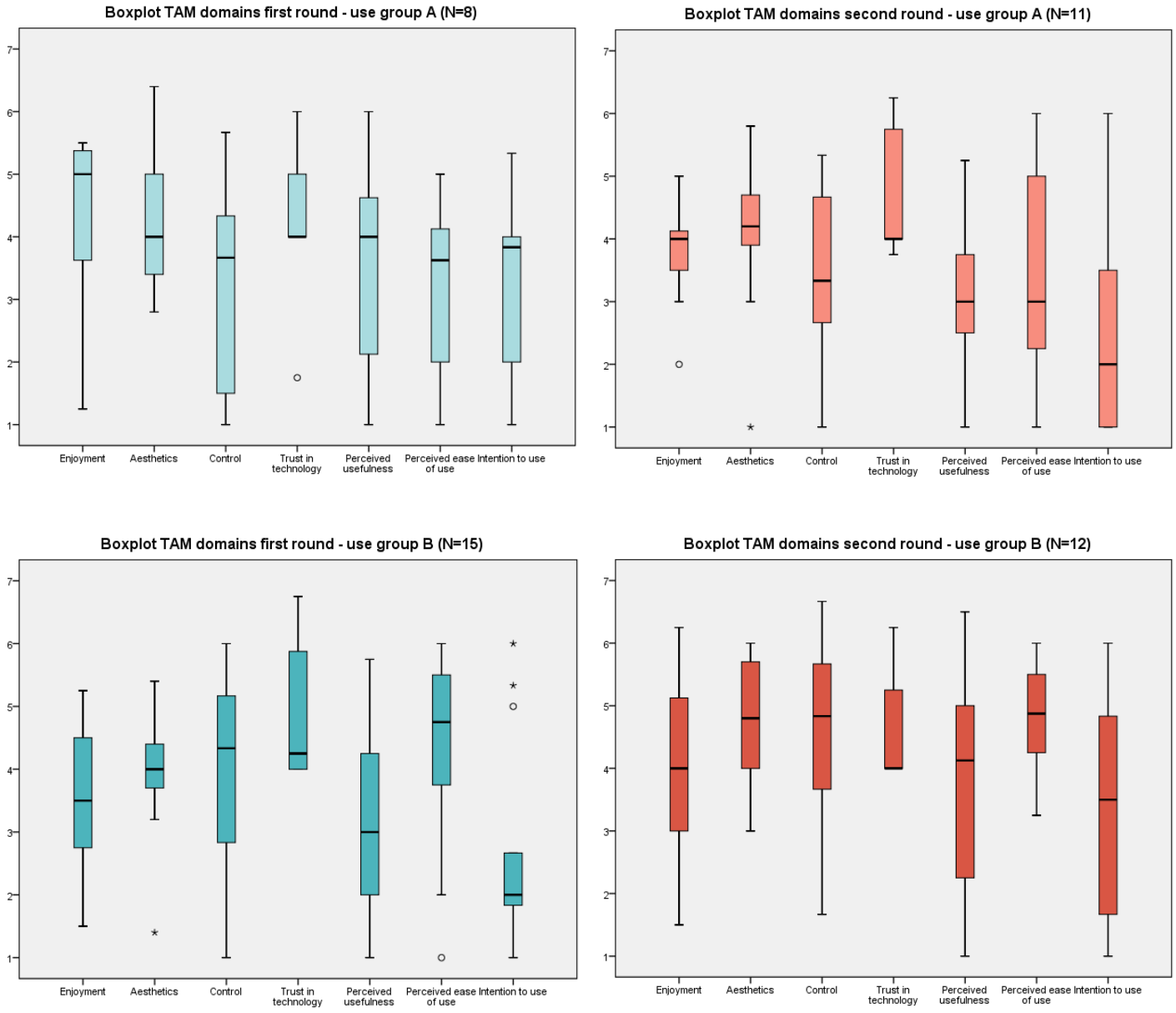

Figure 6: Boxplots user experience assessed by TAM in the first and second round, divided into participants that used the functional demonstrator for less than 4 times (use group A) and for at least 4 times in the implementation phase (use group B).

Besides the user experience questionnaire, interviews were conducted with the participants. Overall, the majority of the participants would not recommend Council of Coaches to others. During the first round,

15 participants out of 24, indicated they would not recommend it, 7 indicated they would, and 2 participants would on the one hand recommend it, but on the other hand not. Reasons they gave for recommending it are shown in Table 8. During the second round 15 participants out of 23, indicated they would not recommend it to others, 5 indicated they would, and 3 participants were neutral about this. The reasoning behind it is also shown in Table 8. Some participants that would not recommend it, gave both reasons for recommending it and not recommending it. Three participants that would not recommend it, said they would recommend it if the coaches have more content.

The mentioned reasons were different between both rounds. During the first round, the three reasons mentioned most to recommend the functional demonstrator to others were: (1) it is a helpful technology (N=5), (2) it is a good application for lonely/less active people who need some support, or people with low literacy (N=4), and (3) it gives the user discipline to follow a particular programme/advice (N=3). The two most mentioned reasons for not recommending it to others were: (1) the coaches have too limited content (N=6), and (2) the coaches give too general information or information not related to personal context (N=6). During the second round, most reasons were just mentioned by one participant. For recommending the functional demonstrator to others, six times it was mentioned that it is good for other people. The most mentioned reason to not recommend the functional demonstrator during the second round was: the coaches give too general information or information not related to personal context (N=9).

Table 8: Reasons mentioned for (not) recommending the functional demonstrator to others during the first round.

First round		
	Reasons for recommending	**Reasons for not recommending**
Mentioned 5 times or more	Helpful technology	Too limited content
		Coaches give too general information or information not related to personal context
Mentioned 2 to 4 times	For people: ▪ who are lonely/less active people and need support ▪ with low literacy	No added value comparing to other sites that are easily available
	It gives the user discipline to follow a particular programme/advice	Childish/patronizing conversations
	Easy access to health advice	Difficulty with log in
		English content
		No need for help in health context
		No answers on specific questions
		No filter in recipe book
Mentioned 1 time	Easy technology	Content
	Information given can be confronting/an eye opener	Too much social talks
	Enjoying the social talks	Cumbersome: wanting to just directly asking a question and receiving an answer instead of a whole conversation
	Improving your health	Layout and working method not good enough
	Coaches are experienced in coaching older adults	Developers have not thought good enough about what the user wants

		No opportunity to really establish a conversation
		Not interactive
		Too slow
		Too simple
		Not the right target group (too healthy and active)
		Poor subdivision in recipe book
		Prefers face-to-face contact
		Experienced too many technical problems
Second round		
	Reasons for recommending	**Reasons for not recommending**
Mentioned 5 times or more	For people who: ▪ have little knowledge about a healthy lifestyle ▪ are older ▪ are lonely/less active who need support ▪ have difficulties changing their lifestyle ▪ want to change their lifestyle	Coaches give too general information or information not related to personal context
Mentioned 2 to 4 times		Childish/patronizing conversations
		Too simple
		Prefers face-to-face contact
		No feedback on performed physical activities
		User design is demotivating/old-fashioned
Mentioned 1 time	Basic information which is important for everyone	Annoying that you need a laptop/PC/tablet to use it
	Information given is clear, with step-by-step explanations	No opportunity to ask something
	Good tips about nutrition	Every time you need to log in again, and start all over again with the conversations
	Good tips to prevent dementia	Not challenging
		No opportunity to really establish a conversation
		Difficulty with log in
		Too little support, no starting initiative from the coaches
		Too limited content
		Does not motivate

Users' reasons for (not) using the functional demonstrator were coded with the six domains of positive health: bodily functions, mental health, meaning, quality of life, social participation, and daily routine (Huber, et al., 2016). Reasons that could not be coded with these domains, were coded as 'other reasons'. Figure 7 shows the percentages each domain was chosen as a reason to (not) use the functional demonstrator. In the first round, 31 reasons were given for using it, and in the second round 19. These reasons were mostly related to the domain bodily functions (N=20), followed by other reasons which are not related to positive health, but to curiosity towards the technology or study, wanting to help researchers and future end-users, and using it for fun (N=6 (first round) and N=2 (second round)). Twenty-seven reasons were given for not using the functional demonstrator in the first round, and 26 in the second round. A small proportion of these reasons were related to the positive health domains (22.2% and 7.1% resp.), of which mostly to bodily functions, participants that perceived themselves already healthy (N=3 and N=2 resp.). Most of the reasons for not using the functional demonstrator were related to the technology (N=19 and N=24 resp.) or not having time to use it (N=2 and N=1 resp.). The most mentioned reasons in the first round were about: content (N=10) and difficulty with log in (N=4), and in the second round about: content as well (N=9) and not stimulating the user enough (N=3).

Some participants indicated they do not have a reason to use or not use the functional demonstrator. Regarding using it, only during the second round seven participants did not have a reason. Regarding not using the functional demonstrator, five participants of the first round and three participants of the second round did not have a reason.

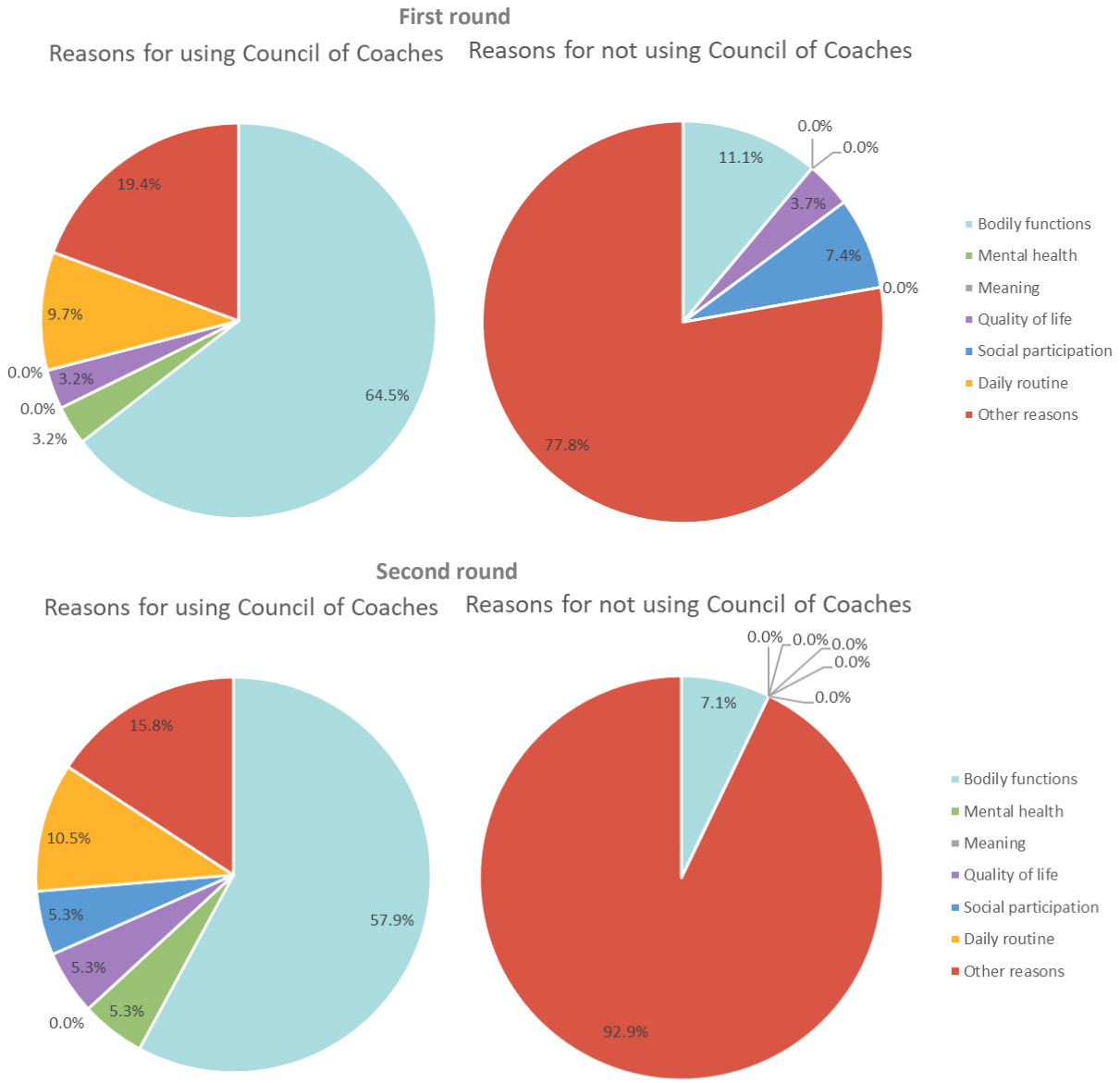

Figure 7: Reasons for (not) using Council of Coaches' functional demonstrator

About the interaction with the system, the participants in both rounds had mixed experiences. In the first round, ten participants indicated that it was easy to use the Council of Coaches' functional demonstrator, two others said it is a good system, and another two participants said the way of interacting is good. One participant said: *"In the beginning, you have to find out how to use it, but then it's quite simple."* (P-6). On the other hand, six other participants experienced some difficulties when using the system: difficulties with logging in into the system (N=4), and difficulties in general with using the system (N=2). Also, two participants indicated they did not like the way of interacting with the system.

In the second round, 14 participants found it easy to use the functional demonstrator, one found it difficult the first time, but after a while it was easy. Fourteen participants found the way of interacting with the system good, and two participants found the way of interacting fun to do. Furthermore, two participants liked the appearance of the functional demonstrator, and one said it was user friendly. However, three participants found it difficult to use it: difficulties in general with using the system (N=2), and difficulties with logging into the system (N=1).

Furthermore, eight participants of the first round and six participants of the second round experienced some problems when using the functional demonstrator. For the first round it was about the following: too slow to work with (N=5), some technical problems (N=2), could not connect the Fitbit with Olivia (N=1), always getting stuck in the begin question of the dialogues (N=1). For the second round these problems were: problems with following the coaching sessions of Helen (N=3), sometimes the system gets stuck (N=2), and problems with logging into the system (N=1).

The majority of participants gave some recommendations for improving the system (N=15 (first round) and N=18 (second round)). The recommendations given in the first round are listed in Table 9 and the ones given in the second round are listed in Table 10. After the recommendations in the first round, some changes were made in the functional demonstrator before the second round of the evaluation started, to improve the system according to the users' feedback. These changes are also listed in the table. Besides those changes more changes have been made according to the problems that users' experienced. These are listed in Appendix A.

Table 9: Recommendations for improving the Council of Coaches functional demonstrator of the participants in the first round, and changes made in the system as a result of their recommendations.

First round		
Recommendations	**Number of times mentioned**	**Changes made in system after the first round, changes that are on the list, or comments to why it does not fit the scope of this product**
Easier log in	6	Improved the error messaging after unsuccessful login attempts. Added an icon to the login password field that allows you to see the password you're typing. Login email address is no longer case sensitive.
More content	3	More content has been added throughout and after the first round. See Appendix A
More personalized advices	3	Throughout the process of defining new content for the application we have always tried to include forms or personalization wherever possible.
More depth in the dialogues	3	More content has been added throughout and after the first round, with also more depth. See Appendix A
To the point advices	2	More advices have been added throughout and after the first round. See Appendix A
Less social talk	2	We consider having social conversations with the coaches to be a core part of the concept. Still, two strategies are in place to address these concerns: (1) add more "serious" content (so the ratio between "social" and "serious" improves), and (2) make sure that social conversations can always be skipped for those that do not appreciate it.
Visual material	2	The focus for the proof of concept study has always been on dialogue-based interaction. We realize the strength and potential of multi-media use, mixing dialogue with visual representations and experimented with this in the Activity Book and Recipe Book widgets. Development time for such features is unfortunately high, so no additional similar features could be implemented.

Faster operation of system	2	From our own internal testing we have noticed good performance on high-end tablet devices (e.g. newer iPad models) and sometimes poor performance on low-end tablet devices. There are two causes for this, and both could be addressed in the next project phase (product development). The first is related to network calls: currently every interaction step requires a call to our server to retrieve the next step (and update internal models of the coaches). The underlying technology platform (WOOL) already supports fully executing dialogues in the client, so network calls are only needed between separate dialogues (e.g. topic switches). However, this would require the server architecture to be updated to support this mechanism as well, which is a relatively large amount of work. The second performance impact is generated by the rendering of graphics on the client device. All graphics in Council of Coaches are Scalable Vector Graphics (SVG), which means that everything will look crisp on a 6" phone screen, and just as crisp on a 70" TV screen. However, client devices will need more effort in rendering these graphics than when using bitmap graphics (e.g. PNGs). The solution that is on the "design table" is to include a settings menu in which graphics quality can be switched from low to high (which is easy), and to include a set of graphic resources where details are toned-down (which is a lot of work).
More interactive (according to the current situations: weather conditions, other problems)	2	Coda gives information about the COVID-19 situation. Integration of a weather-service to personalize advice is on the to-do list.
Speech control system instead of written dialogues	1	We are investigating ways to integrate speech control into the functional demonstrator. The technical basics are simple, but the implications for the human-computer interactions are not. Speech input without speech output (i.e. spoken voices of the coaches) is very awkward. Text to speech in 8 different voices that match the different coach characters (as well as coda) is one of the accompanying issues that need to be solved. Besides that, speech-to-text (for input) does not always work well for older adults, as well as increases the need for a good network connection (audio is sent to server, text is sent back). So, this feature is currently in "R&D" but, in our expectations would negatively impact the usability of the system if not implemented well.
Content about sleeping problems	1	A lot of additional dialogue content was added for Rasmus, who now has new dialogues in which he coaches on sleep: information and quizzes.
Option to print out a recipe	1	This is a nice to have at best. Technically not as easy as it sounds, especially for the wide range of devices that Council of Coaches supports. Instead of providing a print option, we have considered to support two additional options for the recipe book that are meant to provide additional assistance to the preparation of recipes. The first is a mobile app integration that allows you to store

		the list of ingredients into a digital shopping list. The second is to provide a recipe-book mode that lets Francois guide the user through a recipe step-by-step. Both features would require a lot of development time and are saved for the next phase of the project.
More stringent coaching	1	Additional personalizing in the coaching is an overarching aim for the content of Council of Coaches, which includes different coaching styles for different individuals.
Showing where's new content.	1	This is a good idea that was not considered during this evaluation phase. We considered the ongoing evaluation to be a "running beta" version, where updates where happening all the time and all over the place. After an official release, a better "content release plan" should be drafted, whereby regular (e.g. bi-monthly) updates should focus on a specific area of the application, and include some type of messaging to point out to users where this new content can be found. We look, as more often, to video game design for inspiration on how this can be solved in future updates and consider a pop-up like in Figure 8 below (each panel in the image can be interacted with to provide more info on where to find this new content).  **Figure 8: Example of "New Content Highlight" screen from a popular video game (World of Warcraft).**
Instead of real conversations with coaches: having a menu of which you can choose a topic, and directly diving into this topic.	1	This is not in the "spirit of the Council of Coaches". Although we do recognize the desire for users to quickly achieve certain tasks, we look to the design of the "menu" dialogues, and have made changes to streamline these.
Making a version for primary school children with problems they experience	1	The Council of Coaches concept can be applied in many domains, and we strongly believe that it will in the coming years.

Table 10: Recommendations for improving the Council of Coaches' functional demonstrator of the second-round participants.

