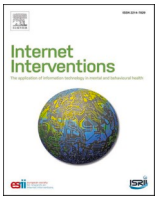
Internet Interventions 27 (2022) 100502

Contents lists available at ScienceDirect

Internet Interventions

journal homepage: www.elsevier.com/locate/invent



Automatic topic selection for long-term interaction with embodied conversational agents in health coaching: A micro-randomized trial

Tessa Beinema [a,b,*], Harm op den Akker [a,b,c], Marian Hurmuz [a,b], Stephanie Jansen-Kosterink [a,b], Hermie Hermens [a,b]

[a] eHealth Group, Roessingh Research and Development, Enschede, the Netherlands
[b] Biomedical Signals and Systems Group, University of Twente, Enschede, the Netherlands
[c] Innovation Sprint, Brussels, Belgium

ARTICLE INFO

Keywords:
Embodied conversational agents
Initiative
Tailoring
Topic selection
Dialogues
Health coaching
micro-randomized trial

ABSTRACT

Introduction: Embodied Conversational Agents (ECAs) can be included in health coaching applications as virtual coaches. The engagement with these virtual coaches could be improved by presenting users with tailored coaching dialogues. In this article, we investigate if the suggestion of an automatically tailored topic by an ECA leads to higher engagement by the user and thus longer sessions of interaction.

Methods: A Micro-Randomized Trial (MRT) was conducted in which two types of interaction with an ECA were compared: (a) the coach suggests a relevant topic to discuss, and (b) the coach asks the user to select a topic from a set of options. Every time the user would interact with the ECA, one of those conditions would be randomly selected. Participants interacted in their daily life with the ECA that was part of a multi-agent health coaching application for 4–8 weeks.

Results: In two rounds, 82 participants interacted with the micro-randomized coach a total of 1011 times. Interactions in which the coach took the initiative were found to be of equal length as interactions in which the user was allowed to choose the topic, and the acceptance of topic suggestions was high (71.1% overall, 75.8% for coaching topics).

Conclusion: Tailoring coaching conversations with ECAs by letting the coach automatically suggest a topic that is tailored to the user is perceived as a natural variation in the flow of interaction. Future research could focus on improving the novel coaching engine component that supports the topic selection process for these suggestions or on investigating how the amount of initiative and coaching approach by the ECA could be tailored.

1. Introduction

Digital behaviour change interventions (DBCIs) are increasingly investigated (Brinkman, 2016) as tools to support people in their health behaviour change process, both as a means of treating health conditions and in preventative contexts. These applications support users as needed and are always available. However, they face challenges in terms of adherence (Wangberg et al., 2008; Nijland, 2011; Crutzen et al., 2011; Kohl et al., 2013; Yardley et al., 2016). Potential causes for this lack of adherence are actively being researched. It appears that contributing factors are the lack of direct involvement of a healthcare professional (*no social incentive*) and content that does not always fit the user's personal situation (*relevance of content*) (e.g., Andersson et al. (2009); Buimer et al. (2017)). Two directions of research aimed at improving the interaction and engagement with these applications are therefore *Embodied Conversational Agents* (ECAs) and *tailoring*.

ECAs are "*more or less autonomous and intelligent software entities with an embodiment used to communicate with the user*" (Ruttkay et al., 2004). In DBCIs these agents can take on the role of a coach (Kramer et al., 2020) and they give a system social ability, which is important for maintaining a collaborative relationship (Bickmore et al., 2010; Kamphorst, 2017; Bickmore et al., 2018). ECAs make the use of health applications easier, more satisfying and less frustrating (André and Pelachaud, 2010; Bickmore et al., 2016) and potentially more effective (Ma et al., 2019). Furthermore, they are always available and their dialogues can be tailored dynamically to the user, for example by

Abbreviations: ECA, Embodied conversational agent; MRT, Micro-randomized trial.

* Corresponding author at: Biomedical Signals and Systems Group, University of Twente, Enschede, the Netherlands.
E-mail address: t.c.beinema@utwente.nl (T. Beinema).

https://doi.org/10.1016/j.invent.2022.100502
Received 26 July 2021; Received in revised form 27 January 2022; Accepted 2 February 2022
Available online 6 February 2022

elaborating on certain topics of discussion when needed (Bickmore and Giorgino, 2006). In health coaching applications, ECAs can be the main component (e.g., Sebastian and Richards (2017); op den Akker et al. (2018)) or they can be part of a broader application (e.g., van Velsen et al. (2020)).

Ultimately, regardless of their specific features and capabilities, health applications need to be *engaging* for a longer period of time to impact users' behaviour change (Yardley et al., 2016; Perski et al., 2017; Cole-Lewis et al., 2019). Or, as stated by Bickmore et al. (2010) in the context of ECAs: "*Engagement is a prerequisite for other system objectives: If a user stops interacting with a system, then it cannot have any further impact.*" For DBCIs, Cole-Lewis et al. (2019) distinguish 'Big E' (health behaviour engagement) and 'Little e' (DBCI engagement), with a subdivision for the latter in user interaction with a) features that encourage frequency of use and b) behaviour change intervention components. They also emphasise that 'Big E' is dependent on 'Little e'. In general for interaction with applications, short-term engagement tends to be characterised as flow (Nakamura and Csikszentmihalyi, 2002; Hamari et al., 2016) or enjoyment (O'Brien and Toms, 2013), while long-term engagement can be seen as the duration and depth of usage of a system over time (Couper et al., 2010). There are several objective measures for long-term engagement, such as the number of voluntary interactions users choose to have, the number of logins, and the amount and type of content used (Perski et al., 2017; Trinh et al., 2018).

Tailoring (op den Akker et al., 2014) is investigated to make a digital health application's content, such as the conversation with an ECA, more personally relevant for the specific user (e.g., Wangberg et al. (2008); Krebs et al. (2010); Ryan et al. (2019)). However, most research on tailoring coaching content has its focus on the last steps of presenting information (e.g., inserting tailored goals or adjusting the wording of sentences). In general, in health coaching applications with ECAs, participants either all follow the same coaching programme or a human expert manually defines the high-level structure of content to be presented (e.g., Abdullah et al. (2018); Benítez-Guijarro et al. (2019); Fadhil et al. (2019)). Where the first approach presents all participants with the same dialogues, the second approach requires continuous involvement of health professionals. Alternatively, participants could be allowed to select what they want to discuss, which can lead to more engagement because the user feels that they are in control (Perski et al., 2017). However, that approach does put the initiative with the user, while the coach should be the expert – or as Kamphorst (2017) states: "*an e-coaching system should be credible and proactive*". Therefore, we think that combining the first and second approaches by letting an ECA automatically suggest a relevant topic to discuss (both immediately and in the long-term), while allowing users to make the final decision, could be an improvement in the interaction with virtual coaches.