Second round	
Recommendations	**Number of times mentioned**
More personalized advices	5
Option to ask questions	3
Keeping a food diary	2
More initiative from the coaches	2
Less social talk	2
Reminders	1
Understandable vocabulary for everyone	1
More attractive lay-out	1
More content	1
Advices based on performed activities	1
More interesting, different coaches like François	1
To the point advices	1
Easier log in	1
Easier accessible	1
Less childish, not too many explanations	1
System needs to remember log in credentials	1
More real look-a-like coaches	1
Visual material	1
More active, compelling coaching	1
Turning it in a kind of general practitioner for social distant older adults	1

In the end of the interviews, participants were asked whether they have any remaining things they want to share with us about the functional demonstrator. Positive things they mentioned were about:

- The recipe book (N=4): "*Nice recipe book. It was concrete, and I could retrieve really great dishes from it.*" (P-10)
- The appearance (N=4): "*I really like the appearance of the system I think the creation of the system is really good.*" (P-1)
- The system as a whole (N=4): "*I think the website is fantastic.*" (P-9) and "*The intent of the application is really good, and I think it could work well for people.*" (P-25)

- The coaches: (N=3): "*They never say: 'You are dumb', or something like that. So that is why I find them friendly*." (P-27)
- The content (N=3): "*I liked Helen's brain teasers. Sometimes, I had to think hard about it.*" (P-37)
- You as a user having control (N=1): "*Each time you have a choice whether you want to continue or stop.*" (P-27)
- A specific function within the system (N=1): "*Last time I logged in, I asked a question to one of the coaches, and then Emma joined the conversation. I thought: 'Oh well, this is a nice function.'*" (P-5)
- The radio (N=1): "*The radio is nice to listen to.*" (P-30)

Besides those, some comments were less positive. Mostly about the content (N=13 (first round) and N=10 (second round)): too little, too simple, not-personalised, low level/childish, wrong answer options, unilateral stories. Other comments were about: too robot-like, the music playing on the radio, the lay-out, François' attitude, not staying logged in, social talks, asking questions.

4.4 Potential health effects

Table 11 shows the mean scores on quality of life, positive health domains, SMAS-s domains, and total SMAS-s score at T0, T1 and T2. These scores are divided into the previous mentioned two use groups. The scores of the positive health domains and of the SMAS-s domains at T0, T1 and T2 are plotted in spider plots (see Figure 9 and Figure 10).

A normality test has been conducted, which showed that the health variables were non-normally distributed. So, a non-parametric test was used to test whether there is a significant difference in health outcomes after using the functional demonstrator. The Friedman test showed that in the first round for both use groups, there was one health variable with a significant difference. For use group A, it was the variable perceived health state measured on a Visual Analogue Scale (VAS) (Chi^2=6.6; df=2; p=0.04). Participants scored themselves highest at T2 (mean rank = 2.60), and lowest at T1 (mean rank = 1.10). At T0, the mean rank score was 2.30. The Wilcoxon Signed-rank test, showed that the difference in this score was significant between T1 and T2 (Z=-2.02, p=0.04). At T2 all five participants in this group had a higher VAS score on their perceived health state, compared to at T1. However, the Holm-Bonferroni correction showed that there were no significant results between the groups.

For use group B, the Friedman test showed that there was a significant difference in mental health (Chi^2=10.6; df=2; p=0.005). Participants scored themselves highest at T1 (mean rank = 2.43), and lowest at T0 (mean rank = 1.47). At T2, the mean rank score was 2.10. The Wilcoxon Signed-rank test, showed that the difference in this score was significant between T0 and T1 (Z=-2.09, p=0.04). At T1, 10 out of 15 participants had a higher mental health score compared to at T0. One participant had a lower score at T1 compared to T0, and between the 4 left participants there was no difference between T0 and T1. However, again the Holm-Bonferroni correction showed that there were no significant results between the groups.

In the second round, the Friedman test showed one health variable in use group A and three variables in use group B with significant differences over time. For use group A, it was the self-efficacy domain of the SMAS (Chi^2=7.8; df=2; p=0.02). Participants scored themselves highest at T0 (mean rank = 2.35), and lowest at T1 (mean rank = 1.40). At T2, the mean rank score was 2.25. The Wilcoxon Signed-rank test, showed that the significant difference in this score was between T0 and T1 (Z=-2.401, p=0.016), and between T1 and T2 (Z=-2.156, p=0.031). At T1, there were 7 participants that scored lower compared to at T0, and 3 participants that scored equal. At T2, there were 6 participants that scored higher compared to at T1, 1 that scored lower, and 3 that scored the same. The Holm-Bonferroni correction showed that there was only a significant difference in self-efficacy between T0 and T1 (0.05/3=0.017 > 0.016).

For use group B, the domains taking initiative and positive frame of mind of the SMAS, and the total SMAS score had a significant difference over time. Regarding the domain taking initiative (Chi^2=6.2; df=2; p=0.045), participants scored highest at T2 (mean rank = 2.46), and lowest at T0 (mean rank = 1.54). The mean rank score at T1 was 2.00. The Wilcoxon Signed-rank test showed this difference was between T0 and T2 (Z=-2.354, p=0.019). At T2, 8 participants scored higher compared to at T0, 1

participant scored lower, and 3 participants scored the same as at T0. However, the Holm-Bonferroni correction showed that there were no significant results between the groups.

Second, regarding the positive frame of mind domain (Chi^2=7.0; df=2; p=0.031), participants scored highest at T2 (mean rank = 2.33), and lowest at T0 (mean rank = 1.46). The mean rank score at T1 was 2.21. The Wilcoxon Signed-rank test showed that the significant difference found was between T0 and T2 (Z=-2.401, p=0.016). Seven participants scored higher at T2, and for five participants the score was the same at T0 and T2. The Holm-Bonferroni correction showed that there was a significant difference between T0 and T2 (0.05/3=0.017 > 0.016).

Finally, the total SMAS score in use group B (Chi^2=9.2; df=2; p=0.010) had a significant difference. Participants scores themselves highest at T2 (mean rank = 2.54), and lowest at T0 (mean rank = 1.33). At T1, the mean rank was 2.13. The Wilcoxon Signed-rank test, showed that the difference in this score was significant between T0 and T1 (Z=-2.121; p=0.034), and between T0 and T2 (Z=-2.910, p=0.004). Regarding the scores between T0 and T1, 9 participants had a higher score at T1, and 3 participants had a higher score at T0. Regarding the scores between T0 and T2, 11 participants had a higher score at T2, and 1 participant had a higher score at T0. The Holm-Bonferroni correction showed that there was a significant result between T0 and T2 for the total SMAS score (0.05/3=0.017 > 0.004).

Table 11: Mean (SD) of health variables at T0, T1 and T2 in the first round and second round, divided into participants that used the functional demonstrator for less than 4 times (use group A) and that used the functional demonstrator for at least 4 times in the implementation phase (use group B).

	First round					
	Use group A			*Use group B*		
	M (SD) at T0 (N=9)	*M (SD) at T1 (N=9)*	*M (SD) at T2 (N=5)*	*M (SD) at T0 (N=15)*	*M (SD) at T1 (N=15)*	*M (SD) at T2 (N=15)*
Perceived health state	0.80 (0.16)	0.80 (0.12)	0.90 (0.09)	0.83 (0.15)	0.86 (0.14)	0.86 (0.15)
Perceived health state on a VAS	77.0 (14.6)	73.7 (13.0)	79.6 (11.3)	78.5 (11.1)	82.2 (15.0)	82.4 (10.0)
Positive health domains						
Bodily functions	6.9 (2.0)	6.9 (1.7)	7.8 (1.3)	7.0 (1.3)	7.5 (1.2)	7.7 (1.0)
Mental health	7.0 (1.7)	6.8 (1.1)	7.4 (1.1)	7.7 (1.2)	8.4 (1.0)	7.1 (1.1)
Meaning	7.1 (2.3)	7.3 (1.9)	7.4 (1.5)	8.1 (1.1)	8.5 (0.8)	8.3 (0.8)
Quality of life	7.0 (1.7)	6.6 (1.8)	7.2 (1.1)	7.9 (1.3)	8.4 (1.0)	8.4 (1.1)
Social participation	7.2 (1.6)	6.9 (1.7)	7.4 (1.1)	8.1 (1.1)	8.5 (1.1)	8.5 (1.1)
Daily routine	8.3 (1.0)	8.3 (1.0)	8.6 (1.1)	8.1 (1.4)	8.4 (1.5)	8.6 (1.1)
SMAS-s domains						
Taking initiatives	63.0 (12.1)	63.7 (10.1)	56.0 (12.1)	66.7 (15.3)	72.4 (13.8)	76.9 (14.7)
Investment behaviour	68.1 (12.4)	67.4 (10.8)	70.7 (13.0)	72.0 (11.6)	76.0 (14.4)	80.4 (13.4)
Variety	63.0 (14.6)	61.5 (17.9)	53.3 (6.7)	72.4 (19.0)	64.4 (18.5)	68.4 (18.1)
Multifunctionality	64.4 (17.0)	67.4 (14.3)	57.3 (3.7)	62.2 (13.9)	60.0 (18.3)	66.2 (18.1)
Self-efficacy	68.9 (9.4)	62.2 (15.6)	60.0 (4.7)	76.9 (14.4)	72.9 (15.6)	79.1 (14.2)
	60.0 (13.7)	61.5 (20.8)	58.7 (12.8)	56.0 (18.1)	61.3 (16.6)	58.7 (15.4)

Positive frame of mind						
SMAS-s total score	64.6 (9.2)	64.0 (11.0)	59.3 (5.7)	68.4 (10.2)	67.9 (11.7)	71.6 (10.1)
	Second round					
	Use group A			*Use group B*		
	M (SD) at T0 (N=12)	*M (SD) at T1 (N=11)*	*M (SD) at T2 (N=10)*	*M (SD) at T0 (N=12)*	*M (SD) at T1 (N=12)*	*M (SD) at T2 (N=12)*
Perceived health state	0.81 (0.12)	0.82 (0.12)	0.83 (0.11)	0.87 (0.18)	0.86 (0.20)	0.88 (0.21)
Perceived health state on a VAS	73.6 (16.2)	76.3 (9.7)	79.9 (11.1) [a]	83.2 (19.6)	82.5 (17.5)	82.0 (18.6)
Positive health domains						
Bodily functions	6.7 (1.3)	7.1 (1.4)	7.2 (1.5) [a]	7.3 (2.0)	7.6 (1.8)	7.3 (2.1)
Mental health	7.3 (1.6)	7.8 (1.2)	7.9 (1.3) [a]	7.9 (1.3)	7.7 (1.9)	8.0 (1.5)
Meaning	7.5 (1.8)	8.0 (1.4)	8.0 (1.3) [a]	7.9 (1.4)	7.7 (2.2)	8.2 (1.6)
Quality of life	7.4 (1.7)	8.0 (1.5)	8.0 (1.6) [a]	8.2 (1.3)	8.2 (1.5)	7.9 (1.4)
Social participation	7.6 (1.5)	8.3 (2.1)	8.0 (1.7) [a]	7.8 (1.4)	7.8 (1.9)	7.8 (1.4)
Daily routine	7.6 (1.3)	7.9 (1.4)	7.7 (1.3) [a]	8.3 (1.5)	8.7 (1.5)	8.5 (1.6)
SMAS-s domains						
Taking initiatives	71.7 (12.1)	70.9 (7.5)	73.3 (11.8)	58.3 (16.1)	62.8 (16.2)	66.1 (19.2)
Investment behaviour	73.9 (10.0)	70.3 (11.7)	73.3 (11.8)	62.8 (21.0)	67.2 (17.4)	71.1 (17.1)
Variety	56.1 (16.2)	57.6 (15.2)	65.3 (18.5)	52.2 (20.1)	58.3 (14.2)	62.2 (13.7)
Multifunctionality	63.9 (12.5)	62.4 (16.4)	62.0 (16.0)	57.8 (14.0)	59.4 (11.5)	58.9 (10.9)
Self-efficacy	72.2 (13.0)	65.5 (13.6)	73.3 (13.7)	61.1 (15.0)	66.1 (16.4)	65.6 (17.3)
Positive frame of mind	61.1 (19.5)	63.6 (16.4)	58.7 (16.3)	57.8 (16.7)	62.8 (21.5)	65.6 (16.3)
SMAS-s total score	66.5 (10.4)	65.1 (10.2)	67.7 (11.9)	58.3 (13.4)	62.8 (12.5)	64.9 (11.9)

[a] N=9

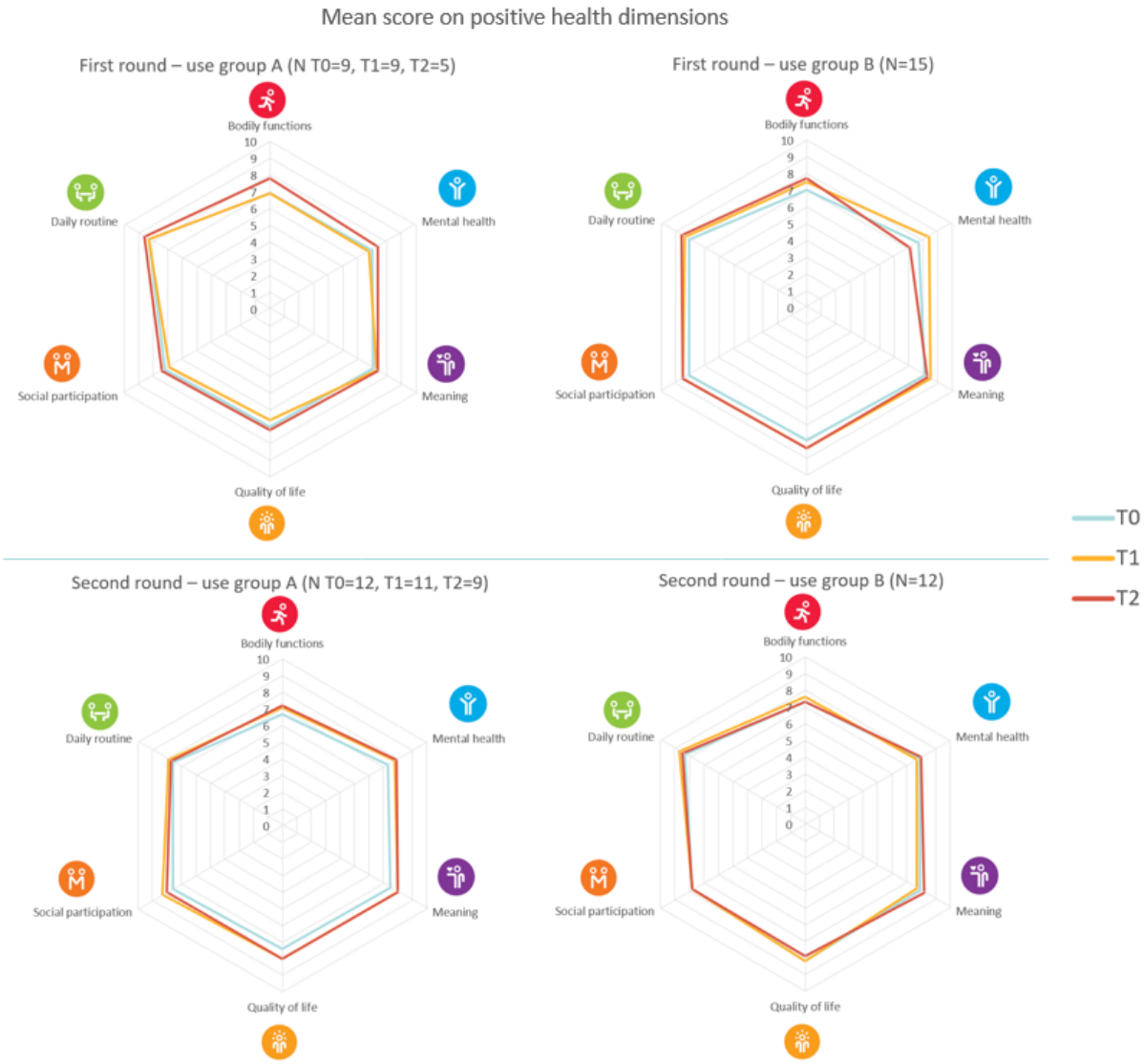

Figure 9: Spider plots mean score on positive health dimensions, divided into first and second round and participants that used the functional demonstrator for less than 4 times (use group A) and that used the functional demonstrator for at least 4 times in the implementation phase (use group B).

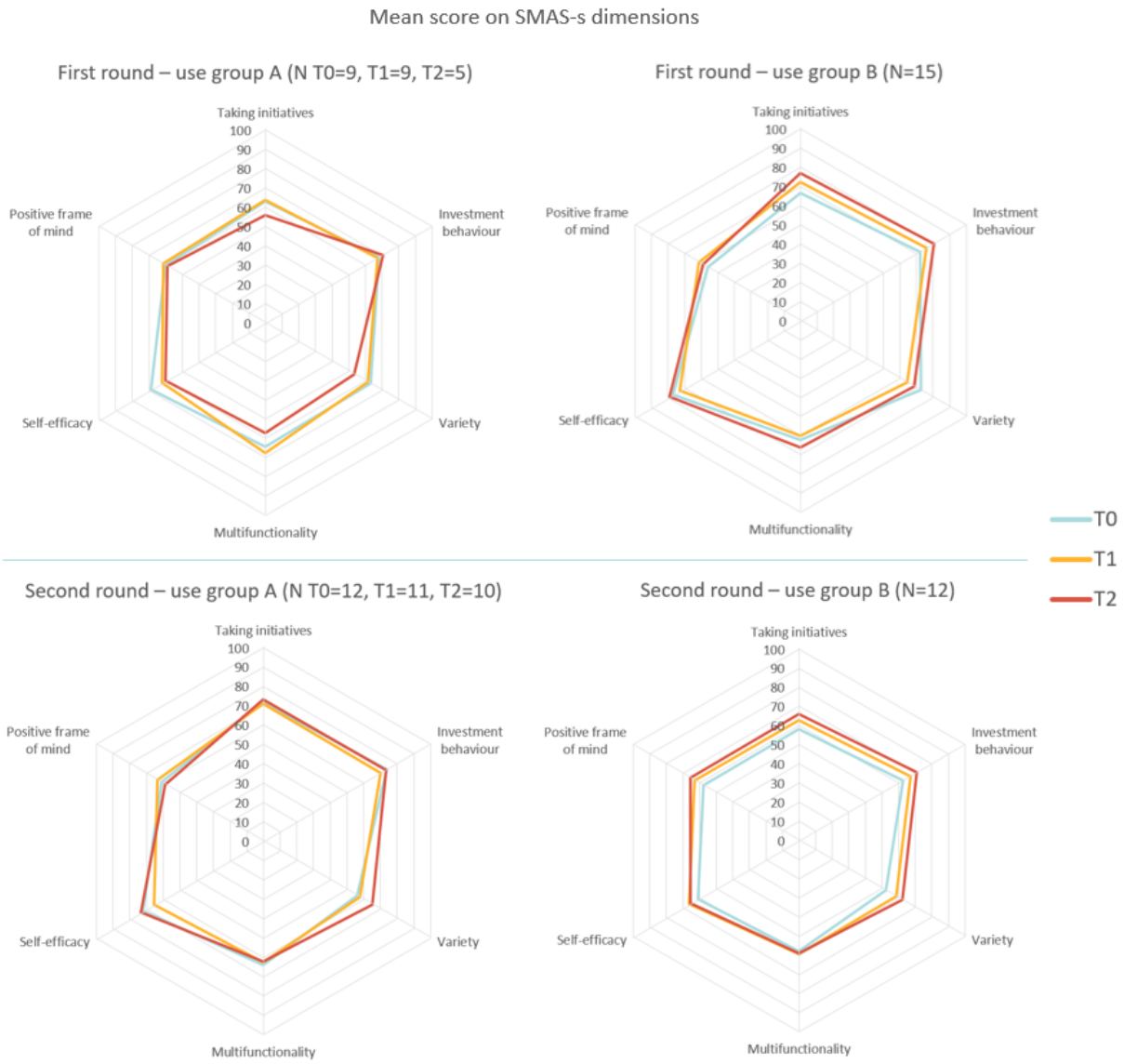

Figure 10: Spider plots mean score on SMAS-s dimensions, divided into first and second round and participants that used the functional demonstrator for less than 4 times (use group A) and that used the functional demonstrator for at least 4 times in the implementation phase (use group B).

4.5 Applicability of the virtual coaches

During both rounds, both primary coaches, Olivia and François, were rated poor among their working alliance. The group of participants that used the functional demonstrator for at least four days during the implementation phase, scored both coaches higher than the group that used it for less than four days. In use group B, both coaches scored highest among the domain Bond, with a mean of 2.2 (SD Olivia=0.9 and SD François=1.1) in the first round, and 2.6 (SD=1.3) for Olivia and 2.4 (SD=1.1) for François in the second round. Looking at the mean total score of their working alliance, François scored slightly higher than Olivia in both use groups in the first round and only in use group A for the second round (see Table 12).

Table 12: Mean (SD) of domains of working alliance of Olivia and François in the first round (N=23) and second round (N=23), divided into participants that used the functional demonstrator for less than 4 times (use group A) and that used the functional demonstrator for at least 4 times in the implementation phase (use group B).

	First round (N=23)				**Second round (N=23)**			
	Use group A (N=8)		*Use group B (N=15)*		*Use group A (N=11)*		*Use group B (N=12)*	
	M (SD) Olivia	*M (SD) François*	*M (SD) Olivia*	*M (SD) François*	*M (SD) Olivia*	*M (SD) François*	*M (SD) Olivia*	*M (SD) François*
Task	1.2 (0.4)	1.3 (0.5)	1.6 (0.6)	1.8 (0.9)	1.5 (0.6)	1.6 (0.8)	1.9 (1.0)	1.9 (0.7)
Bond	1.3 (0.5)	1.3 (0.5)	2.2 (0.9)	2.2 (1.1)	1.9 (0.8)	1.9 (0.8)	2.6 (1.3)	2.4 (1.1)
Goal	1.2 (0.4)	1.3 (0.5)	1.7 (0.6)	1.8 (0.9)	1.7 (0.7)	1.7 (0.8)	2.3 (1.3)	2.1 (0.9)
Total score	1.2 (0.4)	1.3 (0.5)	1.8 (0.6)	1.9 (0.9)	1.7 (0.7)	1.8 (0.8)	2.3 (1.2)	2.1 (0.8)

4.5.1 Satisfaction with the virtual coaches

At T0 of the first round, Emma scored highest on average, and Olivia scored lowest on average (see Table 13 and Figure 11). At T1, François scored highest on average, followed by Olivia. Results of the Related-Samples Wilcoxon Signed-Rank tests show that for every coach (except for Rasmus and Katarzyna, which were not tested due to a low n), the average satisfaction score dropped from T0 to T1.

In the second round, Helen scored highest on average at T0, and Rasmus scored lowest on average (see Table 13 and Figure 12). At T1, Rasmus scored highest on average, and Carlos scored lowest. Results of the Related-Samples Wilcoxon Signed-Rank tests show that for Helen and Carlos the average satisfaction score dropped from T0 to T1. For Olivia, François and Emma, no significant result was found.

Table 13: Results of the Related-Samples Wilcoxon Signed-Rank tests testing for a difference in mean satisfaction score for every coach at T0 and T1 given by the Dutch participants.

	First round			**Second round**		
	M(SD) at T0 (N=23)	*M (SD) at T1 (N=23)*	*p*	*M(SD) at T0 (N=24)*	*M(SD) at T1 (N=24)*	*p*
Olivia	6.4 (1.9)	4.6 (2.1)	*0.01*	6.2 (1.8)	5.3 (2.5)	0.07
François	6.7 (1.5)	5.3 (2.5)	*0.03*	6.3 (1.7)	5.7 (2.5)	0.20
Emma	7.4 (1.2)	4.0 (2.3)	*<0.001*	6.0 (2.1)	4.9 (2.6)	0.07
Helen	6.9 (2.0)	3.7 (2.4)	*<0.001*	6.4 (1.7)	5.1 (2.4)	*0.02*
Carlos	6.5 (1.5)	3.4 (2.5)	*<0.001*	6.2 (1.6)	4.3 (2.1)	*0.001*
Rasmus	6.7 (0.6) [c]	1.0 (-) [a]	-	5.4 (2.6) [f]	6.5 (0.7) [b]	-
Katarzyna	6.7 (1.2) [c]	1.5 (0.7) [b]	-	6.0 (2.1) [e]	4.5 (2.9) [d]	-

[a] N=1 [b] N=2 [c] N=3 [d] N=4 [e] N=12 [f] N=15

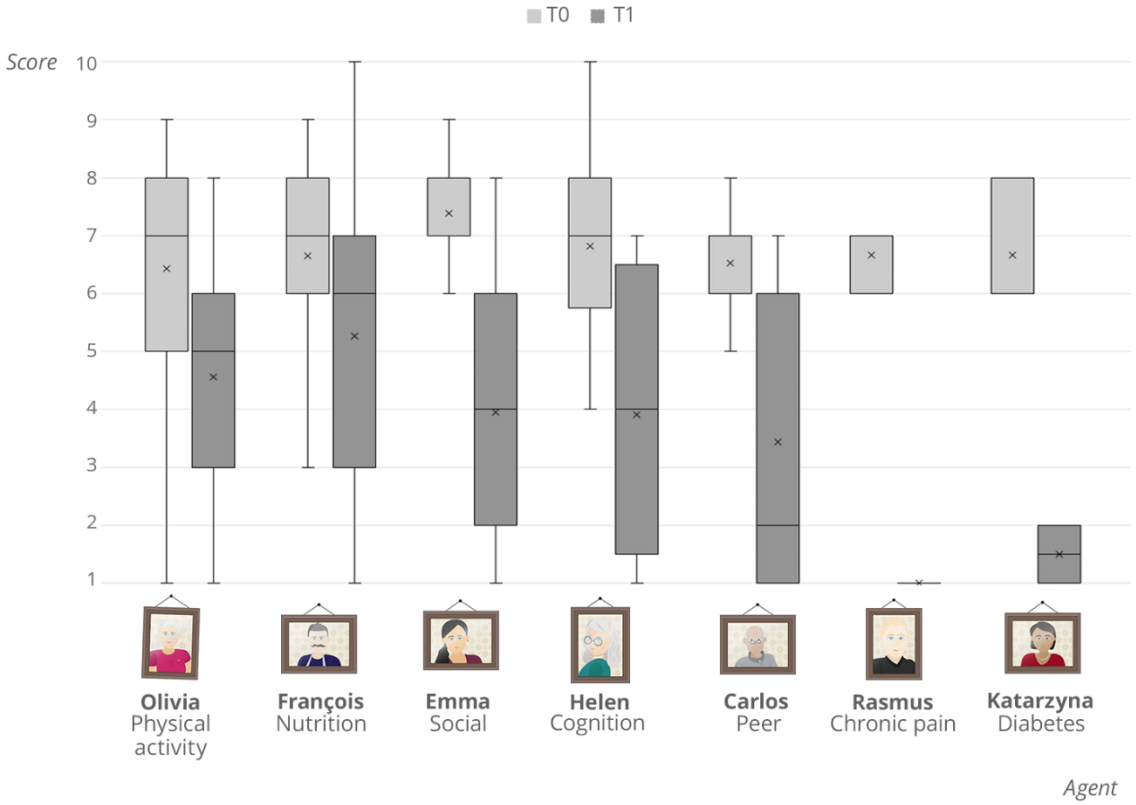

Figure 11: Boxplot showing the satisfaction scores of every coach at T0 and T1 given by the Dutch participants of the first round.

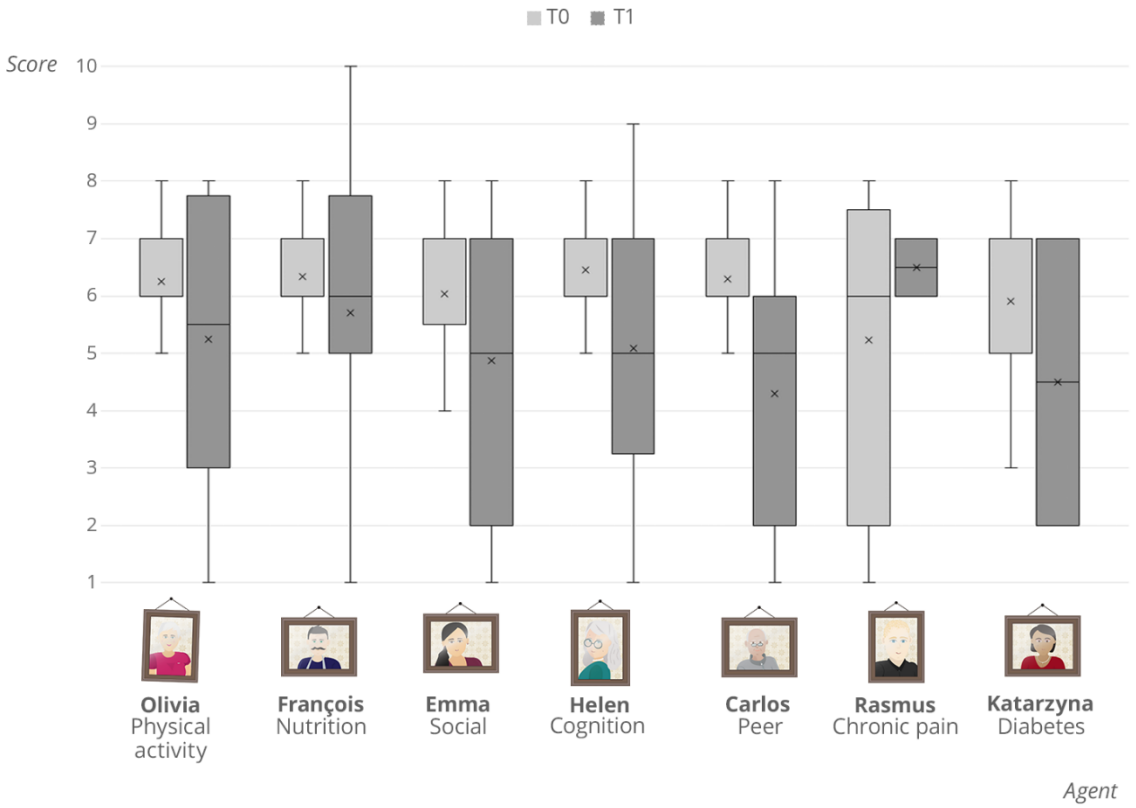

Figure 12: Boxplot showing the satisfaction scores of every coach at T0 and T1 given by the Dutch participants of the second round.

In the interviews of the first round, many participants (N=7) did not prefer one ECA the most. Some found it difficult to make a judgment because they did not interact much. François was preferred most often (N=9). Reasons given related to his content (N=5): providing concrete advice and interesting recipes. Two participants (N=2) particularly liked François' personality: friendly, pleasant, and funny. One participant preferred François, since he or she was interested in the domain of nutrition (N=1). In addition, Olivia was preferred often (N=7). A few participants were especially interested in the domain of physical activity (N=2). Another indicated he or she liked her content (N=1): concrete advice, realistic goals and valuable feedback. One participant (N=1) preferred both Olivia and Helen, liking their background stories. Carlos was preferred just once (N=1), by a participant that enjoyed his small talk.

During the interviews of the second round, 6 participants had no preference for one ECA the most. Olivia was preferred by most participants (N=8), of which 5 participants preferred her because she of the domain she focuses on (physical activity). Three participants preferred her because of her content: nice tips, good suggestions. Furthermore, François was also preferred often (N=6), 3 times because of his content: nice, delicious recipes, and one time because of his personality: funny. Two participants indicated they prefer him the most because both his content and personality: good suggestions, bit different than other coaches, funny jargon. Helen was preferred by two participants, one because of her domain cognition, and one because of her content: learning more about brains. Furthermore, there was one participant who liked all the coaches equally, and one who preferred bot Olivia and François because of their domains (physical activity and nutrition).

The majority of the participants (N=9) in the first round did not prefer one coach the least, since they believed they all had limited content. Two found all coaches equally sweet (N=2). The least preferred was François (N=5). Reasons given included: not liking his background story (N=1) and his content being little concrete, focusing on small talk (N=2). One participant was not interested in the domain nutrition (N=1) and another (N=1) found François' personality annoying, conveyed via stereotype and popular French words. Also, many participants least preferred Carlos (N=3), because of a lack of content (N=3), focusing on small talk instead of concrete goals. Only one participant least preferred Emma (N=1), since she had little content. Just one participant preferred Olivia the least (N=1), not liking her background story.

During the second round, the majority of the participants (N=11) did not prefer one ECA the least. The least preferred coach was Carlos (N=6), all because of his content: no content (N=5), interfering in other conversations (N=1). One participant did not like both Carlos and Helen the most, because of their content: both interfered in other conversations. Furthermore, Olivia, François, Emma and Katarzyna were also mentioned once as ECA who was preferred least. The reason for mentioning Olivia, was because the participant was not interested in her domain. François was mentioned because he was talking about the radio, and once he had no time to give tips. The reasoning behind mentioning Emma was because of her content and domain: not interested in social domain and talking about football and dogs. Finally, Katarzyna was mentioned because of her lack of content.

A more detailed analysis of first round participants' preferences for particular coaches is described in:

ter Stal, S., Hurmuz, M., Jansen-Kosterink, S., Beinema, T., Bulthuis, R., op den Akker, H., Hermens, H., Tabak, M. *Preferences of Older Adults for Embodied Conversational Agents in a Multi-Agent eHealth Application.* (2020). ACM International Conference on Intelligent Virtual Agents (Submitted)

5 Results – Scotland

5.1 Demographics

During the first round, 19 participants were included. Two participants dropped out during this round. The study population of the first round consisted of more women than man (57.9% vs. 42.1%), with a mean age of 63.7 years (SD=7.0). In the second round, 22 participants were included, of which 68.2% were female. The mean age of the participants of this round was 59.5 years old (SD=8.9). Table 14 shows all demographics of both rounds separately.

Table 14: Demographics of the study population in Scotland of the first (N=19) and second round (N=22)

Demographic	First round (N=19)	Second round (N=22)
Gender (%) Male Female	 42.1 57.9	 31.8 68.2
Age (M (SD))	63.7 (7.0)	59.5 (8.9)
Level of education (%) Preparatory secondary vocational education Higher general secondary education, pre-university education Higher vocational education, university	 21.1 5.3 73.7	 4.5 31.8 63.6
Living situation (%) Alone Married/living together	 21.1 68.4 10.5	 9.1 86.4 4.5
Work status (%) Employed Volunteer/caregiver Retired Other	 31.6 5.3 47.4 15.8	 45.5 0 50 4.5
Health literacy (M (SD))	4.8 (0.4)	4.7 (0.4)
Self-reported level of physical activity (%) Not at all Not at all, but thinking about beginning Less than 2.5 hours a week More than 2.5 hours a week in the last six months More than 2.5 hours a week for more than six months	 10.5 5.3 36.8 10.5 36.8	 4.5 9.1 31.8 22.7 31.8
Attitude towards technology (M (SD))	4.7 (1.5)	4.4 (0.7)
Motivation type to live healthy (M (SD)) Intrinsic motivated External regulated motivation A-motivated	 5.4 (0.9) 3.4 (1.3) 2.4 (1.5)	 5.1 (1.0) 3.3 (1.1) 2.3 (1.1)

5.2 Use of Council of Coaches' functional demonstrator

Table 15, Table 16 and Figure 13 below show for each round in both the implementation and facultative phases how participants used the functional demonstrator. Figure 14 shows an overview of the number of participants per phase and how many participants stopped using the functional demonstrator during the study period. In the first round, 17 people used the functional demonstrator during the

implementation phase and 12 used it in the facultative phase. The mean number of days participants used it in the implementation phase was 3.2 (SD=2.1) while the mean number of sessions was 3.5 (SD=2.8). During the facultative phase, the mean number of days participants used the functional demonstrator for was 2.2 (SD=1.3) as was the number of sessions.

In the second round, 19 participants used the functional demonstrator in the implementation phase and 8 used it in the facultative phase. The mean number of days participants used the functional demonstrator in the implementation phase was 5.4 (SD=4.2) while the mean number of sessions was 5.7 (SD=4.6). During the facultative phase, the mean number of days participants used it was 1.9 (SD=1.2) while the mean number of sessions was 2.1 (SD=1.4).

In the first round, the highest mean number of sessions of 2.1 (SD=1.9) took place during week 4. The highest mean length of session of 10.4 minutes (SD=10.6) took place in week 5 and the highest mean number of interactions of 137.9 (SD=108.5) took place during week 2.

In the second round, the joint highest mean number of sessions of 2.7 (SD=1.6 first round, SD=1.3 second round) took place during weeks 2 and 5. The highest mean length of session of 8.2 minutes (SD=6.6) took place in week 2 and the highest mean number of interactions of 141.9 (85.9) took place during week 2.

Table 15: Use data of total system of both rounds: number of participants used the functional demonstrator per phase, mean (SD), min and max number of days used and sessions within the functional demonstrator.

First round					
Phase	*N*	*Mean (SD) number of days used COUCH*	*Range min-max number of days used COUCH*	*Mean (SD) number of sessions within COUCH*	*Range min-max number of sessions within COUCH*
Implementation phase	17	3.2 (2.1)	1 – 7	3.5 (2.8)	1 – 11
Facultative phase	12	2.2 (1.3)	1 – 5	2.2 (1.3)	1 – 5
Second round					
Phase	*N*	*Mean (SD) number of days used COUCH*	*Range min-max number of days used COUCH*	*Mean (SD) number of sessions within COUCH*	*Range min-max number of sessions within COUCH*
Implementation phase	19	5.4 (4.2)	1-15	5.7 (4.6)	1-18
Facultative phase	8	1.9 (1.2)	1-4	2.1 (1.4)	1-5

Table 16: Use data of total system of both rounds: number of participants used the functional demonstrator per week, mean (SD), min and max number of sessions per week, mean (SD), min and max duration in minutes per session per week, and mean (SD), min and max number of interactions per session per week.