1.1. Research aims

In this article, we present a study in which participants interact with an ECA over a longer period of time in a daily life setting. The study investigates the influence of automatically tailoring coaching dialogues at the topic level on users' interaction with the application. Specifically, we let a coaching ECA take the initiative by suggesting a relevant topic to discuss, and we compare this with a more conventional approach in which the user selects a topic themselves. If such suggestion of relevant topics leads to longer interactions between users and the ECA, this would be a step towards extending tailoring methods for effectively coaching people to lead a healthy lifestyle. Our research question therefore is:

RQ: *What is the influence of automatically tailored topic suggestions on the length of interactions with an ECA?*

The use of ECAs, personal relevance and tailoring have all been found to affect engagement with DBCIs (e.g., Wangberg et al. (2008); Krebs et al. (2010); Krämer et al. (2010); Perski et al. (2017); Ryan et al. (2019)). Furthermore, manual tailoring of coaching topics or modules by human experts is also appreciated (Abdullah et al., 2018; Benítez-Guijarro et al., 2019; Fadhil et al., 2019). Therefore, we hypothesise that suggestion of a relevant topic will lead to increased engagement with the ECA, and thus longer interactions:

H1. *Suggestion of automatically tailored topics will lead to longer interactions with the ECA.*

In addition, we take a closer look at users' acceptance of topics that are suggested when the ECA takes the initiative, which gives an indication of the quality or relevance of those suggestions. We also perform an initial exploration of potential demographics that might be of influence on the acceptance of these suggestions, since various demographics have been found to influence engagement and appreciation of DBCIs (e.g., as reported in Hardiker and Grant (2011); Perski et al. (2017); van Velsen et al. (2019); Beinema et al. (2021)) and ECAs (e.g., Payne et al. (2013); Pezzullo et al. (2017)). Both these investigations can provide starting points for future work on automatically tailoring topics of conversation. Therefore, our second and third hypotheses are the following:

H2. *Participants will accept suggested topics more often than not.*

H3. *A participant's demographics affect their acceptance of a suggested topic.*

In the following sections, we will first provide some background on multi-agent health coaching applications, proactiveness in virtual coaches, and the micro-randomized trial method. Then we will provide details on the design and implementation for both conditions of the micro-randomized trial, and our methods for conducting the trial and analysis. Finally, we will present the results and discuss our findings and conclusions.

2. Background

2.1. Multi-agent health coaching

Changing behaviour can prevent or relieve health conditions, but change in the long-term tends to be difficult (Bouton, 2014). As stated previously, health coaching applications with ECAs can support users in this process. Most ECA applications feature a single ECA as a coach who has specific expertise (e.g., physical activity (Watson et al., 2012; King et al., 2017)). However, health often requires support in multiple domains (World Health Organization, 1946; Huber et al., 2016). This has led to single ECAs coaching on multiple domains (Gardiner et al., 2017; Klaassen et al., 2018) and recently multiple coaches coaching on multiple domains (op den Akker et al., 2018; Das et al., 2019; Kramer et al., 2021). Having multiple coaches available at the same time provides opportunities for vicarious persuasion (Kantharaju et al., 2018) and engagement (André and Rist, 2001). Each agent can have a specific expertise and role – for example, a dietitian or personal trainer – and multiple viewpoints can be presented without an ECA contradicting itself (Hayashi and Ogawa, 2012; Kantharaju et al., 2019). We perform our experiment in a setting where multiple ECAs are present and interact with the user so that the results can be incorporated in tailoring approaches for a broad range of ECA health coaching applications.

2.2. Proactiveness in virtual coaches

Taking the initiative or being proactive, is an important property for a virtual coach. As previously stated by Kamphorst (2017), an e-coaching system needs to, for example, invite the user to reflect on their commitment to a goal or warn them at suspected moments of weakness. This requires that the system is flexible enough to respond to new developments and can start communication about those topics.

When it comes to starting interactions, ECAs that are proactive were found to be better in providing support (e.g., on loneliness to older

adults (Ring et al., 2013)). Proactive agents can also be perceived to be more helpful, even if their proactiveness does not immediately improve task performance (Xiao et al., 2002). It is however important to use the right tone of voice when being proactive, for example, when giving reminders to users during working hours (Bickmore et al., 2007). Furthermore, while a proactive coach could provide the right coaching at the right moment, in the end, a virtual coach should support the user and not just dictate how they should behave (Brinkman, 2016). Thus, research on the proactiveness of virtual coaches could learn from developments in shared-decision making research – both between humans (e.g., on facilitators and barriers (Joseph-Williams et al., 2014)) as well as between humans and ECAs (Zhang and Bickmore, 2018).

2.3. Micro-randomized trial

Properly assessing the effectiveness of technology-supported health services in real-world settings is challenging and there is a need for pragmatic study designs (Ekeland et al., 2010, 2012; Kairy et al., 2009; LaPlante and Peng, 2011). The *Micro-Randomized Trial* (MRT) is a method of evaluating interventions originally proposed by Klasnja et al. (2015) for the evaluation of *Just-in-Time Adaptive Interventions* (JITAIs). They found that conventional methods such as randomized trials were not suitable for evaluating these JITAIs. In a micro-randomized trial, an intervention option is randomly selected at every relevant decision point (e.g., whether or not to send a notification). Furthermore, in a MRT, effect is measured after each intervention through a short-term parameter that resembles the intended long-term effect. In our evaluation, the initiative for choosing a topic is randomized every time the user interacts with a coach, and we measure the length of the interaction that immediately follows.

3. Methods

We performed a MRT to compare users' responses to the coach suggesting a topic (coach-initiative) with a more conventional implementation of coaching dialogues in which users could select a topic themselves (user-initiative). The micro-randomized trial was embedded in the final evaluation of the Council of Coaches application (op den Akker et al., 2018), which consisted of two separate rounds with participants. The full protocol for that evaluation is described in an article by Hurmuz et al. (2020). Since the MRT shared the same participants and setup as the full evaluation, we will only summarise the important aspects of the overall procedure that are relevant for the MRT, while elaborating on the design and implementation of the MRT itself.

3.1. The multi-agent eHealth application

In the Council of Coaches application (op den Akker et al., 2018; Hurmuz et al., 2020), users can interact with multiple ECAs. Each of these ECAs has their own role, expertise and backstory. There are six coaches with expertise in the following domains: physical activity, nutrition, social activity, cognition, chronic pain and diabetes. In addition, there was an agent that provided peer support and an agent that guided the user through the application (the assistant). After an intake with the assistant, users could select their council of coaches. The physical activity coach and nutrition coach were obligatory, and the diabetes and chronic pain coaches were only available to those who had indicated in the intake to have those conditions.

An example interaction with the application can be found in Fig. 1. Users could start an interaction with one of the coaches in the application by clicking on a coach of their choice. The main participants in such an interaction are the user and a specific coach, but the other coaches can also join in to provide their own viewpoint on the ongoing conversation. The interactions followed a speech-bubble and reply-button paradigm, as depicted in Fig. 1.

3.2. Study design

The MRT was specifically set up for one of the two obligatory coaches, the physical activity coach (Olivia). In the MRT, we randomized two types of interaction. The first is an interaction in which the user decides what they would like to discuss. The second is an interaction in which the coach automatically suggests a topic to discuss based on current user parameters (see Section 3.3.4 for details). Thus, there were

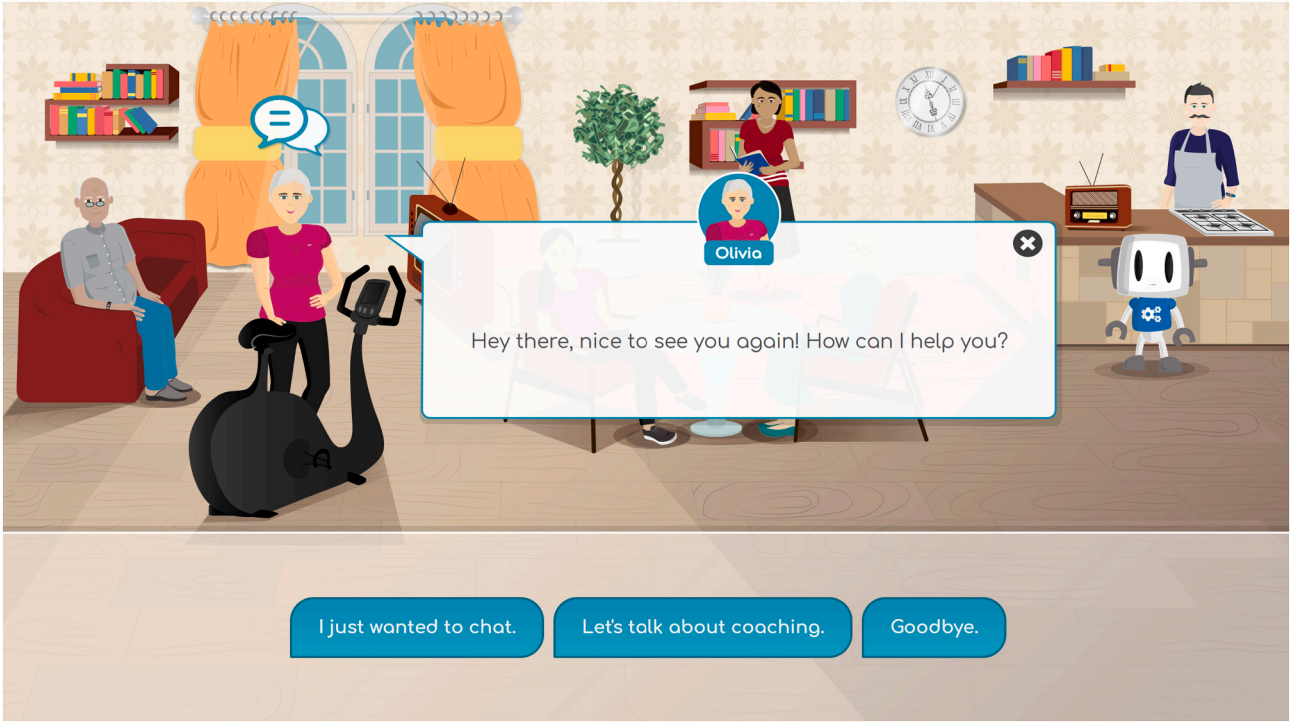


Fig. 1. An example interaction with the physical activity coach (Olivia) in the multi-agent eHealth application.

two conditions:

Condition 1: The user gets the initiative and chooses the topic of conversation (*user-initiative*).

Condition 2: The ECA takes the initiative and suggests the topic of conversation (*coach-initiative*).

Every time the user would click on the physical activity coach (to start the interaction), the system would micro-randomly select one of the two conditions; both with a 50% chance (see Fig. 2). The difference between the two conditions was in the start of the interaction through the preceding menu-dialogue and 'start [topic]' dialogues, while the dialogues that followed were the same.

3.3. Implementation

The health coaching application was implemented for use on a tablet, laptop or desktop using a client-server setup. The WOOL Dialogue Platform (Beinema et al., 2022) was used for dialogue authoring and execution. To facilitate the implementation of both conditions in the MRT, we structured the topics that could be discussed as an hierarchical tree. This tree can be found in Fig. 3.

3.3.1. Content

Dialogue content was created for the eight topics in the topic tree, namely:

Introduction. A conversation between the coach and the user in which the coach introduces herself to the user and provides some information on her background and the type of coaching content the user might expect from her.

Background story. Social dialogues in which the coach shares a part of her background story with the user. For example, a short story about Olivia's dog ('Brian') and how she likes to go running with him.

Discuss sensors. Dialogues that cover subtopics such as 'Connecting your activity tracker.', 'Why should I use an activity tracker?', and 'My sensor is not working, why is that?'

Goal-setting. Dialogues in which the user can set a new long- and short-term goal, or change their current goal.

Feedback. Dialogues that allow the user to view their measured activity data. In these dialogues, the coach can also show the user their 'activity book' widget, which provides an overview of their physical activity (steps taken) over the past week.

Gather information. Dialogues in which the coach asks the user for information that can be used to tailor or personalise the coaching provided. For example, the question 'Do you have a dog?' can be used to suggest that the user walks with their dog more often as a form of increased activity.

Inform 'why'. Dialogues in which the coach explains why it is good to be physically active. For example, 'Being active increases blood flow, which is healthy for your brain'.

Inform 'how'. Dialogues in which the coach gives advice on how to be more physically active. For example, 'Take the stairs instead of the elevator'.

For every topic, dialogue scripts were written, which were available in both Dutch and English. In addition, for the user-initiative condition, a menu-dialogue was created that users could use to select a topic.

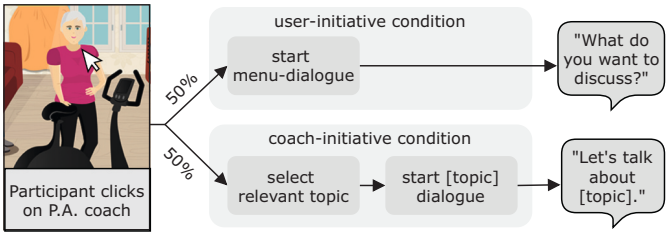


Fig. 2. A schematic representation of the procedure in the MRT.

Furthermore, for the coach-initiative condition a 'start [topic] dialogue' was created to precede each topic. We elaborate on this in the following subsections.

3.3.2. User-initiative

In the user-initiative condition, the user could select a topic to discuss. We defined a 'menu'-dialogue that facilitated this. When the user would click on the physical activity coach and did not yet complete the *Introduction* dialogue, the first step in the menu-dialogue would lead them to that *Introduction* dialogue. For each subsequent interaction, the coach would state 'Hey there, nice to see you again! How can I help you?' (resembling *Start* in the topic model). The user could then respond with: 'I just wanted to chat.' (leading to *Social*), 'Let's talk about coaching.' (leading to *Coaching*) or 'Goodbye' (ending the interaction). If the user would then click on 'Let's talk about coaching', the coach would say 'Let's get down to business and talk about some coaching. What do you want to discuss today?' and the user could select: 'My goals' (*Goal-Setting*), 'My activity tracker' (*Discuss Sensors*), 'Tips and info' (*Health Education*, which would be followed by a choice between *Inform 'Why'* and *Inform 'How'*), 'My progress' (*Feedback*), or 'Goodbye.' (ending the interaction). In this manner, the menu-dialogue would allow the user to navigate towards a dialogue on the topic of their preference in line with the topic structure as depicted in Fig. 3.

3.3.3. Coach-initiative

In the coach-initiative condition, the coach would suggest a topic to discuss. This suggested topic was selected by our coaching engine component (original concept by Beinema et al. (2018)). This coaching engine applied a topic selection algorithm that took into account parameters such as dialogue availability, completion dates and data prerequisites.

In this condition, the dialogues for the different topics would be preceded by a short 1-step dialogue in which the coach suggests the topic. Such a statement could be, for example, 'Would you like me to tell you something about how you can be more active?'. The user could then respond with (a variation of) 'Yes, that would be nice.' (accepting the suggestion) or 'Goodbye.' (ending the interaction). In the second round of the study, users would additionally have the option to reply 'I would like to discuss something else.' (rejecting the suggestion and being forwarded to the menu-dialogue that was also used in the user-initiative condition).

3.3.4. The topic selection algorithm

For the coach-initiative condition, the topic model was implemented as a tree (with the topics as nodes). Each topic was assigned a set of selection parameters that resembled aspects that contributed to that topic's relevance (positively or negatively). These parameters each had a weight assigned to them and their value was dependent on the information that was stored for a specific user (e.g., their age, previously completed dialogues, or available data). In addition to these selection parameters, topics were also assigned an a-priori weight and value. For the first round, these were set to 0.5 and 0.5 for all topics. For the second round, these were changed to 0.0 and 0.0 for all topics. This was done when also adding additional dialogue content between rounds because lowering the residual relevance of topics seemed like an improvement for the resulting suggestions. To give a specific example, selection parameters for the 'Goal-setting' topic included whether an activity tracker had already been connected, if a goal had already been set, and the time since that goal had been discussed last. This would then cause the topic to be most relevant to discuss when a tracker was connected, no goals were set, and the topic had not been discussed yet.