First round							
Week	*N*	*Mean (SD) number of sessions*	*Range min-max*	*Mean (SD) duration in*	*Range min-max duration in*	*Mean (SD) number of*	*Range min-max number of*

			number of sessions	*minutes per session*	*minutes per session*	*interactions per session*	*interactions per session*
1	-	-	-	-	-	-	-
2	16	1.4 (0.5)	1 – 2	5.9 (4.7)	1.2 – 19.8	137.9 (108.5)	22 – 452
3	10	1.8 (1.3)	1 – 4	5.9 (4.9)	1.0 – 16.9	89.2 (73.1)	10 – 301
4	7	2.1 (1.9)	1 – 6	10.3 (12.7)	1.2 – 48.6	105.3 (77.4)	32 – 316
5	4	1.0 (0)	1 – 1	10.4 (10.6)	1.8 – 8.2	137.0 (81.4)	19 – 204
6	9	1.8 (0.8)	1 – 3	5.4 (4.4)	1.4 – 16.1	86.1 (41.7)	13 – 169
7	6	1.2 (0.4)	1 – 2	5.3 (3.0)	1.0 – 9.4	82 (57.6)	18 – 170
8	2	1.0 (0)	1 – 1	3.8 (3.1)	1.6 – 6.0	52.5 (46)	20 – 85
9	1	1.0 (0)	1 – 1	3.2 (0.0)	3.2 – 3.2	86 (0)	86 – 86
Second round							
Week	*N*	*Mean (SD) number of sessions*	*Range min-max number of sessions*	*Mean (SD) duration in minutes per session*	*Range min-max duration in minutes per session*	*Mean (SD) number of interactions per session*	*Range min-max number of interactions per session*
1	-	-	-	-	-	-	-
2	18	2.7 (1.6)	1 – 6	8.2 (6.6)	1.0 – 38.3	141.9 (85.9)	28 – 453
3	11	2.2 (1.2)	1 – 3	5.7 (3.4)	1.7 – 14.4	100.0 (48.5)	27 – 254
4	10	2.0 (1.3)	1 – 5	6.2 (3.5)	1.0 – 13.9	97.5 (55.1)	30 – 252
5	6	2.7 (1.3)	1 – 4	4.1 (1.8)	1.4 – 8.5	81.9 (39.4)	40 – 171
6	8	1.6 (0.9)	1 – 3	4.2 (2.4)	1.2 – 11.0	94.2 (47.7)	8 – 170
7	2	1 (0.3)	1 – 1	2.0 (0.6)	1.5 – 2.6	32.0 (4.0)	28 – 36
8	0	-	-	-	-	-	-
9	2	1 (0.3)	1 – 1	3.6 (0.5)	3.1 – 4.1	57.5 (2.5)	55 – 60

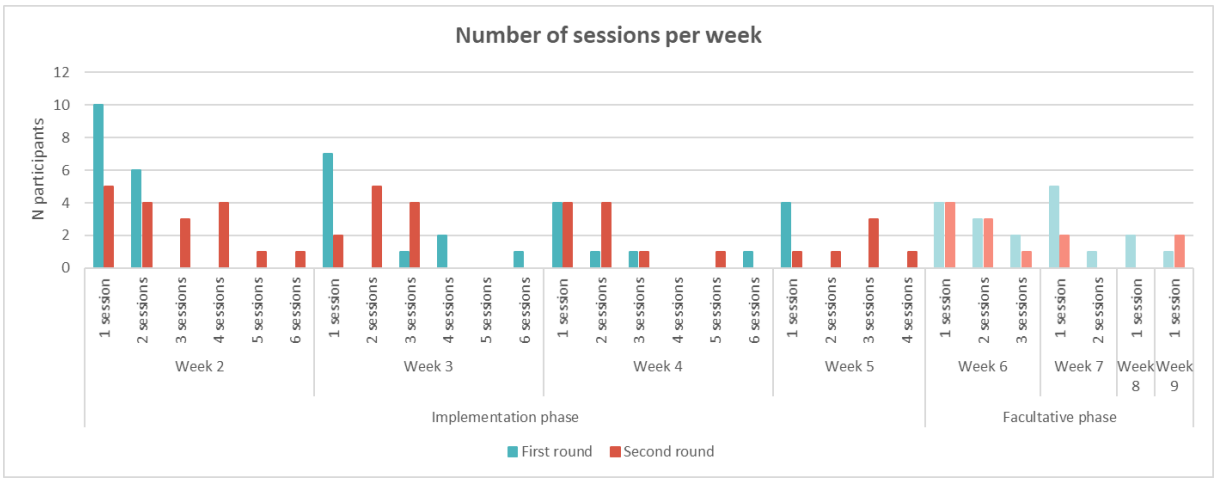

Figure 13: Frequency graph; number of participants vs. number of sessions within Council of Coaches' functional demonstrator per week.

Figure 14: Flowchart number of participants per phase and drop-outs.

5.3 User experience

The usability of the functional demonstrator was scored with a mean of 59.9 (SD=18.2, N=17) by the participants in the first round, which means the acceptability of the usability was marginally low. Seven of the 17 participants (41.2%) said they were willing to pay to use the functional demonstrator. Of the 7 participants who were willing to pay to use it each month, the average amount they indicated they would be willing to spend was £4.29 (€4.79) per month.

The usability of the functional demonstrator was scored with a mean of 64.5 (SD=18.0, N=15) by the participants in the second round, which means the acceptability of the usability was marginally high. For the second round, 1 out of the 15 participants (6.7%) said (s)he is willing to pay to use the functional demonstrator each month, indicating a price of £5 (€5.59) per month.

Regarding the user experience measured with the TAM, the first round participants were generally quite positive about the functional demonstrator (see Table 17 and Figure 15). Most of the participants were positive in 4 of the 7 domains. Trust in technology had the most positive results (M=5.4, SD=1.1) with 70.6% of participants positive. The joint highest percentage of participants who were negative was in the enjoyment and control domains (23.5%).

For the second round, user experience measured with the TAM was mostly quite neutral about the functional demonstrator. Most of the participants were neutral in 4 of the 7 domains while the other 3 domains were very close to being mostly neutral with 46.7% for each of them. The most positive domain was trust in technology (M=5.3, SD=1.0) with 53.3% of participants positive while the most negative was intention to use with 20.0%.

Table 17: User experience assessed on 7 domains in the first (N=17) and second round (N=15).

	First round (N=17)				
	M	*SD*	*% positive*	*% neutral*	*% negative*
Enjoyment	3.2	1.4	11.8	64.7	23.5
Aesthetics	5.1	1.2	58.8	41.2	0
Control	4.1	1.9	41.2	35.3	23.5
Trust in technology	5.4	1.1	70.6	29.4	0
Perceived usefulness	4.7	1.7	52.9	35.3	11.8
Perceived ease of use	4.6	1.7	52.9	29.4	17.6
Intention to use	4.6	2.0	47.1	35.3	17.6
	Second round (N=15)				
	M	*SD*	*% positive*	*% neutral*	*% negative*
Enjoyment	3.5	1.5	13.3	73.3	13.3
Aesthetics	4.9	1.2	40.0	60.0	0.0
Control	4.1	1.5	20.0	73.3	6.7
Trust in technology	5.3	1.0	53.3	46.7	0.0
Perceived usefulness	4.1	1.4	26.7	60.0	13.3
Perceived ease of use	4.4	1.7	40.0	46.7	13.3
Intention to use	3.9	1.9	33.3	46.7	20.0

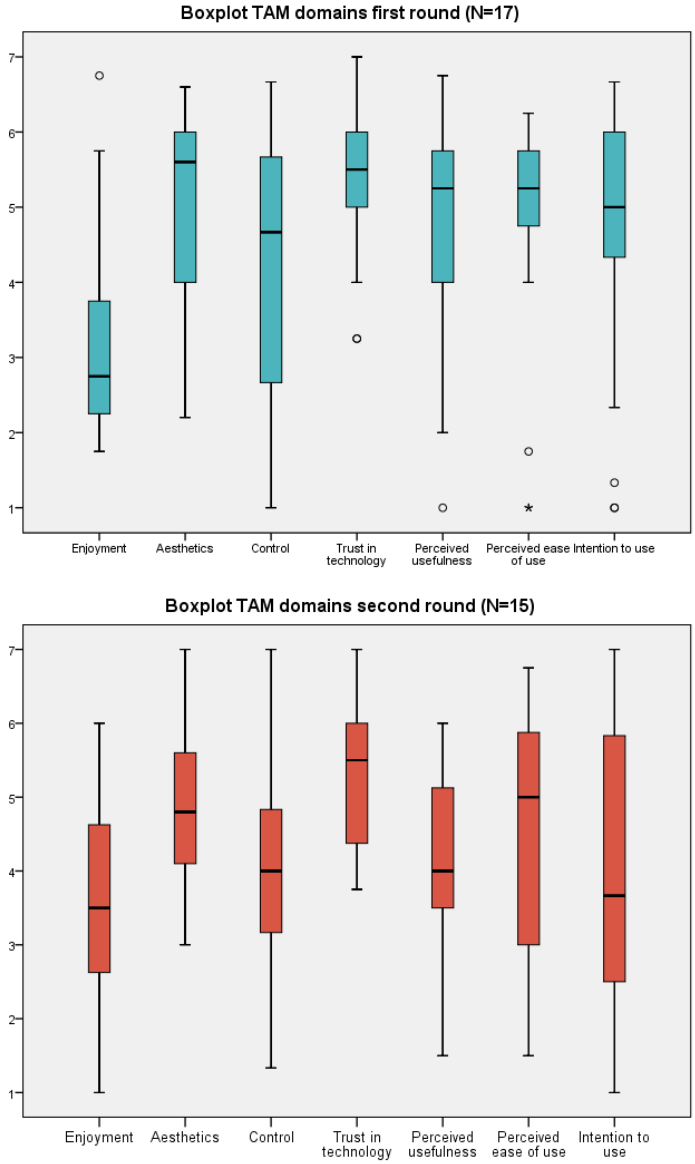

Figure 15: Boxplots user experience assessed by TAM in the first and second round.

In the first round, 15 participants completed the interview and 9 of them said they would recommend the functional demonstrator. In the second round, 14 participants completed the interview and 5 said they would recommend it. Some participants gave reasons for both recommending it and not recommending it. The reasons for recommending or not recommending the functional demonstrator are shown below in Table 18.

In the first round, the most common reasons for recommending the functional demonstrator were that it was convenient, it was easy to use and it was enjoyable. The most common reasons for not recommending it were that the technology was difficult to use, the coaching dialogue was repetitive, and the technology did not work properly.

In the second round, the most common reasons for recommending the functional demonstrator were to track the user's progress, to get good advice from the coaches and to get good recipes. The most common reasons for not recommending it were that the technology was difficult to use, the lack of content, not being able to ask your own questions, the coaching was patronizing, the technology was slow and the application was very time consuming.

Table 18: Reasons mentioned for (not) recommending the functional demonstrator to others during the first round and second round.

First round		
	Reasons for recommending	**Reasons for not recommending**
Mentioned 5 times or more		Technology was difficult to use
Mentioned 2 to 4 times	It was convenient	Coaching dialogue was repetitive
	It was easy to use	Technology did not work properly
	It was enjoyable	
Mentioned 1 time	Good for people who need specific advice	The coaching was patronizing
	You can repeat parts you did not understand	Does not add anything to an activity tracker
	Encouraged me to be more active	The interface was old fashioned
	Encouraged me to Improve my health	There was no option to ask your own questions
	It's a great idea	
Second round		
	Reasons for recommending	**Reasons for not recommending**
Mentioned 5 times or more		Technology was difficult to use
Mentioned 2 to 4 times	To track my progress	Lack of content
	Coaching advice was good	The coaching was patronizing
	The recipes were good	Technology was slow
		Application was very time consuming
		There was no option to ask your own questions
Mentioned 1 time	The application was easy to use	Didn't get the answers I was looking for
	To improve my health	Not enough personalisation
	It was encouraging	The application was irritating
	It is good for changing people's habits	Would be quicker looking up the information myself
	I liked the memory sessions	Not sure who it's aimed at

	It feels like you're speaking to the coaches	Recipes were not aimed at a UK audience
	It's a good idea	Technology didn't work
	It looks good	There were too many dialogue loops
	I lost weight	The text was too small
	I'm more aware of my fitness	No option to have diabetes and chronic pain coach
	I'm more aware of my nutrition	

Users' reasons for (not) using the functional demonstrator were coded with the six domains of positive health: bodily functions, mental health, meaning, quality of life, social participation, and daily routine (Huber, et al., 2016). Reasons that could not be coded with these domains, were coded as 'other reasons'. Figure 16 shows the percentages each domain was chosen as a reason to (not) use the functional demonstrator. In the first round, 28 reasons were given for using it and 11 not to. In the second round 20 reasons were given for using it and 15 not to.

In round 1, 19 of the reasons to use the functional demonstrator were for bodily functions, 4 for quality of life, 1 for meaning and the rest were coded as others. 3 of the reasons not to use it were for daily routine, the remaining 8 were coded as other. In round 2, 8 of the reasons to use the functional demonstrator were for bodily functions, 2 for quality of life and 1 for social participation. The remaining 9 were classified as others. 2 of the reasons not to use it were for daily routine while the remaining 13 were coded as other (see Figure 16).

In the first round, 10 participants indicated that the functional demonstrator was easy to use while 4 said it was difficult to use. One participant said it was too simplistic, while in total there were 10 problems identified. Four participants had difficult setting their accounts up, and 4 had problems syncing their Fitbit with the functional demonstrator. One each had problems saying that it simply did not work, it was slow, it was difficult to use on tablet or mobile and there was Fitbit data missing.

In the second round, 12 participants said it was easy to use while 2 found it difficult. There were 3 problems mentioned once each. There was a problem logging in, a problem with syncing the Fitbit and one participant found it slow.

The participants were asked for recommendations on how we could improve the functional demonstrator. A total of 32 recommendations were given in round 1 and 36 in round 2. The recommendations for the first round are given in Table 19 below, while the recommendations for the second round are in Table 20.

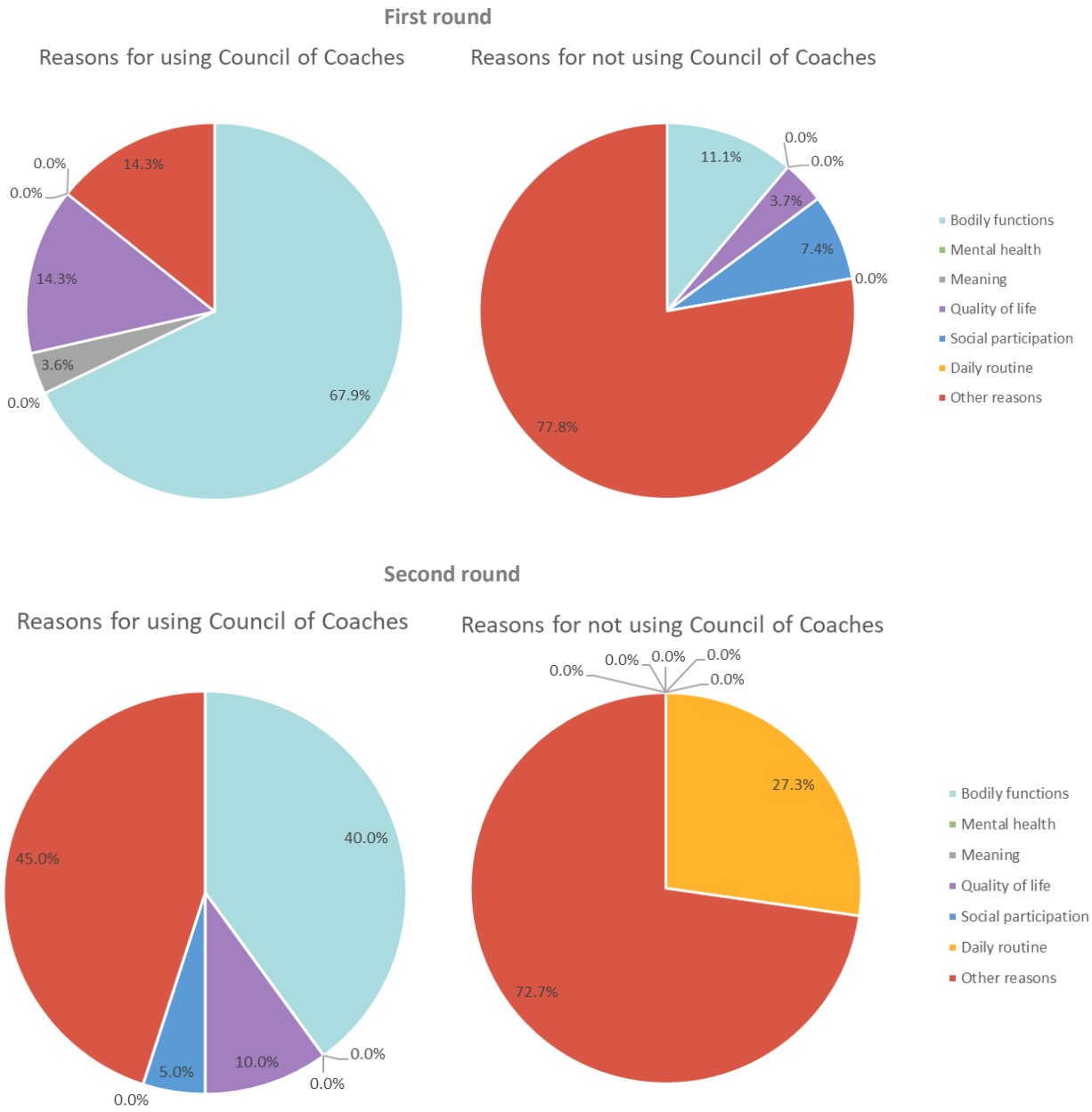

Figure 16: Reasons for (not) using Council of Coaches' functional demonstrator.

Table 19: Recommendations for improving the Council of Coaches' functional demonstrator of the participants in the first round, and changes made in the system as a result of their recommendations.

First round		
Recommendations	**Number of times mentioned**	**Changes made in system after the first round**
More content	7	More content has been added throughout and after the first round. See Appendix A.
Make the account creation process easier	3	The design philosophy for the account creation process is that it should be seamless, easy and does not "feel like a mandatory task". This design philosophy

		resulted in the procedure of creating an account in a conversational style with Coda the robot. While we are happy with this, we do realize it is still a barrier to entry – having to provide your email address before having had the opportunity to "try out" the application. Ideally and eventually we want to permit users to start using Council of Coaches without providing any information, and only later (when users are convinced that they want to continue using the service, store their information in a persistent account. In the short term however, this was out of scope to implement.
More personalized advice	3	Throughout the process of defining new content for the application we have always tried to include forms or personalization wherever possible.
More goal setting	2	Olivia's dialogues are improved to allow the user to set a long-term physical activity goal.
Improve the recipes	2	An enormous amount of effort has been put into the provision of a serious set of recipes to test this feature of the application. A total of 280 recipes, all tagged with nutrient and allergy information and translated from Dutch to English were available in the application since the start of the evaluations. We realize the most important component of any "recipe feature" is its database of recipes. In the next phase of the project, moving towards product development, this has to be taken up (i.e. by providing back-end tools for adding new recipes).
Make the system more intuitive	2	Interaction paradigms and user interface has been tested in several rounds, and improving how easy, intuitive and enjoyable the UI works is always a top priority.
You should be able to ask your own questions	1	This is an interesting suggestion but will need serious research and experimentation before being "production ready".
More songs on the radio	1	All the radio channels have been upgraded with brand new music. All four channels have a collection of 20 different tracks in their genre to play.
Dialogue is too simplistic	1	More content has been added throughout and after the first round. See Appendix A.