The *relevance* for a topic was computed by taking the weighted average of the parameters' weights and values using the following formula (with p being the number of selection parameters for a topic):

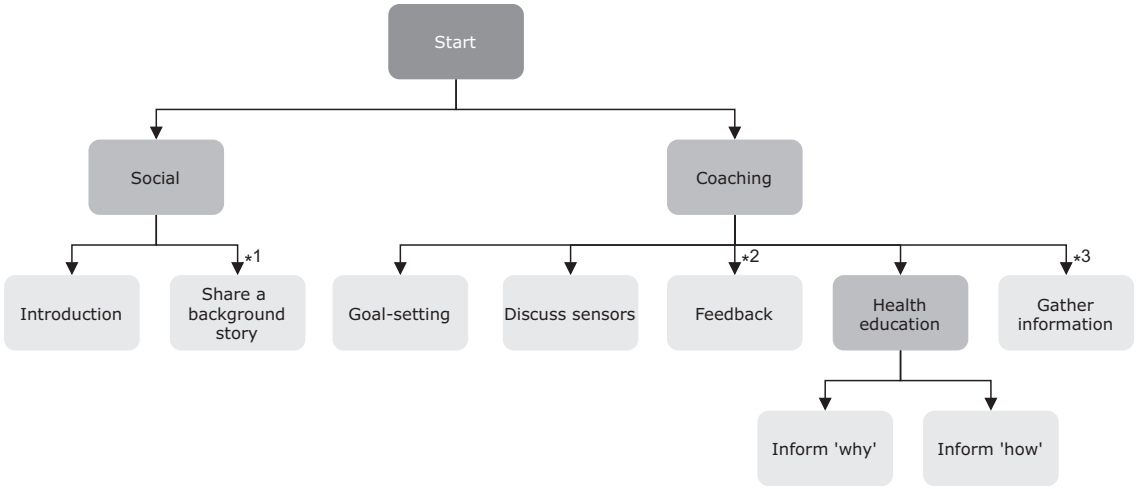


Fig. 3. The topic model featuring the topics for which dialogues could be held with the physical activity coach. *[1] The *Background Story* topic was added for the second round of the study, and *[2] the *Feedback* topic was extended. *[3] The *Gather Information* topic was only available in the coach-initiative condition.

$$Relevance = \frac{w_{a-priori} \cdot v_{a-priori} + \sum_{i=1}^{p} w_i \cdot v_i}{w_{a-priori} + \sum_{i=1}^{p} w_i} \quad (1)$$

Whenever a topic has to be suggested, the selection algorithm starts at the top of the tree (*Start*), computes the relevance for all the direct subtopics and then select one of those subtopics (in this case *Social* or *Coaching*). Depending on the balance between exploration and exploitation that is set for the algorithm, the selection is made from a weighted distribution (more relevant topics have a higher chance of being selected) or by selecting the topic with the highest relevance. Between the first and second round of this study, the exploration probability was increased from 0% to 25%, since it provided a little more variation of selected topics if parameters were close together in terms of relevance. The selection process continues until a topic is selected that has no further subtopics.

3.4. Participants

The study was aimed at older adults (55 years or older). Inclusion criteria were that they had to be able to read and speak Dutch or English, had a WiFi connection at home, were able to provide informed consent, and were able to see a smartphone/tablet screen clearly. In the end, due to difficulty of recruitment, age requirements were slightly relaxed in Scotland for the second round. Participants were recruited through advertisements in local newspapers and on social media in the Netherlands and Scotland. Beforehand participants were informed that they would receive a small gift to thank them for participating, independent from how actively they participated. Recruitment took place in two rounds, each preceding the corresponding round in the study.

3.5. Procedure

The MRT was conducted in two rounds as was the procedure for the overall evaluation of the Council of Coaches application (Hurmuz et al., 2020) (see Fig. 4). That evaluation in a real-life setting was conducted in two rounds for pragmatic reasons such as to ensure that enough human support and devices (e.g., tablets) were available.

At T_0, a researcher met with participants to provide them with the technology, create an account and complete the intake with the assistant agent, and let the participant complete the baseline questionnaire (on demographics and health status). In the one-week baseline phase that followed, participants wore the activity tracker, but they did not yet use the coaching application. This phase was included to ensure activity

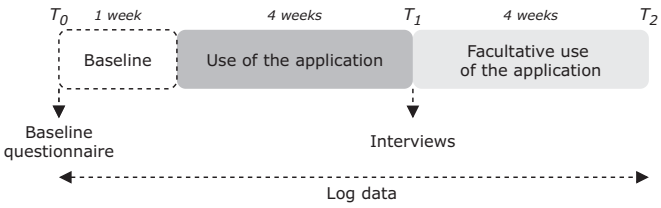


Fig. 4. A timeline showing the procedure for the micro-randomized trial.

data would be present when they would start using the application and allowed for the novelty effect of the tracker to wear off. After this week, the participants were asked to use the application for four weeks as they wanted (they received no instructions for frequency). At T_1 participants were interviewed (additional questions for the MRT were added in round 2) and they could indicate whether they wanted to use the application for an additional four weeks (the facultative use phase). If participants did not want to continue using the application the study would end at T_1, otherwise, it would end at T_2.

Due to Covid-19 and our target population of older adults, for the second round, the same overall procedure was followed, but direct contact between researchers and participants was limited. Materials were sent to participants by post, explanations were given over the phone, and interviews were also conducted by phone.

3.6. Outcomes and measurements

The main outcomes of this study were length of the interaction (i.e., number of dialogue steps), users' acceptance of topic suggestions by the coach (i.e., accepting/declining the topic suggested by the coach) and influence of demographics on acceptance of suggestions. Secondary outcomes were users' experience of the coach suggesting topics (i.e., noticing it, and if they were positive, neutral or negative about it).

Three types of data were collected for the MRT. The first type of data came from the demographic questions that were included in the baseline questionnaire at T_0. Participants were asked for their gender (male/female), age, highest level of education (primary education, secondary education, further education or higher education) and living situation (alone, married/living together, living together with my caregiver, other). They were also asked the three items from the health literacy scale by Chew et al. (2004) (on a five-point Likert scale), four items on attitude towards technology (Agarwal and Prasad, 1998) (on a seven-point Likert scale), questions on their motivation to live healthy (van

Velsen et al., 2019) (on a seven-point Likert scale) and for their self-reported physical activity level (not at all, not at all but thinking about beginning, less than 2.5 h a week, more than 2.5 h a week in the last six months, more than 2.5 h a week for more than six months).

The second type of data came from the system's interaction logs. These logs included a record for every dialogue that was started with a coach. Stored information per dialogue included:

- *Dialogue steps*. Each statement by a coach or reply by a user is counted as one dialogue step.
- *'Cancellation' boolean*. Whether the dialogue was cancelled by clicking on the 'X'-button of the coach's speech-bubble.
- *'Completed' boolean*. Whether the dialogue was completed, that is, the dialogue ended because it was finished or the user ended the dialogue by responding with a 'Goodbye'-reply.
- *Condition*. Indicating experimental condition: user-initiative or coach-initiative.
- *Referrer*. If the dialogue was started because the user was referred to it from another dialogue (and if so, which one).

The number of dialogue steps was the parameter that we used to quantify *interaction length*, as it provided a clear short-term parameter for the amount of interaction with the agent. The 'Cancellation' and 'Completed' booleans and the first user-response to a suggestion were used to analyse the acceptance of suggestions in the coach-initiative condition.

The third type of data was participants' responses to interview questions. We added these questions to the interviews that were conducted at T_1 for the second round. These questions were the following: 'Did you interact with the physical activity coach?', and (if 'yes') 'Did you notice that the physical activity coach sometimes came with a suggestion for a topic?' and (if 'yes') 'What did you think of this?'.

3.7. Pre-processing of log data

The outcome parameter that we used to determine the length of an interaction was *the number of dialogue steps*. An *interaction* started when the user clicked on a coach, and could then involve a chain of several dialogues and ended when the last dialogue was completed (e.g., the user choosing 'Goodbye'), cancelled (closing the speech-bubble with the 'X'-button) or when it was not completed or cancelled, but a new dialogue was logged (e.g., when the browser was refreshed). A user-initiative interaction started with a menu-dialogue that had no referrer and a coach-initiative interaction started with one of the 'Shall we discuss X'-dialogues.

We applied the following pre-processing steps:

1. We excluded logs for dialogues that were the result of a 'double click' error. That is, two dialogues are started within 1 s of each other, with the first log only including the agent's first statement.
2. The 'sensor connection completed' dialogue that was initiated by the system after successful connection of a sensor was not automatically labelled with the correct condition. We manually relabelled dialogues following this event with the same label as given to the preceding 'connect sensor' dialogue.
3. The 'sensor connected' dialogue was triggered when the system registered that it could retrieve data. In some cases, there was a delay for the start of that dialogue after participants returned from the external connection page. Sometimes, the participant had started a new dialogue with the coach themselves, which would then be interrupted. We removed such interrupted dialogues if they were only one dialogue step long. If the new dialogue was longer than one step, we removed both the interfering and the interfered dialogue, since neither could reach their full number of dialogue steps (and in some cases, both dialogues belonged to different conditions).

To conclude pre-processing, we aggregated the information for dialogues that were part of one interaction. The resulting data set contained one row per interaction with information such as a participant identifier, experimental condition, number of dialogue steps, whether the last dialogue in the interaction was completed or cancelled, and a transcript of statements and replies.

3.8. Analysis

IBM SPSS Statistics version 25 was used for data analysis. Since we made adjustments to the available content and the settings of the topic selection algorithm between the two evaluation rounds, we analysed both rounds separately, except for the analysis about investigating the influence of demographics on acceptance.

Even though we analysed both rounds separately, we did compare demographics for participants between both rounds to get an insight into how these two groups compared. For the continuous demographics age, health literacy, attitude towards technology, intrinsic motivation, external regulation and a-motivation we performed independent-samples *t*-tests. For the categorical demographics gender and country we performed Pearson Chi-Square tests, but for living situation we performed a Fisher's exact test since some cells contained less than 5 items. Finally, for the ordinal demographics level of education and self-reported physical activity we performed Mann-Whitney *U* tests.

To test our first hypothesis that interactions in the coach-initiative condition would be longer (in terms of dialogue steps) than interactions in the user-initiative condition, a generalised estimating equations (GEE) analysis was performed. Of the two methods named by Klasnja et al. (2015) in the paper introducing the MRT (the GEE and MLM), the GEE supposedly makes less explicit assumptions on distribution and can handle smaller clusters better. Participant numbers were added as the subject variable. An exchangeable structure was selected for the working correlation matrix. We checked our intended dependent variable (number of steps) for normality and decided to include it with a natural log transformation applied. Condition (user-initiative or coach-initiative) was included as the main predicting factor in our model. All other settings were set to the standard options.

The interview data from the added questions in the second round were analysed to gather insights into users' experience of the change in initiative. The full interviews were recorded and transcribed for the larger evaluation (Hurmuz et al., 2020). The specific answers to the three interview questions that we added for this study were listed in an Excel file. In that file, we counted the number of 'yes' and 'no' responses for the 'Did you interact with the coach?' and 'Did you notice the coach suggesting a topic?' questions. Two authors (TB and HO) categorised participants' responses to the 'What did you think of this?' follow-up question as positive, neutral or negative. Differences in category assignment (25%) were discussed and a final category was assigned.

To assess the acceptance of topics that were suggested by the coach in the coach-initiative condition (as is relevant for our second hypothesis), we created an overview table that listed the number of accepted and rejected suggestions per topic for each participant. From that table we then computed the overall number of suggestions that were accepted (user agreed to discuss the topic by selecting the 'Yes, that would be nice.' response) or rejected (user chose the 'goodbye' response, closed the speech-bubble, did not respond at, or selected the 'I would like to discuss something else' option in the second round). We also computed the percentage of user-initiative interactions in which the coach only made one statement to get an idea of the rejection rate in that condition.

Finally, to explore if certain user characteristics might be linked to acceptance or rejection of suggested topics (our third hypothesis) we observed individual users' responses. We included participants from both rounds that had had more than 10 total interactions in the coach-initiative condition. A non-parametric Kendall's Tau correlation was performed between ordinal and continuous demographics and the percentage of accepted suggestions. Specifically, these demographics were:

age, self-reported physical activity, health literacy, education, attitude towards technology, intrinsic motivation, external regulation, and a-motivation. Mann-Whitney *U* tests were performed to test if there was a difference in acceptance for country (NL, SC) and gender (male, female). We did not include the living situation demographic in our tests, since almost all participants were married and/or living together.

3.9. Ethical approval

As previously stated, the micro-randomized trial was included in a larger evaluation (Hurmuz et al., 2020). That evaluation 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, the 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 the University of Dundee. Each participant gave his/her informed consent on paper beforehand.

4. Results

4.1. Participants

In the first round, 44 participants created an account and 40 interacted with the ECA that embodied the MRT (23 NL, 17 SC). A full overview of all demographics can be found in Table 1. The mean age of these MRT participants was 65.35 (SD = 7.35). Most of them were female (67.5%). They had a good health literacy (M = 4.35, SD = 0.67) and had a slightly positive attitude towards technology (M = 4.46, SD = 1.17). Their levels of intrinsic motivation were high and their levels of external regulation or a-motivation relatively low when it came to living healthy. Most of them (57.5%) had completed higher vocational education or university-level education. Most were married or living together (75%) and they were quite active with 52.5% being active for more than 2.5 h per week.

In the second round, 46 participants created an account and 42 interacted with the ECA that embodied the MRT (24 NL, 18 SC). The mean age of participants was 62.12 (SD = 8.68). Although the target population was adults ageing 55 years and older, 4 participants were included that were 40, 42, 47 and 54. We decided to keep these participants included. Most of the participants were female (71.4%). They had a good health literacy (M = 4.32, SD = 0.62) and a slightly positive attitude towards technology (M = 4.57, SD = 1.53). This second group also had high levels of intrinsic motivation and relatively low levels of external regulation or a-motivation when it came to living healthy. The level of education was high (61.9% higher vocational or university), they were mostly married or living together (76.2%) and they were quite active with 59.5% being active for more than 2.5 h per week.

Analyses showed that there were no significant differences in participants' demographics between the two rounds.

Table 1
Demographics of participants in the MRT.

Demographic	Category	Round 1 (N = 40)	Round 2 (N = 42)
		M (SD) or N (%)	M (SD) or N (%)
Age		65.35 (SD = 7.35)	62.12 (SD = 8.68)
Gender	Male	13 (32.5%)	12 (28.6%)
	Female	27 (67.5%)	30 (71.4%)
Country	Netherlands	23 (57.5%)	24 (57.1%)
	Scotland	17 (42.5%)	18 (42.9%)
Health literacy		4.35 (SD = 0.67)	4.32 (SD = 0.62)
Attitude towards technology		4.46 (SD = 1.17)	4.57 (SD = 1.53)
Motivation to live healthy	Intrinsic motivation	5.19 (SD = 1.11)	5.07 (SD = 0.92)
	External regulation	2.82 (SD = 1.26)	3.14 (SD = 1.19)
	A-motivation	2.28 (SD = 1.45)	2.19 (SD = 1.06)
Level of education	Preparatory secondary vocational education	8 (20.0%)	3 (7.1%)
	Higher general secondary education, pre-university education	9 (22.5%)	13 (31.0%)
	Higher vocational education, university	23 (57.5%)	26 (61.9%)
Living situation	Married/living together	30 (75%)	32 (76.2%)
	Alone	9 (22.5%)	10 (23.8%)
	Other	1 (2.5%)	0 (0.0%)
Self-reported physical activity	Not at all	4 (10.0%)	1 (2.4%)
	Not at all, but thinking about beginning	1 (2.5%)	3 (7.1%)
	<2.5 h a week	14 (35.0%)	13 (31.0%)
	>2.5 h a week in the last six months	12 (30.0%)	14 (33.3%)
	>2.5 h a week for more than six months	9 (22.5%)	11 (26.2%)

4.2. Collected log data

In the first round, 6077 logged dialogues were collected for all participants, and for the second round, 6222 dialogues were collected for all participants. Pre-processing of those logged dialogues led to 2384 and 2210 dialogues with the ECA who embodied the MRT. These dialogues amounted to 568 interactions in the first round and 443 in the second round. Fig. 5 shows flowcharts illustrating these steps for both rounds.