More feedback	1	More contents are continuously being added.
Improve the appearance of the coaches	1	The appearance of the coaches is part of one of the underlying research questions of the project, and thus could not have been changed during the project's evaluation.
1 polished coach would have been better than multiple coaches	1	No, but 3 even better polished coaches would have probably been better than the 7 coaches we offer now. The handcrafting of quality content for the coaches has proven to take a lot of time. As the project progresses, the tools have improved, and dialogue writers have gained more experience, content creation did go faster, but the point partly remains true that "less may be more" when it comes to the number of coaches provided.
Make coaches less stereotypical	1	This is very much a personal preference. From our experience, the most "stereotypical" coach (François) is at least the most often remembered, which is good. One could argue that François is the most memorable because he has the most "character".
Make it obvious that your Fitbit is not synced	1	The procedure of connecting the Fitbit with Olivia has been improved. When returning from the Fitbit website (after connecting your Fitbit), Olivia gives more detailed information on the success or potential failure of the procedure.
Make the coaches 3D	1	Three-dimensional coaches (and all the technical complexities and possibilities that come with it) are being investigated under the Agents United initiative (see www.agents-united.org).
Make the coaches speak	1	We are investigating ways to integrate speech control into the functional demonstrator. The technical basics are simple, but the implications for the human-computer interactions are not. Speech input without speech output (i.e. spoken voices of the coaches) is very awkward. Text to speech in 8 different voices that match the different coach characters (as well as coda) is one of the accompanying issues that need to be solved. Besides that, speech-to-text (for input) does not always work well for older adults, as well as increases the need for a good network connection (audio is sent to server, text is sent back). So, this feature

		is currently in "R&D" but, in our expectations would negatively impact the usability of the system if not implemented well.
More user support	1	Coda has slowly received updates to his content that provides help with various system features or device (e.g. Fitbit connections). Additional support content may be added in the future.
Fix the login problems	1	Improved the error messaging after unsuccessful login attempts. Added an icon to the login password field that allows you to see the password you're typing. Login email address is no longer case sensitive.
The coaches should move about	1	Yes, we agree that animations would "liven up" the application. We have worked on facial / communicative animations as a first step, but these were unfortunately not ready for testing during the final evaluation.

Table 20: Recommendations for improving the Council of Coaches' functional demonstrator of the second round participants.

Second round	
Recommendations	**Number of times mentioned**
More content	8
Needs better personalization	4
Give user notifications	3
Ask own questions	3
Coaches should get to the point	3
Coaches should be more realistic	2
Films would be better than coaching dialogue	1
Recipes should be more personalized	1
Needs a back button	1
Coaches should have more individual personalities	1
Goals should be set in smaller increments	1
More support	1

Would be better as a mobile app	1
Would prefer animated coaches	1
Heart rate should be discussed	1
Recipes should be written for UK audience	1
More encouragement	1
More interactive	1
Needs more depth	1

5.4 Potential health effects

Table 21 shows the mean scores on quality of life, positive health domains, SMAS domains, and total SMAS score at T0, T1 and T2. Figure 17 and Figure 18 show the spider plots of the positive health and SMAS domains. We had 19 participants at T0, 17 at T1 and 4 at T2 for the first round. There was an increase in perceived Health State throughout the study while there was a slight decrease in all the positive health domains from T0 to T1 before it increased again in T2. It should be noted that after T0 the UK went into lockdown due to COVID-19.

Because only 4 participants completed the T2-questionnaire, we will leave this out of the statistical analyses for assessing the potential health effects. 17 participants completed both T0 and T1. The Wilcoxon Signed-rank test, showed that there was a significant difference in two health variables. First, the mental health domain within positive health (Z=2.65, p=0.008). At T1, 10 participants had a lower mental health score compared to at T0. One participant had a higher score at T1, and between the left 6 participants there was no difference between T0 and T1 on mental health. Second, the Wilcoxon Signed-rank test, showed that there was a significant difference in the total SMAS-s score of T0 and T1 (Z=-2.69, p=0.007). At T1, 13 participants had a lower SMAS-s score compared to T0. Three participants had a higher score at T1, and between the left 3 participants there was no difference between T0 and T1 on the total SMAS-s score.

For the second round, we had 22 participants at T0, 15 at T1 and 5 at T2. Because we again have little participants completed the T2-questionnaire (N=5), we will leave this out of the statistical analyses for assessing the potential health effects. The Wilcoxon Signed-rank test showed that 4 health variables significantly increased at T1 compared to T0. First, the perceived health state on a Visual Analogue Scale (Z=-2.407, p=0.016). At T1, 9 participants had a higher perceived health state compared to at T0, 4 participants had a lower score at T1, and 2 participants had the same score. Second, the bodily function domain within positive health (Z=-2.456, p=0.014). At T1, 7 participants had a higher score than at T0, and 8 participants had the same score. Third, the quality of life domain within positive health (Z=2.310, p=0.021). At T1, 8 participants had a higher quality of life score, 1 participant had a lower score, and 6 participants had the same score compared to T0. Finally, the Wilcoxon Signed-rank test showed a significant increase in the variety domain within the Self-Management Ability Scale (-2.081, p=0.037). At T1, 8 participants had a higher score on this domain, 3 had a lower score, and 4 participants had the same score as at T0.

Table 21: Mean (SD) of health variables at T0, T1 and T2 in the first round and second round.

	First round (N T0=19, N T1=17, N T2=4)		
	M (SD) at T0	*M (SD) at T1*	*M (SD) at T2*
Perceived health state	0.79 (0.14)	0.81 (0.17)	0.88 (0.14)
Perceived health state on a VAS	84.3 (12.3)	84.4 (12.4)	87.5 (8.7)
Positive health domains			

Bodily functions	7.1 (1.7)	7.2 (1.8)	8.0 (0.8)
Mental health	8.2 (1.7)	7.5 (2.1)*	8.8 (0.5)
Meaning	8.0 (2.1)	7.8 (1.9)	9.3 (0.5)
Quality of life	8.1 (1.7)	7.8 (2.1)	9 (0)
Social participation	7.6 (2.6)	7.3 (2.2)	7.8 (1.0)
Daily routine	8.7 (1.3)	8.6 (1.3)	9.8 (0.5)
SMAS-s domains			
Taking initiatives	78.6 (19.6)	74.1 (18.5)	70.0 (17.6)
Investment behaviour	78.2 (18.4)	72.5 (16.6)	88.3 (3.3)
Variety	62.5 (27.0)	58.4 (18.3)	63.3 (16.8)
Multifunctionality	70.2 (17.8)	60.0 (18.4)	68.3 (10.0)
Self-efficacy	77.2 (18.6)	71.4 (17.3)	71.7 (19.9)
Positive frame of mind	66.3 (17.8)	60.8 (24.0)	81.8 (28.0)
SMAS-s total score	72.2 (19.9)	66.2 (15.4)*	73.9 (15.9)
	Second round (N T0=22, N T1=15, N T2=5)		
	M (SD) at T0	*M (SD) at T1*	*M (SD) at T2*
Perceived health state	0.72 (0.26)	0.75 (0.16)	0.61 (0.35)
Perceived health state on a VAS	69.8 (22.5)	76.7 (21.8)*	71.0 (35.1)
Positive health domains			
Bodily functions	6.3 (2.2)	7.2 (2.1)*	6.6 (2.6)
Mental health	7.1 (2.3)	7.7 (2.1)	8.2 (1.3)
Meaning	7.3 (2.1)	8.2 (1.5)	8.4 (1.5)
Quality of life	7.3 (2.1)	7.9 (1.8)*	7.8 (1.6)
Social participation	6.4 (2.7)	7.2 (2.3)	7.8 (1.6)
Daily routine	8.1 (1.9)	8.7 (1.2)	8.2 (2.5)
SMAS domains			
Taking initiatives	63.3 (20.2)	72.0 (19.2)	70.7 (27.7)
Investment behaviour	63.0 (16.7)	72.4 (20.1)	72.0 (23.3)
Variety	49.1 (14.5)	57.3 (14.2)*	64.0 (16.1)
Multifunctionality	56.4 (11.8)	55.1 (15.4)	53.3 (17.6)
Self-efficacy	58.8 (16.2)	64.0 (16.5)	68.0 (23.3)
Positive frame of mind	61.8 (21.7)	59.6 (23.3)	60.0 (22.6)
SMAS total score	58.7 (12.9)	63.4 (15.1)	64.7 (18.7)

* $p<0.05$ (Wilcoxon Signed-rank test)

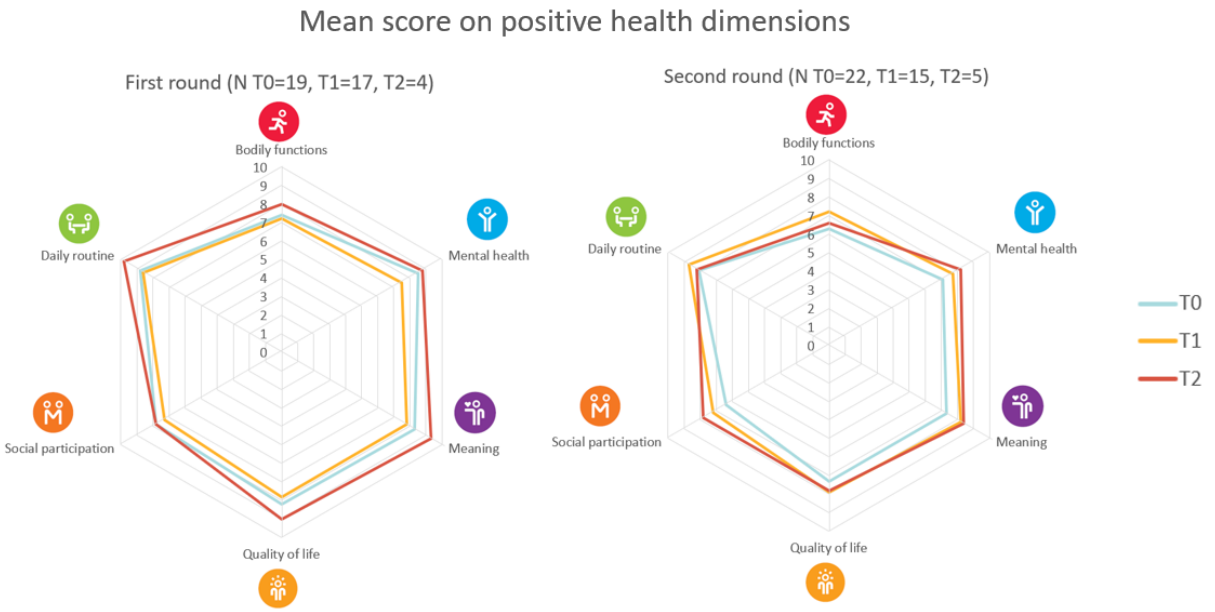

Figure 17: Spider plots mean score on positive health dimensions, divided into first and second round.

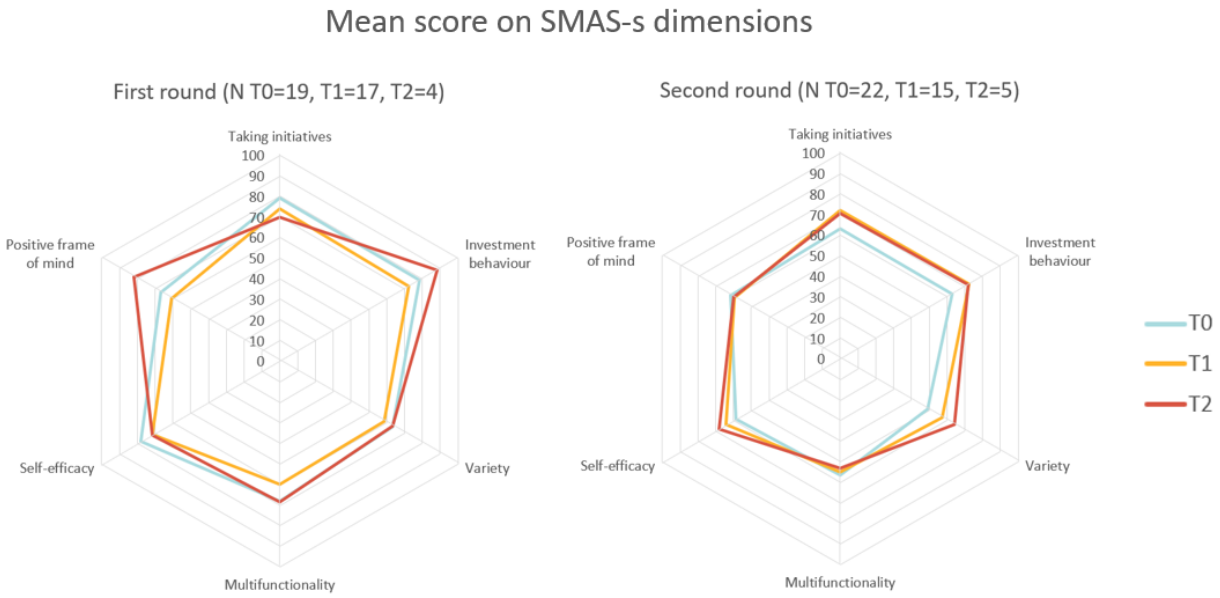

Figure 18: Spider plots mean score on Self-Management Ability dimensions, divided into first and second round.

5.5 Applicability of the virtual coaches

During the first round, Olivia and François both scored almost the same on their working alliance. Olivia scored slightly better, but both coaches were not rated very poor or good on their working alliance. During the second round, Olivia scored a bit better than François. During both rounds, both coaches scored highest among the domain Bond, with a mean of 2.7 (SD Olivia=1.0 and SD François=1.4) for both coaches in the first round, and 2.6 (SD1.5) for Olivia and 2.4 (SD=15) for François in the second round. Looking at the mean total score of their working alliance, Olivia scored slightly higher than François in both rounds. See Table 22 for the mean scores on each domain.

Table 22: Mean (SD) of domains of working alliance of Olivia and François in the first round and second round.

	First round				Second round			
	Olivia		François		Olivia		François	
	N	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*
Task	16	2.6 (1.1)	16	2.5 (1.2)	14	2.0 (1.1)	13	1.8 (0.8)
Bond	16	2.7 (1.0)	15	2.7 (1.4)	14	2.6 (1.5)	13	2.4 (1.5)
Goal	16	2.4 (1.0)	15	2.3 (1.2)	14	2.1 (1.3)	14	1.8 (1.0)
Total score	16	2.5 (0.9)	15	2.4 (1.2)	14	2.2 (1.2)	13	2.0 (0.9)

5.5.1 Satisfaction with the virtual coaches

At T0 of the first round, Emma scored highest on average, and Olivia scored lowest on average (see Table 23 and Figure 19). At T1, Katarzyna (only rated by one participant) and Olivia scored highest, and Rasmus sored lowest (only rated by one participant). Too little participants filled out the satisfaction scores for both T1 and T0, therefore, we did not test for significant differences between the mean scores of the coaches at T0 and T1.

In the second round, Emma scored highest on average at T0, and Carlos scored lowest on average (see Table 23 and Figure 20). At T1, François scored highest on average, and Katarzyna scored lowest. Results of the Related-Samples Wilcoxon Signed-Rank tests show that for all coaches the average satisfaction score dropped from T0 to T1.

Table 23: Satisfaction scores at T0 and T1 for every coach given by the Scottish participants.

	First round				Second round				
	N (T0)	*M(SD) at T0*	*N (T1)*	*M (SD) at T1*	*N (T0)*	*M(SD) at T0*	*N (T1)*	*M(SD) at T1*	*p*
Olivia	6	7.2 (2.6)	*15*	6.1 (2.9)	21	7.0 (2.0)	14	5.1 (2.8)	0.02
François	5	7.8 (2.3)	15	4.7 (2.9)	21	7.0 (1.7)	13	5.1 (2.2)	0.03
Emma	4	8.0 (0.8)	13	5.2 (2.7)	21	7.5 (1.6)	13	4.9 (2.4)	0.03
Helen	3	7.7 (0.6)	12	4.8 (2.4)	21	7.3 (1.6)	13	4.1 (2.5)	*0.03*
Carlos	5	8.0 (2.4)	14	4.9 (2.4)	21	6.9 (1.7)	14	4.9 (2.1)	*0.006*
Rasmus	1	8.0 (-)	1	3.00 (-)	21	7.3 (2.0)	6	4.0 (1.7)	0.03
Katarzyna	1	7.0 (-)	1	7.00 (-)	21	7.3 (1.8)	9	3.8 (1.6)	0.02

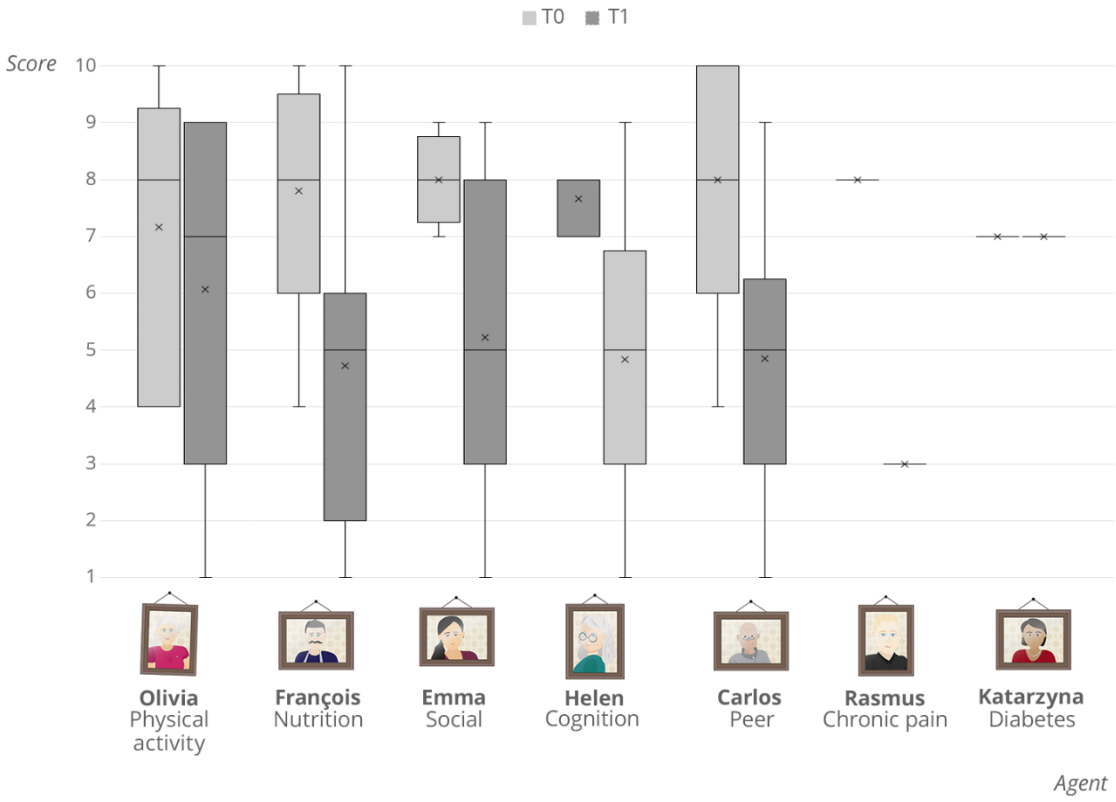

Figure 19: Boxplot showing the satisfaction scores of every coach at T0 and T1 given by the Scottish participants of the first round.