4.3. Comparing interaction length

The distribution of dialogue steps for both conditions in both rounds can be found in Table 2. The GEE showed that there was no significant difference in length between interactions in the user-initiative and coach-initiative conditions for either of the two rounds (see Table 3). Therefore, we *cannot accept our hypothesis that the coach taking the initiative will lead to longer interactions (H1)*. User-initiative and coach-initiative interactions *were of equal length*.

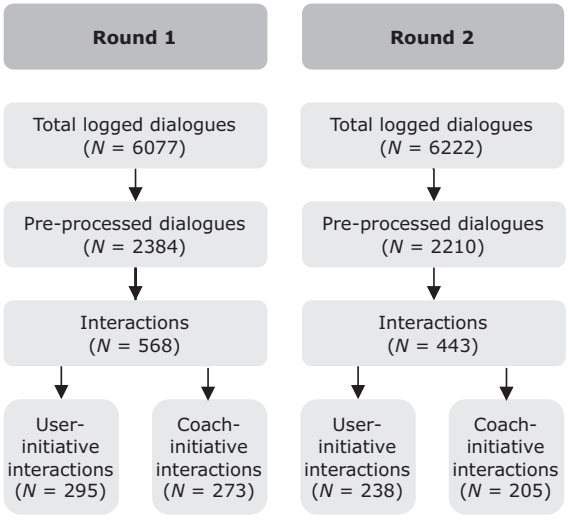


Fig. 5. Two flowcharts illustrating the number of collected dialogues, dialogues after pre-processing and interactions for both conditions. Note that multiple dialogues chained together form one interaction.

Table 2
Distribution of the number of dialogue steps in interactions for both conditions in both rounds, and the natural log transform of the number of dialogue steps (which was used in the GEE).

Round	Initiative	*N* steps	*Ln*(*N* steps)
		M (*SD*)	*M* (*SD*)
1	User	25.29 (22.53)	2.70 (1.21)
	Coach	22.09 (19.25)	2.47 (1.36)
2	User	28.47 (28.47)	2.90 (1.14)
	Coach	24.91 (20.29)	2.72 (1.20)

Table 3
Results for the generalised estimating equations (GEE) analysis for both rounds.

Round	Beta	Std. Error	*p*	Wald χ^2
1	.239	.1272	.060	3.531 (1)
2	.186	.1640	.256	1.290 (1)

4.4. Interview results

Three interview questions were included in the interviews for the second round. Fig. 6 shows a flowchart that provides an overview of the participant numbers as described in this section. Of the 42 participants, five participants did not have an interview (11.9%). Two participants indicated that they did not interact with the ECA that embodied the MRT (4.8%), two said they only completed the introduction dialogue with her (4.8%), and one said they did not interact with the coach, but actually did interact twice according to the logs (2.4%).

The other 32 participants had interacted with the physical activity coach. Almost all of them had been exposed to the coach-initiative dialogue condition (N = 30, 71.4%), and most of them also had accepted and discussed a suggested topic with the coach (N = 28, 66.7%). However, only 16 of them (38.1%) stated that they had noticed that the coach sometimes took the initiative.

Of the 16 participants that noticed an initiative change, eight did not have a preference for either user- or coach initiative and five seemed to like suggestions by the coach (e.g., '*Thought that was good. Also directed me to the other coaches on a few occasions.*' [P13]), motivating that it provided the option to not have to think about something and that the coach was pleasantly surprising. One participant stated that it gave a stimulus to discuss something and indicated that some people need that (himself included). Three participants were not interested in suggestions (e.g., '*I was more interested in my progress.*' [P19]).

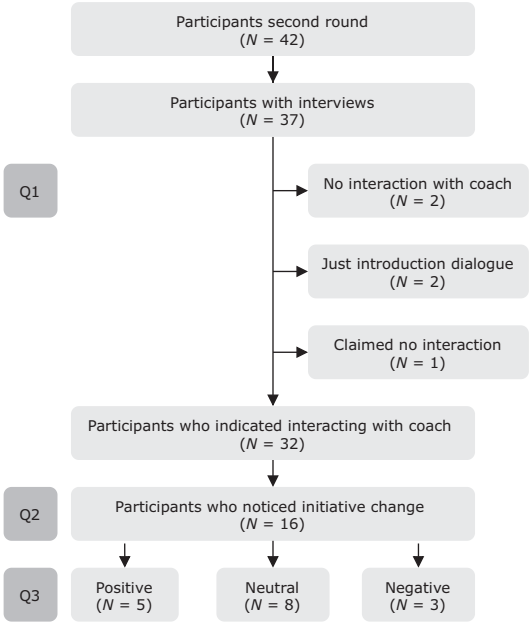


Fig. 6. A flowchart proving an overview of participant numbers for the interviews. (Q1, Q2, and Q3 refer to the first, second and third interview questions respectively.)

Participants' opinions also seemed to be reflected in their responses to the suggestions, for example, the person who thought it was pleasantly surprising and enjoyed the interaction accepted 13 out of 14 suggestions, while the person interested in their progress accepted 5 of the 12 suggestions, chose the '*I want to discuss something else*' reply 6 times, and cancelled the dialogue after 1 suggestion.

In two cases there was a clear discrepancy between logged responses to suggestions and participants' expressed opinion. For example, one did not like suggestions by the coach, and motivated this by stating that she wanted direct and specific coaching advice and not social conversations. However, her log data did show that she agreed to discuss 75% of the suggested topics. Another stated that she always had a topic in mind that she wanted to discuss and that she therefore did not accept suggestions from the coach. She however had agreed to discuss all suggested topics about coaching, and only rejected all social topics.

4.5. Acceptance of suggested topics

In the first round, the coach took the initiative by suggesting a topic to discuss in 273 interactions. As can be seen in Table 4, the overall acceptance rate for these suggestions was high with 213 accepts (78.0%). In 60 cases (22.0%), the suggested topic was not accepted by the user. We must note, however, from the 16 'Goodbye.' rejections, 8 were in response to a dialogue where 'no additional information was available', and 4 of the 27 cancellations were in response to that statement.

Overall, most suggested topics were well received in the first round. The acceptance rate was high for the topics *Introduction* (26 out of 38; 68.4%), *Goal-Setting* (32 out of 35; 91.4%), *Gather Information* (20 out of 22; 90.9%), and *Inform* (69 out of 79; 87.4%). The *Sensors* topic was accepted less by participants (66 out of 99; 66.7%). Of those 33 rejections, 15 were cancellations in response to the suggestion to connect a sensor; mostly by participants that on a later prompt did agree to discuss that topic.

In the second round, the coach took the initiative by suggesting a topic to discuss in 205 interactions. As can be seen in Table 4, the overall acceptance rate for suggestions in this round was lower than that of the first round (62.0%). There was also a large difference in acceptance between suggestions for social and coaching topics (48.6% and 69.2%, respectively). Furthermore, in this round, participants had the option to indicate that they wanted to discuss something else. This meant that even though the suggestion by the coach was not accepted, the conversation still continued.

Overall, suggestions for coaching topics were again well received. *Goal-Setting* was accepted 2 out of 3 times (66.7%), and the *Feedback* topic 17 out of 18 (94.4%). While the *Gather Information* topic was accepted 18 out of 38 times (47.4%), its suggestion also resulted in 9 changes of topic (23.7%) and 9 cancellations (23.7%). The *Inform* topics were accepted 31 out of 42 times (73.8%), and 7 times the topic was changed (16.7%).

Suggestions for social topics were not always welcomed. Participants accepted the *Introduction* topic 13 out of 23 times (56.5%), and the suggestion for the coach to *Share a Background Story* about herself was only accepted 22 out of 49 times (44.9%), while a change of topic was requested 19 times (38.7%).

To get a sense of rejections in the user-initiative condition, the percentage of cases in which the coach only got to make one statement in that condition was also computed. For the first round, this was the case in 11.2% of interactions (33 out of 295), and for the second round 9.7% (23 out of 238) of the interactions were rejected, which amounted to

Table 4
The acceptance of topic suggestions by a coach in the coach-initiative condition. We present acceptance and rejection numbers for both rounds separately, and both combined. Furthermore, for the two separate rounds, we also provide details on how a topic was rejected, namely: *changed* when the user selected 'I want to discuss something else', *goodbye* if they choose to end the interaction with 'Goodbye.', *cancelled* if they closed the coach's speech-bubble, and *nothing* if a user did not respond at all.

Round	Topics	*N*	*N* accepted	*N* rejected	*N* rejected reason			
					Changed	Goodbye	Cancelled	Nothing
1	Social	38	26 (68.4%)	12 (31.6%)	n.a.	5 (13.2%)	7 (18.4%)	0 (0.0%)
	Coaching	235	187 (79.6%)	48 (20.4%)	n.a.	16 (6.8%)	27 (11.5%)	5 (2.1%)
	Overall	273	213 (78.0%)	60 (22.0%)	n.a.	21 (7.7%)	34 (12.5%)	5 (1.8%)
2	Social	72	35 (48.6%)	37 (51.4%)	25 (34.7%)	7 (9.7%)	4 (5.6%)	1 (1.4%)
	Coaching	133	92 (69.2%)	41 (30.8%)	23 (17.3%)	3 (2.3%)	14 (10.5%)	1 (0.7%)
	Overall	205	127 (62.0%)	78 (38.0%)	48 (23.4%)	10 (4.9%)	18 (8.8%)	2 (0.9%)
Both	Social	110	61 (55.5%)	49 (44.5%)				
	Coaching	368	279 (75.8%)	89 (24.2%)				
	Overall	478	340 (71.1%)	138 (28.9%)				

10.5% of user-initiative interactions over both rounds.

The *high acceptance rates* for the coach-initiative condition indicate that even though suggestions by the coach did not lead to longer interactions, the suggestions that the coach made *were suitable*. We therefore accept our second hypothesis (*H2: Participants will accept suggested topics more often than not.*).

4.6. Demographics and individual acceptance

Finally, a possible influence of demographics on acceptance of coach-suggestions was explored. From both rounds, 18 participants had 10 or more coach-initiative interactions with the MRT coach. The correlations between ordinal or continuous demographics and the percentage of accepted suggestions can be found in Table 5. As can be seen, there was a moderately strong correlation between attitude towards technology and percentage of accepted suggestions ($r_\tau = 0.48, p = .007$).

The Mann-Whitney U test for country showed that there was no significant difference ($U = 30.0$, $p = .574$) between the acceptance of Dutch ($Mdn = 0.78$, $IQR = 0.55–0.82$) and Scottish ($Mdn = 0.59$, $IQR = 0.39–0.90$) participants. For gender there also was no significant difference ($U = 34.0$, $p = .851$) between the acceptance of Male ($Mdn = 0.75$, $IQR = 0.55–0.84$) and Female ($Mdn = 0.70$, $IQR = 0.46–0.89$) participants.

The correlation between attitude towards technology and percentage of accepted suggestions leads us to *accept our hypothesis that there is an influence of demographics on topic acceptance* (H3).

5. Discussion

In this article we investigated the influence of a virtual coach taking the initiative and suggesting a relevant (tailored) topic to discuss on users' interaction with that coach. To that end, we implemented two versions of a coach's dialogue (coach-initiative and user-initiative) and compared these in a micro-randomized trial where participants interacted with the coach in their daily life over a longer period of time.

Table 5
Kendall's Tau correlation of users' demographics with percentage of accepted suggestions.

Demographic	Correlation	
	r_τ	p
Age	.09	.620
Self-reported physical activity	−.06	.743
Health literacy	−.11	.561
Education	−.12	.551
Attitude towards technology	.48	.007
Intrinsic motivation	.12	.492
External regulation	.06	.732
A-motivation	.21	.255

Our first hypothesis was that the coach taking the initiative by suggesting a relevant topic to discuss would lead to longer interactions than when a user had to select a topic themselves. An underlying assumption for this hypothesis was that the suggestions to discuss specific and relevant topics could have an engaging effect. The topics selected by our algorithm were selected using up-to-date parameter values from each user's profile. Previous studies showed that manual adjustment of coaching content for virtual coaches was beneficial (e.g., Abdullah et al. (2018); Fadhil et al. (2019)). However, our results showed that there was no difference in length between interactions in the user-initiative and coach-initiative conditions. While this does not support the hypothesis, our findings do show that a coach taking the initiative was equally engaging as asking the user to indicate what they would like to discuss. Likewise, the results from the interviews taught us that participants either did not notice that the coach took the initiative or were fine with it. This suggests that the change of initiative was perceived as a natural variation in the interaction and that the coach taking the initiative was not perceived as being disruptive to the flow of conversation. While this lack of awareness of the manipulation seems to stand out, a similar finding was reported by Olafsson et al. (2019). They performed a manipulation in which they removed the possibility of participants to be able to respond negatively to suggestions by a health coaching ECA, and found that participants did not notice that lack of choice. Furthermore, while there might not have been a difference in dialogue length, the suggestions by the coach might have influenced relational parameters that we were not able to measure after every interaction (e.g., perceived helpfulness as found by Xiao et al. (2002), or preference for an ECA (Olafsson et al., 2019)).

The equality between the coach-initiative and user-initiative conditions, both objectively (interaction length) and subjectively (user experience), is interesting since they are actually quite unequal when it comes to 'freedom of choice'. That is, where the algorithm in the coach-initiative condition only suggests one specific topic, in the user-initiative condition users have the full set of topics to choose from. Having a coach suggest a relevant topic to discuss could lead to higher engagement with DBCIs because of personal relevance and tailoring of content, and perhaps also because of novelty, a sense of narrative and guidance (Perski et al., 2017). However, suggesting a specific topic has to be done right. That is, if that single topic you suggest is not relevant for the participant, they will rightfully reject your suggestion. The high acceptance rate (71.1%) for topic choices by the coach suggests that our algorithm did select topics that were relevant and suitable for users. This also supports our second hypothesis. One might wonder if users would just accept all suggestions, but the difference between the acceptance of social (55.5%) and coaching topics (75.8%) indicates that users did care which topics were suggested. Overall, we conclude that our underlying coaching engine component seemed to have selected topics with a high enough relevance for users. Potential factors that could have influenced topic acceptance could be the task-mindedness and independence of

users, since some indicated in the interviews that they already had a clear purpose in mind when starting an interaction. We therefore advise to include these in future research. In addition, the correlation between a positive attitude towards technology and acceptance of topic suggestions suggests that participants' openness to using a health coaching application to begin with is also a factor. Furthermore, even though finding a correlation between one of the demographics and acceptance of topic suggestions was enough to accept our third hypothesis, we do note that this was just one demographic and strongly recommend to perform further investigations for all demographics in future research as their influence may also be dependent on the application, target population or domain.