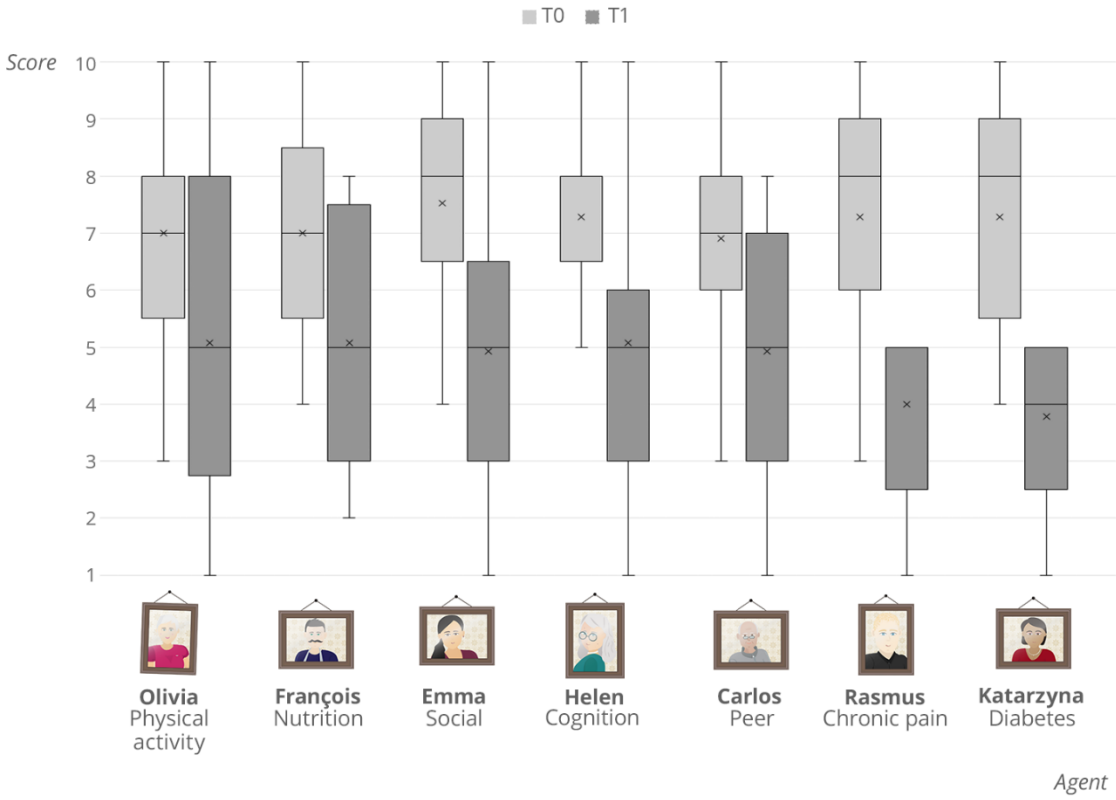

Figure 20: Boxplot showing the satisfaction scores of every coach at T0 and T1 given by the Scottish participants of the second round.

In the interviews of the first round, many participants (N=5) preferred Olivia, because of her content: good advice (N=1), the domain physical activity (N=1) and her personality: positive and friendly (N=2). François was preferred three times (N=3), because of its content: nice recipes (N=1), the domain nutrition (N=1) and his personality: a jovial character (N=1). Emma was preferred twice (N=2), both participants indicated they liked her content: providing valuable advice (N=2) and her personality: she was found to be friendly (N=2). Also, one of them preferred Emma, since she talked about the domain he or she needed most help with (N=1). Only one participant (N=1) did not indicate any preference, since he or she indicated not to have interacted with all the coaches.

Many participants preferred Carlos the least (N=5), mostly because of a lack of content (N=4), or a lack of personality (N=1). François was preferred the least once (N=1), by a participant that found his personality annoying: stereotyped and offensive. One participant (N=1) liked all the coaches. The rest of the participants did not prefer one coach the least, since they did not interact with all of them, or did not get to know all the coaches.

In the interviews of the second round, Olivia and François were preferred the most, both by 5 participants. Olivia was preferred because of her domain: interested in physical exercise (N=3), her content: most useful coach (N=1), and one because (s)he interacted most with Olivia. François was preferred because of his content: a lot of content (N=3), good recipes and tips (N=1), and useful coach (N=1). Two participants preferred Helen the most, one because of her content; interesting and taught new things, and one participant said (s)he wanted to keep speaking to her. Rasmus was preferred once because of his content: most useful. One participant had no preference for any coach because (s)he did not need much support.

The least preferred coach in the second round was Carlos (N=5), because of his content; all participants did not see the purpose of him being in the council, they did not need him. Katarzyna was preferred least by three participants, because of lack of content (N=2), and not feeling needing her (N=1). Two participants least preferred François one because of his content; not very useful, and one did not gravitate towards him. Emma and Helen were preferred least both by one participant, both because of their appearances. Emma looked too young, and Helen looked a little too intense like a psychiatrist. Two participants could not point out who they least preferred, because they did not interacted enough with the coaches.

A more detailed analysis of first round participants' preferences for particular coaches is described in:

ter Stal, S., Hurmuz, M., Jansen-Kosterink, S., Beinema, T., Bulthuis, R., op den Akker, H., Hermens, H., Tabak, M. *Preferences of Older Adults for Embodied Conversational Agents in a Multi-Agent eHealth Application.* (2020). ACM International Conference on Intelligent Virtual Agents (Submitted)

6 Results – MRT

In this section, we will present some preliminary results for the Micro-Randomized Trial. Further results will be published in the paper referenced in section 3.6.1. Since the facultative phase of the second round for the Scottish participants was still ongoing for some participants at the time of writing, for these participants data collected until August 10th (2020) was included. Furthermore, the demographic information for the participants will not be repeated in this section, since this information has been fully reported in sections 4.1 (for Dutch participants) and 5.1 (for Scottish participants).

6.1 Raw log data

In this subsection the collected raw log data for both evaluation rounds is described. This raw log data was the input for the pre-processing phase, which is described in the next subsection.

Most participants in the evaluation study created an account (the number of missing accounts matched the number of drop-outs) and these participants generated a large number of logged dialogues while interacting with the system. Table 24 provides an overview of the total number of accounts and logged dialogues for both evaluation rounds before pre-processing was performed.

Site	Round 1	Round 2*
Number of accounts		
The Netherlands	26	25
Scotland	19	22
Total	*44*	*47*
Number of logged dialogues		
Netherlands	4180	3406
Scotland	1897	2701
Total	*6077*	*6107*

Table 24: Number of accounts that had logged dialogues and the number of logged dialogues in total for both countries per round. *Note that the Scottish data for round 2 is not fully complete.

A boxplot showing the spread in the total number of completed dialogues per participant (split by country and round) before pre-processing can be found in Figure 21. As can be observed, in the first round the mean number of logged dialogues per participant was higher for Dutch participants (M = 160.77) than for Scottish participants (M = 99.84). In the second round this difference is smaller (M = 136.24 and M = 122.72, respectively). Please also note that not all Scottish participants have finished their facultative phase for round 2.

In terms of dialogue steps, for round 1, the Dutch participants completed a total of 892 dialogue steps on average and Scottish participants completed a total of 691 dialogue steps on average; the Scottish completing a slightly higher number of steps per dialogue (NL = 5.55, SC = 6.91). For round 2, the Dutch participants completed a total of 950 dialogue steps on average and the Scottish participants completed 830 (NL = 6.97, SC = 6.77). Please also note that not all Scottish participants have finished their facultative phase for round 2.

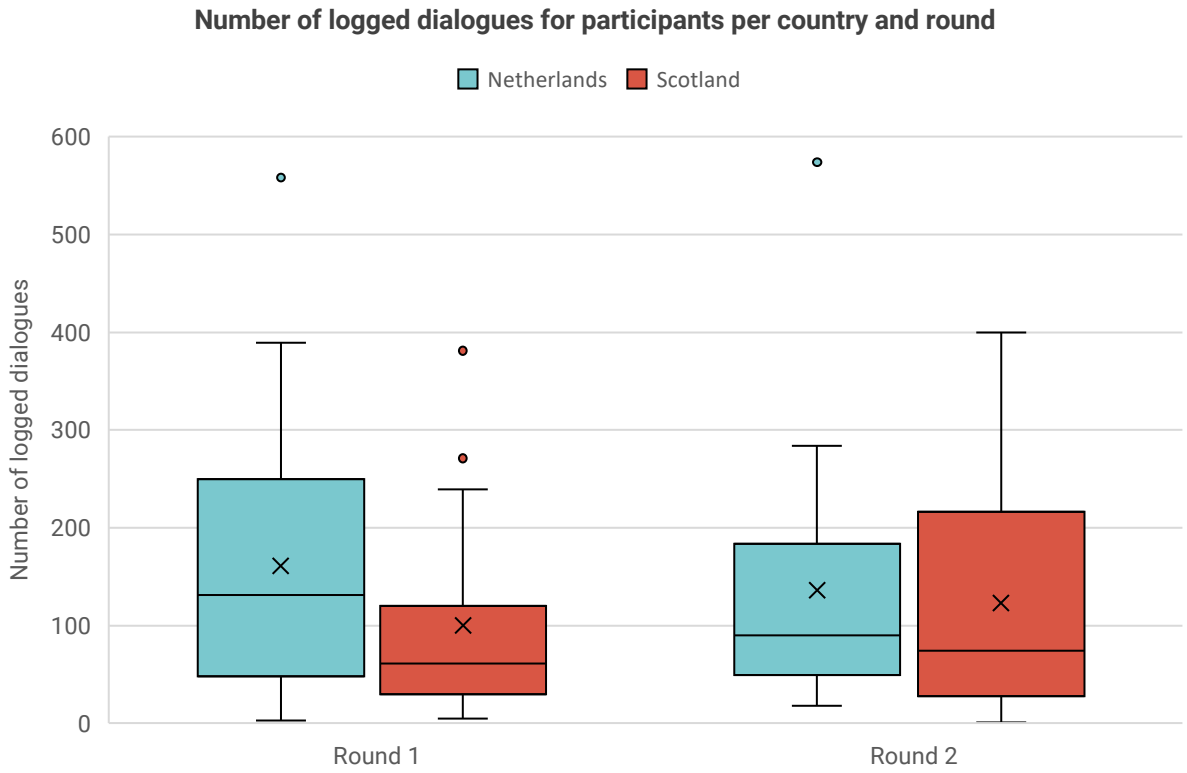

Figure 21: The spread for the number of logged dialogues per participant for both rounds and both countries. The crosses indicate the mean.

6.2 Raw log data and pre-processing

In this subsection we elaborate on the pre-processing that was applied to extract the interactions per experimental condition for the MRT from the raw log data.

As described in section 3.6.2, two pre-processing steps were performed to clean-up the data for analysis. In the first step, all dialogues were removed that were not part of an interaction with the physical activity coach. This was done because the physical activity coach was the coach for which the MRT was implemented; thus, logged dialogues for the other coaches were not relevant for the analysis of the MRT.

In the second step, all logged dialogues were removed that were caused by a by a double click error. Furthermore, an additional clean-up requirement was added to this step, namely removing all logs that were part of an interrupted interaction caused by a 'Fitbit connected delay'. This clean-up requirement was added since some participants experienced a delay when returning to the application after being referred to the Fitbit website to connect their Fitbit. Upon return, they should have immediately been shown a 'Fitbit connected' dialogue automatically, but in some cases, this dialogue started after 3-10 seconds when participants already had started another interaction with the physical activity coach. Since these new interactions were interrupted by the 'Fitbit connected' dialogues and could never reach their full potential length, and interrupting dialogues could also be part of the other experimental condition, we decided to remove both the interrupted and interrupting interactions from the data set.

The number of dialogues after each step and the number of removed dialogues in each step can be found in Figure 22.

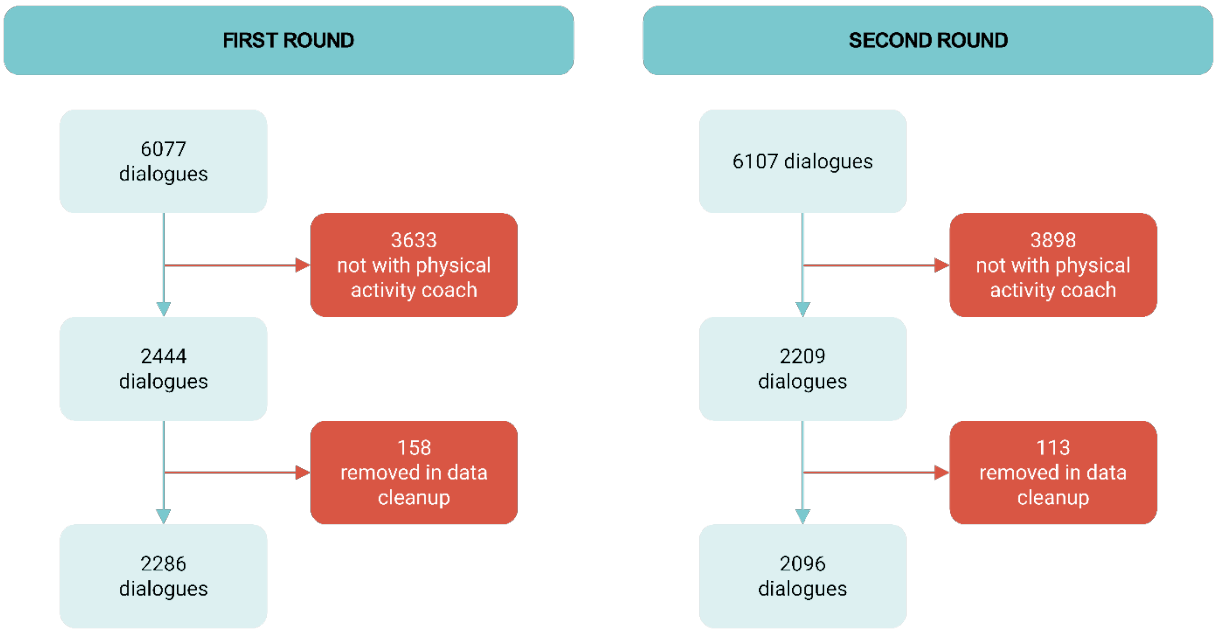

Figure 22: The evolution of the total number of dialogues during the pre-processing process. The sum of the dialogues for both countries was used in both rounds.

6.3 Interactions

Since the aim of the MRT was to compare the *coach initiative* and *user initiative* conditions, each logged dialogue belonged to one of these conditions. Using the definitions described in section 3.6.2 for interactions (groups of logged dialogues that were part of one 'conversation' with a coach) that corresponded to either of these conditions, we labelled the dialogues that were the start of these interactions. The total number of interactions for both conditions (split for the two rounds) can be found in Table 25 below.

Condition	Round 1	Round 2
Coach initiative	270	200
User initiative	289	228

Table 25: The number of interactions collected for both conditions of the MRT (split by evaluation round).

Figure 23 shows the spread for the number of interactions for both conditions that each participant had. As can be seen, the mean number of coach initiative interactions in both rounds is slightly lower than the mean number of user initiative interactions.

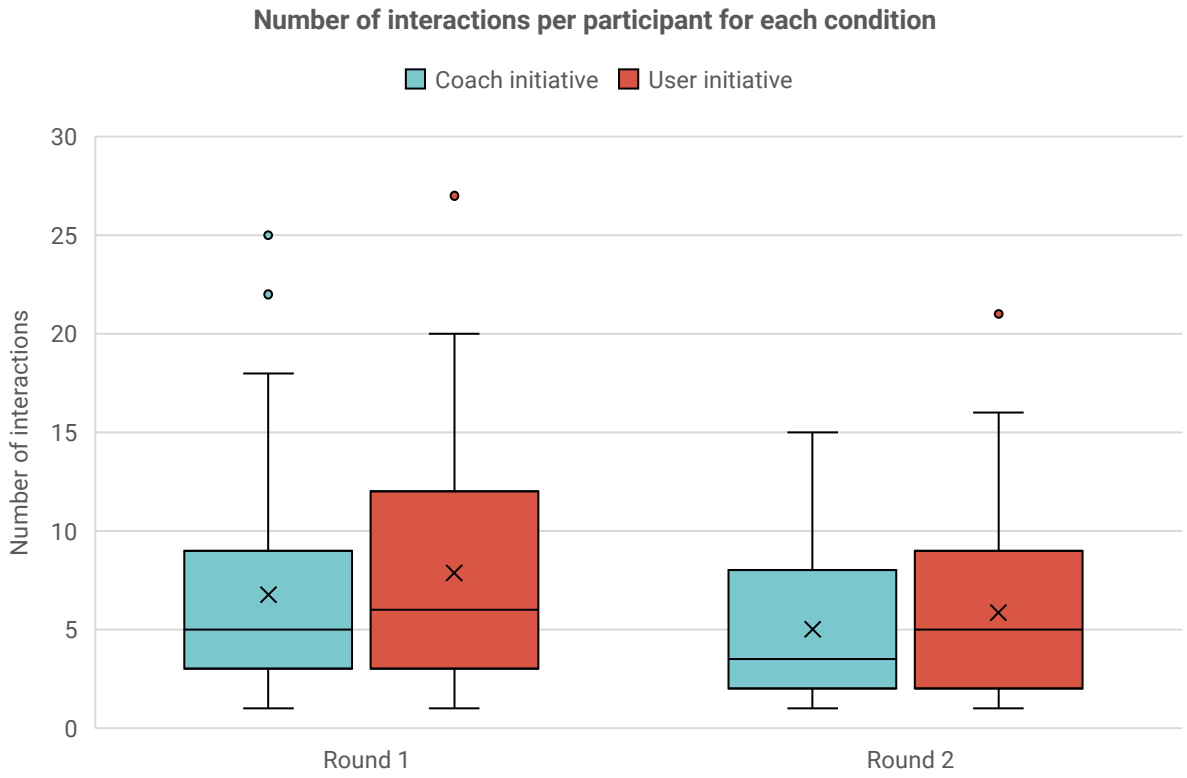

Figure 23: Boxplots showing the spread for the number of interactions per participant for both conditions (and split over the two evaluation rounds). The crosses indicate the mean.

6.4 Dialogue length

To compare the length of the interactions between the coach initiative condition and the user initiative condition we calculated the mean number of dialogue steps for both conditions for all users. Figure 24 shows this mean number of dialogue steps (coach statements and user replies) per interaction for the coach initiative and user initiative conditions for all the users in both evaluation rounds.

We performed a paired-samples t-test to compare the mean number of dialogue steps per interaction between both conditions for our participants. For the first round there were 36 participants that had interactions in both conditions. The mean number of dialogue steps for the coach initiative condition was 23.74 (*SD* = 11.82) and the mean number of dialogue steps for the user initiative condition was 24.01 (*SD* = 9.61). The test showed no significant difference in the mean number of dialogue steps between both conditions ($p > 0.05$).

For the second round there were 37 participants that had had interactions in both conditions. The mean number of dialogue steps for the coach initiative condition was 27.96 (*SD* = 16.17) and the mean number of dialogue steps for the user initiative condition was 26.50 (*SD* = 12.10). The test showed no significant difference in the mean number of dialogue steps between both conditions ($p > 0.05$).

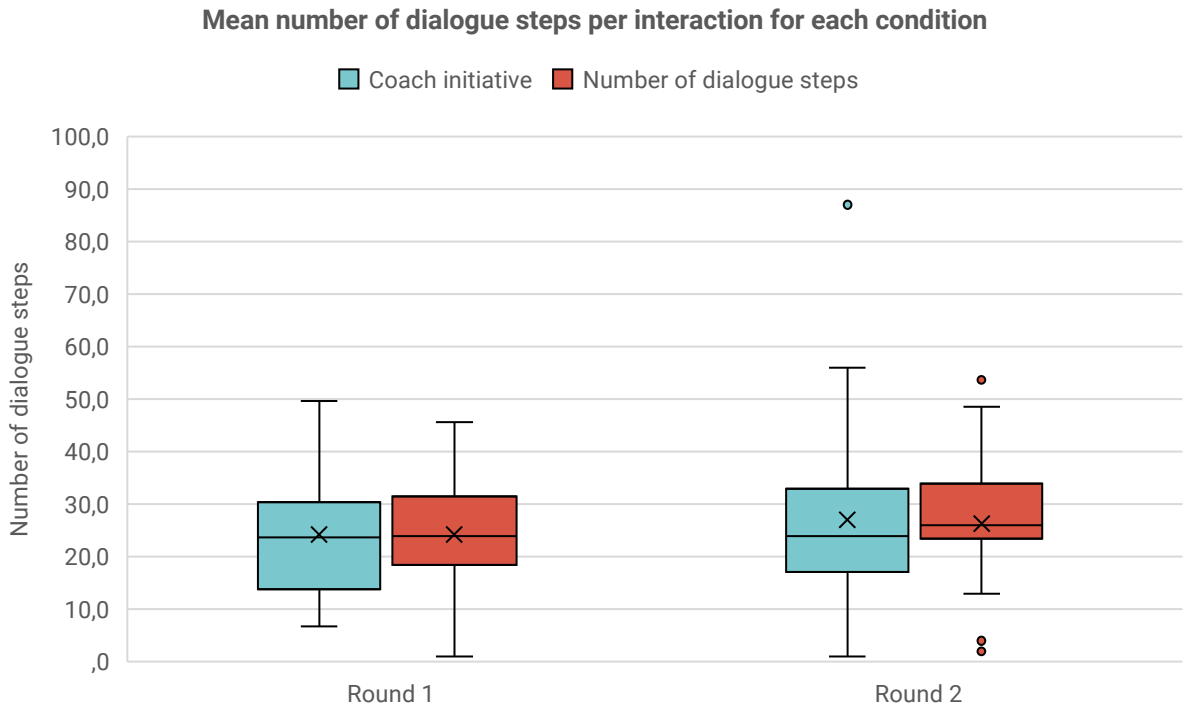

Figure 24: Boxplots showing the spread for the number of dialogue steps per interaction for both conditions (split by evaluation round). The crosses indicate the mean.

7 Discussion

The aim of this evaluation was to assess the use, user experience with, and the potential health effects of the Council of Coaches' functional demonstrator, which was implemented in a real-world setting among older adults aging 55 years or older. An observational cohort study was conducted in the Netherlands and Scotland, which consisted of 5-9 weeks. In both countries, we split the study population in two rounds, with the aim to include 25 participants in each round in each country. At the end of the study, we had a total of 92 participants. The study consisted of first a baseline week (participants only wore a Fitbit tracker), then four weeks of implementation phase (participants wore a Fitbit and used the functional demonstrator), then they could choose whether they wanted to be included in the follow-up phase, which consisted of four additional weeks (participants wore only a Fitbit tracker or a Fitbit and used the functional demonstrator).

During the set-up of this evaluation, we chose to also focus more on the virtual coaches and the effectivity of the coaching messages, next to the standard evaluation of the use, user experience and health effects. The findings of these multiple studies are divided. Study 1 is about the use, user experience and potential health effects, study 2 is about the virtual coaches, study 3 about the effectivity of coaching messages.

7.1 Study 1

7.1.1 Principal findings

First of all, regarding the outcome use we found that the mean number of use sessions of the functional demonstrator was higher in the implementation phase. During the first week of the implementation phase, most participants used the technology. In the Netherlands, the participants of the first round used it on average more times than in the second round, and in Scotland it was the other way around: 6.9 (SD=5.4) vs. 5.3 (SD=4.0) sessions during the implementation phase in the Netherlands and 3.5 (SD=2.6) vs. 5.7 (4.6) sessions in Scotland.

Next to this, according to the study population, the acceptability of the usability of the functional demonstrator was on average low. The average scores on the user experience domains were between 2.6 and 5.4. The domain trust in technology scored best, participants were on average not distrustful regarding the security of the functional demonstrator, and the use of their personal data. The domain intention to use scored the least in the Netherlands, they do not have the intention to use the system or recommend it to others, and enjoyment in Scotland, they do not think the system is entertaining or exciting to use. With interviews we tried to get more insights into these scores. We noticed that participants focused a lot on the content, and not on the system itself, especially regarding the usability. Not a lot of technical problems occurred during the implementation phase, but participants expected to get more in depth and personal advices. This gap between participants' expectations and the reality made participants probably less positive about the overall working of the system. Reasons given for not using the system was mostly related to content or difficulties with logging in (long email address/password), but not really regarding the working operation or usability of the system.

Looking at all the health variables we measured, we see that on average, participants score quite high. In the first round, none in group potential health effects were found in the Netherlands. In Scotland, there were health variables that decreased on average in the whole group: mental health and the total score of the Self-Management Ability Scale. We do not expect this is due to using the functional demonstrator. Instead, we think this can be a cause from the total lockdown Scotland experienced. In the second round, the participants of both countries increased significantly on different health variables as a group. In the Netherlands, the participants that used the functional demonstrator for at least four times during the implementation phase, had increases in two variables: positive frame of mind (a domain of the SMAS), and the total score of the Self-Management Ability Scale. Second-round participants of the Netherlands that used the functional demonstrator for less than four times, had a small decrease in one health variable: self-efficacy (a domain of the SMAS). In Scotland, participants health scores increased in four variables: perceived health state on a Visual Analogue Scale, the bodily function and quality of life domains of Positive Health, and the variety domain of the SMAS. This shows that using the functional

demonstrator had most influence on self-management. The SMAS-questionnaire measures self-management abilities in older adults: taking initiative, investment behaviour, variety, multifunctionality, self-efficacy and positive frame of mind (Schuurmans, et al., 2005). A higher score, shows a better self-management. Furthermore, using the functional demonstrator also has a potential effect on users' self-reported positive health and health state.

When we look at the individual health scores of the participants, we see that in the first round in the Netherlands there were 22 (out of 24) participants that had a clinically relevant increase in at least one of the health variables, and in Scotland 10 (out of 17) participants with such an increase. In the second round, 20 (out of 23) Dutch participants and 10 (out of 15) Scottish participants had a clinically relevant increase. Thus, this shows that when looking at individual level, 78,5% of the study population experienced a clinically relevant improvement in one of the health domains after using Council of Coaches' functional demonstrator.

7.1.2 Comparison with prior research

The law of attrition, as stated by Eysenbach (2005), describes the use data we found in our study. This law implies that in using eHealth, there are two attrition trends; one in which participants drop out completely (i.e. not using the application and not completing questionnaires), and one in which participants stop using the eHealth application, but are still completing questionnaires (Eysenbach, 2005). In our evaluation, we see this trend as well. We had participants that dropped out completely, and participants who did not use the functional demonstrator anymore, but were completing questionnaires. The decline in use data over time in this evaluation thus confirms the law of attrition in eHealth studies.

Comparing the usability score of the functional prototype used in this evaluation with the one evaluated in D2.6 (Van der Kamp, et al., 2019), the usability of the functional prototype scored lower during the current evaluation. In D2.6 the average usability score of both countries together was 73.9 (SD=17.9), and in this Deliverable the average usability scores were between 35.3 (SD=19.6) and 64.5 (SD=18.0). A possible explanation could be the setting in which the two studies have been conducted. In D2.6 the evaluation focused on usability in particular, and participants used the functional prototype for approximately 30 minutes in a lab setting. The current evaluation is the first evaluation in a real-world setting for at least 4 weeks, in which participants could choose themselves how to use it. Another possible explanation, is the gap between participants' expectations and the reality mentioned above.

Little research has been conducted towards the effect of virtual coaching systems on older adults' quality of life. Looking at the average health state of our study population, our participants perceive their quality of life quite high. Some participants indicated themselves during the interviews that they already are very active and living healthy, which could have influenced the small in group results we found. But looking at the individual scores, we see that even though the average health scores are high, the majority of the population experienced an improvement in at least one of the measured health variables.

7.2 Study 2

7.2.1 Principal findings

In this study, participants scored the virtual coaches at start of the study, and after the implementation phase. We found that after interacting with the virtual coaches for four weeks, participants rated the coaches lower than at first glance. A possible explanation for this can be that participants had high expectations about being coached toward a specific personal aim. As described above in the principal findings of the first study, during the interviews, participants were talking a lot about content. At first glance, participants scored them based on a picture with a name and coaching domain. But after the implementation phase, they scored them more based on the content they had. When asking them which coaches they liked the most and the least, most participants based this on the content they had or their coaching domain. So, if a coach had little content, most times (s)he was mentioned as the least preferred one.

For measuring the applicability of the virtual coaches, the Working Alliance Inventory questionnaire was used. This questionnaire measures the therapeutic alliance; the higher the score, the more likely someone will follow the advices given (Paap, Schrier, & Dijkstra, 2018). In the Netherlands, the primary

coaches, Olivia and François, were rated poorly. In Scotland, both coaches scored better, but still not very high. This questionnaire is based on a human health professional, we now used it for virtual coaches. The participants did not find these questions really related to them. During the interviews, some participants said it was hard to answer these questions, because it is a computer with who they are talking, so they did not really feel a connection. For some this was a reason to score both coaches on all questions the lowest score.

7.2.2 Comparison with prior research

Little research is done regarding difference in rating virtual coaches before interacting with them, and after interacting for some amount of time. Komatsu et al. (2012) looked at expectations of people towards the functioning of a robotic agent and how people actual perceived the functioning of the robotic agent. They saw that when people had low expectations about the functioning, and were positive about the actual functioning, they had a higher acceptance rate. In our study, we saw that people had high expectations, but were not positive about the actual functioning of the virtual coaches. This can be the reason for the lower satisfaction scores after interacting with them.

7.3 Study 3

7.3.1 Principal findings

In this study, we micro-randomized the initiative that was taken at the start of the interaction with the physical activity coach. Every time a participant would start an interaction by clicking on her, they had a 50% chance that she would suggest a topic for them to discuss (coach initiative) and a 50% chance that they were asked what they wanted to discuss (user initiative). Our hypothesis was that participants might be more engaged in the conversation if the coach took initiative, thus leading to longer conversations.

However, when comparing the mean length of dialogues that participants had between these conditions, no significant results were found for both rounds. For the first round, the mean length of dialogues seemed slightly higher in the user initiative condition, while in the second round this mean length was slightly higher for the coach initiative condition. The lack of difference suggests that participants might not have noticed a difference in both conditions and thus the coach suggesting a topic was not experienced negatively, but apparently blended well with the user initiative interactions, which can be seen as the baseline.

Furthermore, we also suspect that there might be two effects that had an influence on the number of dialogue steps completed for coach initiative and user initiative dialogues in general. First, in the interviews some participants indicated that they had a clear goal for what they wanted to discuss and the topic selection for the physical activity coach did (and could) not take this into account. In the user initiative condition participants would select the dialogues themselves, but in the coach initiative condition they would cancel the dialogue and try again.

Second, a difference in dialogue steps might have been created by the pre-dialogues that we constructed for both conditions. When a coach initiative dialogue was started it would take users about 2 dialogue steps to reach the dialogue with the content for that topic (‘Shall we discuss your activity data?’ and ‘Yes’). However, when a user initiative dialogue was started it would take users at least 4 dialogue steps to reach the content for that topic (E.g. ‘What do you want to discuss’, ‘Let’s talk about coaching’, ‘How can I help you with coaching’, ‘Can you show me my activity data?’).

Future work should investigate these effects and take them into account. Extension of the physical activity coach’s content and topic selection algorithm with dialogues that more extensively discuss the participant’s personal aims seems a promising direction to improve the relevance of suggested topics.

7.3.2 Comparison with prior research

In our study, we tailored conversational topics to participants in the coach initiative condition, while the user initiative condition implemented one of two often used traditional approaches (the other being to provide all participants with the same program and order of dialogues). Some examples can be found in literature in which the topics of conversation are tailored by letting a human (e.g. a health care

professional) define these in the system's backend (e.g. (Abdullah, Gaehde, & Bickmore, 2018) and (Fadhil, Wang, & Reiterer, 2019)) and these approaches are well received by participants, but this is a labour intensive process. This does however suggest that the principle of tailoring dialogues holds and that our results are mainly influenced by participant attention and extensiveness of tailoring.

From the preliminary results of the study so far, we conclude that some prerequisites for human-human conversation and tailoring also remain essential for the tailoring of topics that we implemented. In her definition of a topic for human-human conversation, Riou (2015) stated that a conversational topic is what a portion of the interaction is about, but also that it *has to be the centre of shared attention* by participants in the conversation. Since our participants were in some cases focussed on finding the specific topic, they wanted to discuss instead of being open to discuss anything else the coach might suggest, this shared attention was never established.

Furthermore, tailoring involves the adjustment of timing, intention, content, and representation (op den Akker, Jones, & Hermens, 2014). While timing in our coach initiative condition was supported by the participant clicking on the coach and representation was largely covered by the application's look and feel, tailoring the intention and content to participants requires 'a database of information about the user and prior interactions with them' (Bickmore, Gruber, & Picard, 2005). We assume that an improved topic suggestion algorithm that could take into account more information about participants and thus would suggest more relevant and tailored topics to the participants would also improve the participants' 'openness' to topics that the coach suggests.

7.4 Strengths and limitations

A very important strength of this study and other evaluation studies within the Council of Coaches project, is the iterative evaluations. During each evaluation we included the target group, to give us advices, input, recommendations to improve the system. This end user involvement gave us the opportunity to really improve the functional demonstrator according to their needs. In the Netherlands, the participants of the first three studies, were recruited from a research panel. They participate also in other studies at RRD and are used to this kind of studies. For this last study, we chose to recruit participants in other ways. We involved a whole new population in this study, which gave us new insights, as an enrichment to our previous evaluations. Second, during this evaluation we did not only focus on effectivity, as in a lot of other evaluation studies. We chose to use a pragmatic form of evaluating, which gave us the space to not only focus on health effects, use and user experience, but also perform other studies within one evaluation. Finally, in this study, we involved not only healthy older adults, but also adults with health conditions: 17 participants diagnosed with Diabetes Mellitus Type 2, 6 diagnosed with chronic pain, and 3 diagnosed with both Diabetes Type 2 and chronic pain.

This study had some limitations. First, a lot of participants had other expectations about the functional demonstrator, and about virtual coaches. They thought that there was a real person behind the computer, and some thought it would really help them with losing weight. Furthermore, during the interviews, we noticed that participants find it hard to see such a technology to use it for themselves. They quickly think about other people who would benefit from it, but not about themselves. People mostly tend to think in solving problems, instead of preventing problems. We are not always used to think in the preventive side of health care. Older adults tend to live from day to day, they do not think a lot about problems that can occur in the near future. If someone thinks (s)he is healthy, or does not have any problems, it is harder to see such a technology as something that would work for him/her.

For future research with a virtual coaching system, it is important to give a lot of attention to expectations management; to give clear information about what (not) to expect and the importance of such a system, to avoid these problems.

8 Overall conclusion

This deliverable describes the results of the final evaluation with the fourth Council of Coaches' functional prototype. This evaluation shows that not a lot of problems occurred during the use of the functional demonstrator in participants' home setting. Almost half of the participants used the functional demonstrator also at least once in the facultative phase. A lot of participants saw potential in implementing this demonstrator in real life, maybe not for themselves but for people they know who can benefit from using this. This evaluation also shows that to get such a virtual coaching system on the market, it is important for the coaches to have more personalized content, i.e. more tips and advices related to the user's situation, and for the users to have an opportunity to ask questions to the coaches. This is an important next step for further use of the Council of Coaches' functional demonstrator and for broadening the older adults target group that can benefit from this.

The results of this evaluation will be published in scientific journals/at conferences:

- Hurmuz, M.Z.M., Jansen-Kosterink, S.M., De Franco, D., op den Akker, H., Hermens, H.J. *User Experience, Use, and Potential Health Effect of a Conversational Agent-Based Electronic Health Intervention: An Observational Cohort Study*. International Journal of Medical Informatics (probably).
- ter Stal, S., Hurmuz, M., Jansen-Kosterink, S., Beinema, T., Bulthuis, R., op den Akker, H., Hermens, H., Tabak, M. *Preferences of Older Adults for Embodied Conversational Agents in a Multi-Agent eHealth Application.* (2020). ACM International Conference on Intelligent Virtual Agents (Submitted).
- Beinema, T., op den Akker, H, Hurmuz, M., Jansen-Kosterink, S., Hermens, H. (2020). *Automatic topic selection for embodied conversational agents in health coaching: a micro-randomized trial.* Journal of *XXX* (to be decided).

9 Bibliography

Beinema, T., Op den Akker, H., Broekhuis, M., Huizing, G., van Velsen, L., ter Stal, S., & van Loon, J. (2019). *D2.4: Evaluation results of first functional prototype and updated requirements.* The Council of Coaches Consortium.

Beinema, T., Broekhuis, M., Op den Akker, H., Van Velsen, L., Ter Stal, S., Pease, A., . . . Van Loon, J. (2019). *D2.5: Evaluation of the second functional prototype and updated requirements.* The Council of Coaches Consortium.

Van der Kamp, M., Beinema, T., Broekhuis, M., Op den Akker, H., De Franco, D., Pease, A., . . . Welan, E. (2019). *D2.6: Evaluation of third functional prototype and updated requirements.* The Council of Coaches Consortium.

Brooke, J. (1996). SUS - A quick and dirty usability scale. In J. Brooke, P. Jordan, B. Thomas, I. McClelland, & B. (. Weerdmeester, *Usability evaluation in industry* (pp. 189-194).

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science, 35*(8), 982-1003. doi:10.1287/mnsc.35.8.982

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319. doi:10.2307/249008

Van Reenen, M., & Janssen, B. (2015). *EQ-5D-5L User Guide: Basic information on how to use the EQ-5D-5L instrument.* Retrieved from https://euroqol.org/publications/user-guides/

Schuurmans, H., Steverink, N., Frieswijk, N., Buunk, B., Slaets, J., & Lindenberg, S. (2005). How to Measure Self-management Abilities in Older People by Self-report. The Development of the SMAS-30. *Quality of Life Research, 14*(10), 2215-2228.

Huber, M., van Vliet, M., Giezenberg, M., Winkens, B., Heerkens, Y., Dagnelie, P., & Knottnerus, J. (2016). Towards a 'patient-centred' operationalisation of the new dynamic concept of health: a mixed methods study. *BMJ open, 6*(1), e010091.

Paap, D., Schrier, E., & Dijkstra, P. (2018). Development and validation of the Working Alliance Inventory Dutch version for use in rehabilitation setting. *Physiotherapy Theory and Practice, 35*(12), 1292-1303.

Komatsu, T., Kurosawa, R., & Yamada, S. (2012). How Does the Difference Between Users' Expectationsand Perceptions About a Robotic Agent Affect Their Behavior?: An Adaptation Gap Concept for Determining Whether Interactions Between Usersand Agents Are Going Well or Not. *International Journal of Social Robotics, 4*, 109-116.

Beinema, T., Op den Akker, H., Kosterink, S., ter Stal, S., & van den Boer, J. (2019). *D3.4: Final coaching actions and content.* The Council of Coaches Consortium.

Riou, M. (2015). A Methodology for the Identification of Topic Transitions in Interaction. *Discours*(16).

Abdullah, A. S., Gaehde, S., & Bickmore, T. W. (2018). A tablet based embodied conversational agent to promote smoking cessation among veterans: A feasibility study. *Journal of Epidemiology and Global Health*, 225-230.

Fadhil, A., Wang, Y., & Reiterer, H. (2019). Assistive Conversational Agent for Health Coaching: A Validation Study. *Methods of Information in Medicine*, 9-23.

op den Akker, H., Jones, V. M., & Hermens, H. J. (2014). Tailoring real-time physical activity coaching systems: a literature survey and model. *User Modeling and User-Adapted Interaction, 24*(5), 351-392.

Bickmore, T. W., Gruber, A., & Picard, R. (2005). Establishing the computer - patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, 21-30.

Hurmuz, M., Jansen-Kosterink, S., Op den Akker, H., & De Franco, D. (2020). *D7.6: Demonstration protocol, ethical approval.* The Council of Coaches consortium.

Hurmuz, M. Z., Jansen-Kosterink, S. M., Op den Akker, H., & Hermens, H. J. (2020). User Experience and Potential Health Effects of a Conversational Agent-Based Electronic Health Intervention: Protocol for an Observational Cohort Study. *JMIR Research Protocols, 9*(4), e16641. doi:10.2196/16641

Eysenbach, G. (2005). The Law of Attrition. *Journal of Medical Internet Research, 7*(1), e11.

10 Appendix A

Throughout the final evaluation of the functional demonstrator, changes are made to improve the demonstrator (based on recommendations of the participants), dialogues have been updated and more dialogues have been added. First, we show a table with system updates done after 30th of January (see Table 26), and then a table with all dialogues that have been added/updated also after 30th of January 2020 (see Table 27).

Table 26: Showing all system updates conducted after 30th of January

System improvements	When?
Fixed a bug in intake for evaluation study in the Netherlands	February
Olivia's topic selection mechanism now chooses between social talk, goal setting, discussing sensors, providing feedback, gathering information, informing how and informing why.	February
Fixed an issue with the language fallback mechanism. When a dialogue was not available in a non-default language, the system would fall back to an English dialogue (intended) but would subsequently not change back to the preferred language (fixed).	February
Fixed an issue where Emma, Helen and Carlos sporadically initiated an incomplete dialogue.	February
When Olivia takes the initiative in the conversation, she will now use the correct language.	February
A new mechanism for handling dialogue translations (from English to Dutch) has been implemented in the background. This should make it easier to make changes to existing dialogues and to add additional content, hopefully speeding up the rate at which we can add more dialogue content to the Council of Coaches B.E.T.A.	March
The radio will now continue playing after the first song has finished.	March
All the radio channels have been upgraded with brand new music. Each channel now randomly plays between a collection of 20 different tracks.	March
Improved the error messaging after unsuccessful login attempts (in case of empty email address or empty password fields).	March
Removed the amount of client-side logging that is shown during normal operation	April
Added an icon to the login password field that allows you to see the password you are typing.	May
Updated the Credits section to include the Dutch Voedingscentrum (recipes) and the Internet Archive (music).	May
You can now go back to the main menu during the initial account creation and intake procedure with the back button in the bottom left corner.	May
Updated the intake procedure to support the second Council of Coaches evaluation round. Research participants can enter a 4-letter code to be automatically enrolled in the correct part of the evaluation, while other users can skip this.	May
Fixed an issue for evaluation participants trying to login to the Council of Coaches functional demonstrator during the baseline week. They are now correctly given the message indicating that they have to wait for their baseline week to be finalized.	May
Added support for the "Monitoring Human Behaviour During the COVID-19 Lockdown" study.	May

Fixed some textual issues in coach statements, where statements constructed from multiple parts were concatenated without a separating space.	May
When entering the research code for the evaluation participants, the user can only enter capital letters, and only exactly 4 characters.	May
Login email address is no longer case sensitive when creating an account or logging in.	June
The timeslot in which Carlos, Helen and Emma can ask their questions has been extended from 18h-24h to 18h-06h.	June
Olivia is now slightly smarter when suggesting a topic to discuss herself.	June
When Olivia suggests something to talk about, in all cases there is now a reply option to say to talk about something else.	June
Fixed some cases where Helen would use an English sentence in a Dutch dialogue.	June

Table 27: Which dialogues are added/updated, and what was the addition/update

Coach	Dialogue name	Month added	Month updated	Dialogue changes
Olivia	olivia-coaching-feedback-ai		February March	Olivia offers the user the opportunity to talk about his/her physical activity progress.
	olivia-coaching-feedback-data		February April	We lowered the number of steps that François has taken in his last week, used to demonstrate how Olivia's activity book works (he was setting the bar a little too high). Olivia can properly explain how the activity book works, even if a user did not select François as a coach.
	olivia-coaching-feedback-menu		April	When a user has not connected his/her activity tracker, Olivia gives information about an activity tracker and explains how she can help the user with his/her physical activity when (s)he connects his/her activity tracker.
	olivia-coaching-gatherinformation-ai		February	Olivia can start a dialogue to gather information about the user.
	olivia-coaching-gatherinformation-dog	February		A new dialogue in which Olivia will ask the user about dogs.
	olivia-coaching-gather-information-environment	February		A new dialogue in which Olivia will ask the user about his/her living environment.
	olivia-coaching-goals-ai		February March	Olivia offers various dialogues regarding goal setting.
	olivia-coaching-goals-long-term-change	February		Olivia can help the user in changing the long-term physical activity goal that has been set earlier.

	olivia-coaching-goals-long-term-set		February March	Olivia can help the user in setting a long-term physical activity goal in number of steps or number of active minutes. Improved Olivia's dialogues that allow the user to set a long-term physical activity goal.
	olivia-coaching-goals-menu		February March	Olivia's menu of talking about physical activity goals has been updated.
	olivia-coaching-goals-notracker		February	When trying to set a goal without having an activity tracker connected, Olivia will help the user connect a tracker
	olivia-coaching-inform-how-friends-and-physical-activity		February	Olivia has a new advice on how to improve your physical activity together with your friends.
	olivia-coaching-inform-how-lunch-walk		February	Olivia has a new advice on how to improve your physical activity during your lunch walks.
	olivia-coaching-inform-how-reduce-sedentary-behaviour-evening		February	Olivia has a new advice on how to improve your physical activity in the evenings.
	olivia-coaching-inform-how-selector		February	Olivia offers advices on how to improve your physical activity. Olivia can repeat previous given tips on how to become more physically active.
	olivia-coaching-inform-how-stairs		February	Olivia has a new advice on how to improve your physical activity by using stairs.
	olivia-coaching-inform-menu		February	The user can ask Olivia now about the why physical activity is good and how to become more active.
	olivia-coaching-inform-why-ai		February	Olivia offers advice on why physical activity is good for you.
	olivia-coaching-inform-why-archieve		February	When having received an advice on why physical activity is good for you, you can review previous advice given.
	olivia-coaching-inform-why-biological-age		February	Olivia has a new advice on why physical activity is good for you, related to your biological age
	olivia-coaching-inform-why-brain		February	Olivia has a new advice on why physical activity is good for you, related to your brain activity
	olivia-coaching-inform-why-selector		February	Olivia's dialogue about selecting a tip has been updated.
	olivia-coaching-menu		February April	You can ask Olivia about 'Tips and info'.
	olivia-coaching-sensors-fitbit-completed		April	When returning from the Fitbit website (after connecting your Fitbit), Olivia will give more detailed information on the

				success or potential failure of the procedure.
	olivia-coaching-sensors-fitbit-faq	April		Olivia has a new dialogue about frequently asked questions about which Fitbit the user uses.
	olivia-coaching-sensors-fitbit-faq-inspire	April		Olivia has a new dialogue about frequently asked questions about the Fitbit Inspire the participants of the evaluation are using.
	olivia-coaching-sensors-fitbit-setup		April	Minor improvements to the Fitbit setup procedure from Olivia. Minor textual fix to the English setup dialogue by Olivia.
	olivia-coaching-sensors-menu		April	Additional questions about activity tracker brand.
	olivia-menu		April	Improved the flow of conversation with Olivia when the user wants to 'discuss something else'.
	olivia-social-story-1	June		Olivia has a new background story available about her dog.
	olivia-social-story-2	June		Olivia has a new background story available about her living situation.
	olivia-social-story-3	June		Olivia has a new background story available about her ex-husband.
François	francois-coaching-inform-menu		March	Added option to this menu to talk about importance healthy diet.
	francois-coaching-inform-why-archieve	March		François can repeat a previous given tip if requested.
	francois-coaching-inform-why-fruit	March		François can inform the user about fruit.
	francois-coaching-inform-why-salt		March	Fixed a bug in François' tip on eating less salt.
	francois-coaching-inform-why-selector	March		Added dialogue about the tips François has about a healthy diet.
	francois-coaching-inform-why-water	March		François can inform the user about water.
	francois-coaching-menu		March	François can discuss 'Tips and info'.
	francois-coaching-recipes-menu		February	The user can do the food preferences intake with François, the user can browse through the recipes, or ask François to help him/her choose a recipe.
	francois-coaching-recipes-selectrecipe		February	François' recipe selection procedure was fixed in English and translated to Dutch. François can guide the user in English and Dutch through selecting a

				recipe out of a database of 280 delicious and healthy recipes.
	francois-menu		March	François talks about the radio.
	francois-social-story-2	June		François has a new background story available about his favourite cheese.
	francois-social-story-3	June		François has a new background story available about his background as a chef.
	francois-social-story-4	June		François has a new background story available about a fight with his dad when he was a teenager.
Emma	emma-coaching-gather-information-ai		February March	Emma asks whether the user has time for some questions.
	emma-coaching-gather-information-esm	March	March	Emma asks the user questions to gather information about his/her social activities.
	Emma-coaching-inform-tips	June		Emma has a tips dialogue with then tips for being social.
	emma-coaching-menu	February	March April	Added a menu with things Emma can discuss: tips, coaching sessions, something else.
	emma-coaching-weekly-1	February	April	Emma's first coaching session is online, which focuses on your social network. Fixed a Dutch translation issue in Emma's first weekly coaching session.
	emma-coaching-weekly-2	March	March April	Added Emma's second coaching session about social activities. Emma's second coaching session is more intelligent now. Updated Emma's second weekly coaching session to include content that is better tailored to the British and Dutch evaluation participants. Fixed an issue with Emma's second weekly coaching session when revisiting the session through the coaching session archive.
	emma-coaching-weekly-3	February	May	Emma has her third weekly social coaching session enabled about making a plan to broaden the user's social network.
	Emma-coaching-weekly-4	June		Emma has her fourth weekly social coaching session enabled about reviewing the previous sessions.
	emma-menu		February March	Emma's menu is updated to talk about coaching and do a coaching session

	Emma-social-story-1	June		Emma has a new background story available about her love for music instruments.
	Emma-social-story-2	June		Emma has a new background story available about her shoe collection.
	Emma-social-story-3	June		Emma has a new background story available about her parents.
Helen	helen-coaching-gather-information-ai		February March	Helen asks whether the user has time for some questions.
	helen-coaching-gather-information-esm	March	March	Helen asks the user questions to gather information about his/her cognitive activities.
	helen-coaching-menu	June		Updated Helen's 'coaching session available' dialogue a little to add some Dutch translations.
	helen-coaching-weekly-1	May		Helen has her first weekly cognitive coaching session enabled about brains and memory.
	helen-coaching-weekly-2	June		Helen has her second weekly cognitive coaching session enabled about how our memory works, and the difference between forgetfulness because of ageing and because of dementia.
	helen-coaching-weekly-3	June		Helen has her third weekly cognitive coaching session enabled about external strategies to improve our memory.
	helen-coaching-weekly-4	June		Helen has her fourth weekly cognitive coaching session enabled about internal memory strategies to improve our cognitive health.
	helen-menu		February March May	Helen's menu is updated to ask the user how to help her/him. The user can choose to do a coaching session with Helen.
	Helen-social-story-2	June		Helen has a new background story available about how she grew up.
	Helen-social-story-3	June		Emma has a new background story available about her travel experience to China.
Carlos	carlos-coaching-gather-information-ai		March	Carlos asks whether the user has time for some questions.
	carlos-coaching-gather-information-esm	March		Carlos asks the user questions to gather information about his/her overall mood.
	carlos-social-story-1	June		Carlos has a new background story available about football.

	carlos-social-story-2	June		Carlos has a new background story available about his grandchildren.
	carlos-social-story-3	June		Carlos has a new background story available about his unhealthy living.
Rasmus	Rasmus-coaching-exercise-introduction	June		Rasmus has new exercises to improve sleep. This introduction shows all exercises the user can choose between.
	Rasmus-coaching-exercise-mindful	June		Rasmus gives information about mindfulness exercises to sleep better.
	Rasmus-coaching-exercise-relax	June		Rasmus gives information about tense and relax exercise to sleep better.
	Rasmus-coaching-exercise-stomach	June		Rasmus gives information about stomach breathing exercise to sleep better.
	Rasmus-coaching-introduction	June		The user can choose what (s)he wants to discuss with Rasmus: learning about sleep or helping with a relaxation exercise.
	Rasmus-coaching-weekly-1-bioclock	June		This is part of Rasmus' first weekly coaching session and is about the biological clock.
	Rasmus-coaching-weekly-1-importance	June		This is part of Rasmus' first weekly coaching session and is about the importance of sleep.
	Rasmus-coaching-weekly-1-insomnia	June		This is part of Rasmus' first weekly coaching session and is about insomnia.
	Rasmus-coaching-weekly-1-introduction	June		Rasmus has four weekly sessions in which he coaches on sleep. Each of Rasmus' sessions has a set of informative subtopics that he will discuss. The subtopics of the first weekly session are told in this dialogue.
	Rasmus-coaching-weekly-1-phases	June		This is part of Rasmus' first weekly coaching session and is about sleep phases.
	Rasmus-coaching-weekly-1-quiz	June		Each of Rasmus' sessions also has a quiz at the end about the information he provided. This is the first quiz.
	Rasmus-coaching-weekly-1-sleeppressure	June		This is part of Rasmus' first weekly coaching session and is about sleep pressure.
	Rasmus-coaching-weekly-2-bedtime	June		This is part of Rasmus' second weekly coaching session and is about bedtime.
	Rasmus-coaching-weekly-2-introduction	June		Rasmus has four weekly sessions in which he coaches on sleep. Each of Rasmus' sessions has a set of informative subtopics that he will discuss. The subtopics of the second weekly session are told in this dialogue.
	Rasmus-coaching-weekly-2-naturallight	June		This is part of Rasmus' second weekly coaching session and is about natural light and melatonin.
	Rasmus-coahing-weekly-2-quiz	June		Each of Rasmus' sessions also has a quiz at the end about the information he provided. This is the second quiz.

	Rasmus-coaching-weekly-2-restriction	June		This is part of Rasmus' second weekly coaching session and is about sleep restriction.
	Rasmus-coaching-weekly-2-rituals	June		This is part of Rasmus' second weekly coaching session and is about sleep rituals.
	Rasmus-coaching-weekly-2-surroundings	June		This is part of Rasmus' second weekly coaching session and is about elements in the surrounding to create a good place to sleep.
	Rasmus-coaching-weekly-3-activity	June		This is part of Rasmus' third weekly coaching session and is about physical activity and its positive effects on body.
	Rasmus-coaching-weekly-3-introduction	June		Rasmus has four weekly sessions in which he coaches on sleep. Each of Rasmus' sessions has a set of informative subtopics that he will discuss. The subtopics of the third weekly session are told in this dialogue.
	Rasmus-coaching-weekly-3-quiz	June		Each of Rasmus' sessions also has a quiz at the end about the information he provided. This is the third quiz.
	Rasmus-coaching-weekly-3-relaxation	June		This is part of Rasmus' third weekly coaching session and is about physical activity and relaxation.
	Rasmus-coaching-weekly-3-stress	June		This is part of Rasmus' third weekly coaching session and is about the influence of stress on insomnia.
	Rasmus-coaching-weekly-4-alcohol	June		This is part of Rasmus' fourth weekly coaching session and is about the influence of alcohol on sleep.
	Rasmus-coaching-weekly-4-caffeine	June		This is part of Rasmus' fourth weekly coaching session and is about the influence of caffeine on your body.
	Rasmus-coaching-weekly-4-introduction	June		Rasmus has four weekly sessions in which he coaches on sleep. Each of Rasmus' sessions has a set of informative subtopics that he will discuss. The subtopics of the fourth weekly session are told in this dialogue.
	Rasmus-coaching-weekly-4-nicotine	June		This is part of Rasmus' fourth weekly coaching session and is about nicotine and sleep.
	Rasmus-coaching-weekly-4-nutrition	June		This is part of Rasmus' fourth weekly coaching session and is about the influence of food types on sleep.
	Rasmus-coaching-weekly-4-quiz	June		Each of Rasmus' sessions also has a quiz at the end about the information he provided. This is the fourth quiz.
	Ramus-social-story-1	June		Rasmus has a new background story available about sailing.
	Ramus-social-story-2	June		Rasmus has a new background story available about his travels as a student.

	Ramus-social-story-3	June		Rasmus has a new background story available about his internship in Tanzania.
Katarzyna	katarzyna-social-story-1	June		Katarzyna has a new background story available about her love for baking.
	katarzyna-social-story-2	June		Katarzyna has a new background story available about her education and occupation.
	katarzyna-social-story-3	June		Katarzyna has a new background story available about how she ended up becoming such a caretaker.
Coda	coda-explain-corona	March		Coda gives information about the COVID-19 situation and disclaimers to the advices given by the coaches.
	coda-explain-radio	March	April	Coda gives explanation about the radio. Minor textual fixed to the Dutch explanation of the radio by Coda.
	coda-explain-radio-ai	March		Coda can initiate a conversation about explaining the radio.
	coda-intake		April	Fixed a Dutch translation issue in Coda's intake dialogue.
	coda-menu		February March	Coda now offers additional dialogue to participants of the Council of Coaches evaluation to assist with questions or possibilities for providing feedback.
	coda-social-story-1	June		Coda has a new background story available about how it ended up in the Council of Coaches.
	coda-social-story-2	June		Coda has a new background story available about its experience with water.
	coda-social-story-3	June		Coda has a new background story available about its thoughts.

Acknowledgements

The Council of Coaches project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Headings and titles in this document, as well as the Council of Coaches logo use the Comfortaa font, designed by Johan Aakerlund and Cyreal and licensed under the Open Font License[1].

Additional text in this document uses the Roboto font, designed by Christian Robertson and licensed under the Apache License, Version 2.0[2].

The Council of Coaches logo and Blobmen graphics were *drawn freely* in Inkscape, licensed under the GNU General Public License[3].

[1] Open Font License: http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=OFL_web
[2] Apache License, Version 2.0: http://www.apache.org/licenses/LICENSE-2.0
[3] Inkscape License Information: https://inkscape.org/about/license/