We see at least two possible directions for future research that relate to the improvement of the coaching engine's topic selection and the more general content design processes. First, the balance between social and coaching topics could be improved. While our results show that suggestions for social topics were not always welcomed, previous research found that background stories and other relational behaviours are important for enjoyment and engagement with an ECA (Bickmore et al., 2010; Trinh et al., 2018). We therefore suggest that instead of simply adjusting the frequency of social topic suggestions, a system could learn a user's interest in social dialogue by taking into account their responses to social comments. These responses can be measured when a topic is suggested, but also for all similar remarks or social sidesteps in dialogues about coaching topics. Furthermore, other predictors may be used in modelling a participant's interest in social interactions. One example of such a predictor in the context of social robots was whether a participant greeted a robot before interacting (Lee et al., 2010).

Second, a similar approach could be applied to improving the suggestion of coaching subtopics. The ECA and the user could have an explicit discussion on preferences for coaching style or strategy (e.g., as investigated in Beinema et al. (2021)). For example, when it comes to deciding what to discuss or the balance between coach- and user-initiative. These investigations could benefit from including lessons from research on shared decision making (Joseph-Williams et al., 2014; Zhang and Bickmore, 2018). Another option could be to implement classification functionality or self-learning mechanisms to determine different types of users based on previous digital health coaching research (e.g., type of motivation to live healthy (van Velsen et al., 2019) or stages of change (de Vries et al., 2016)). This would support further tailoring of initiative and strategies. Such models could benefit from our finding that participants with a more positive attitude towards technology could have a higher acceptance rate for suggested topics, which supported our third hypothesis (influence of demographics on acceptance).

5.1. Strengths and limitations

A strength of this study was that it was a study conducted over a longer period of time (4–8 weeks) in users' daily life. Participants were asked to use the application at will over a period of at least four weeks, which could be extended by another four weeks. This meant that every recorded interaction with the system was a) voluntary and without the possible influence of a researcher's presence, and b) these interactions extended past the first two weeks in which a novelty effect can still be present (Nijland, 2011). Furthermore, we evaluated a novel implementation of an interaction condition in which coaching dialogues were automatically tailored at the topic level (by introducing a coaching engine), and we conducted the (to our knowledge) first micro-randomized trial in the context of embodied conversational coaches.

There were also some limitations. First, participants were recruited through advertisements in local newspapers and on social media, which meant that selection bias was an issue as discussed in the published study protocol (Hurmuz et al., 2020). This probably caused our participants to have relatively high scores on intrinsic motivation and health literacy, and relatively high levels of education. These are all factors that are associated with participants being more active in managing their own health, which might have influenced the way they interacted with the system. Second, during the study, the COVID-19 outbreak reached the Netherlands and Scotland. From that point on, the study was performed remotely, that is, materials were sent to participants by post and interviews were conducted by phone. Due to difficulty of recruitment, age requirements were slightly relaxed in Scotland for the second round. Nevertheless, we have no indication that these procedural changes affected participants interactions with the application.

6. Conclusion

Tailoring coaching conversations with embodied conversational agents (ECAs) has the potential to increase the engagement of users with those coaches, which is deemed a prerequisite for desired behaviour change. The main finding from this micro-randomized trial is that coaching conversations with ECAs can be automatically tailored on a topic level, and that the resulting suggestions by the coach were perceived as a natural variation in the flow of the interaction with a high user acceptance of those suggestions. This is an important step towards more intelligent and engaging health coaching applications. Future work should investigate how to further improve the automatic topic suggestion process, and how the amount of initiative, the coaching strategies and the coaching style applied by the coach could be adjusted to specific types of users.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553 (Council of Coaches).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Dominic De Franco, Ellis Oude Kempers, and Katrien Fischer for their part in participant recruitment and conducting and transcribing the interviews.

References

Abdullah, A.S., Gaehde, S., Bickmore, T.W., 2018. A tablet based embodied conversational agent to promote smoking cessation among veterans: a feasibility study. J. Epidemiol. Glob. Health 8, 225–230. https://doi.org/10.2991/j.jegh.2018.08.104.

Agarwal, R., Prasad, J., 1998. A conceptual and operational definition of personal innovativeness in the domain of information technology. Inf. Syst. Res. 9, 204–215. https://doi.org/10.1287/isre.9.2.204.

Andersson, G., Carlbring, P., Berger, T., Almlöv, J., Cuijpers, P., 2009. What makes internet therapy work? Cogn. Behav. Ther. 38, 55–60. https://doi.org/10.1080/16506070902916400.

André, E., Pelachaud, C., 2010. Interacting with embodied conversational agents. Speech Technol. 123–149. https://doi.org/10.1007/978-0-387-73819-2_8.

André, E., Rist, T., 2001. Presenting through performing: on the use of multiple lifelike characters in knowledge-based presentation systems. Knowl.-Based Syst. 14, 3–13. https://doi.org/10.1016/S0950-7051(00)00096-4.

Beinema, op den Akker, H., Hermens, H.J., 2018. Creating an artificial coaching engine for multi-domain conversational coaches in eHealth applications. In: André, E., Bickmore, T.W., Vrochidis, S., Wanner, L. (Eds.), ICA-HoGeCa '18: Proceedings of the AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications co-located with the Federated AI Meeting, CEUR, pp. 35–39.

Beinema, T., op den Akker, H., van Velsen, L., Hermens, H.J., 2021. Tailoring coaching strategies to users ' motivation in a multi-agent health coaching application. Comput. Hum. Behav. 121, 106787 https://doi.org/10.1016/j.chb.2021.106787.

Beinema, T., op den Akker, H., Hofs, D., van Schooten, B., 2022. The WOOL dialogue platform: enabling interdisciplinary user-friendly development of dialogue for

conversational agents [version 1; peer review: awaiting peer review]. Open Res. Eur. 2, 1–11. https://doi.org/10.12688/openreseurope.14279.1.
Benítez-Guijarro, A., Ruiz-Zafra, A., Callejas, Z., Medina-Medina, N., Benghazi, K., Noguera, M., 2019. General architecture for development of virtual coaches for healthy habits monitoring and encouragement. Sensors 19, 108. https://doi.org/10.3390/s19010108.
Bickmore, T.W., Giorgino, T., 2006. Health dialog systems for patients and consumers. J. Biomed. Inform. 39, 556–571. https://doi.org/10.1016/j.jbi.2005.12.004.
Bickmore, T.W., Mauer, D., Crespo, F., Brown, T., 2007. Persuasion, task interruption and health regimen adherence. In: PERSUASIVE '07: Proceedings of the 2nd International Conference on Persuasive Technology. Springer, pp. 1–11. https://doi.org/10.1007/978-3-540-77006-0_1.
Bickmore, T.W., Schulman, D., Yin, L., 2010. Maintaining engagement in long-term interventions with relational agents. Appl. Artif. Intell. 24, 648–666. https://doi.org/10.1080/08839514.2010.492259.
Bickmore, T.W., Utami, D., Matsuyama, R., Paasche-Orlow, M.K., 2016. Improving access to online health information with conversational agents: a randomized controlled experiment. J. Med. Internet Res. 18, e5239 https://doi.org/10.2196/jmir.5239.
Bickmore, T.W., Trinh, H., Asadi, R., Olafsson, S., 2018. Safety first: conversational agents for health care. In: Studies in Conversational UX Design, pp. 33–57. https://doi.org/10.1007/978-3-319-95579-7_3.
Bouton, M.E., 2014. Why behavior change is difficult to sustain. Prev. Med. 68, 29–36. https://doi.org/10.1016/j.ypmed.2014.06.010.
Brinkman, W.P., 2016. Virtual health agents for behavior change: research perspectives and directions. In: GREATS '16: Proceedings of the Workshop on Graphical and Robotic Embodied Agents for Therapeutic Systems, Held During the 16th International Conference on Intelligent Virtual Agents (IVA '16), pp. 1–17.
Buimer, H.P., Tabak, M., van Velsen, L., van der Geest, T., Hermens, H.J., 2017. Exploring determinants of patient adherence to a portal-supported oncology rehabilitation program: interview and data log analyses. JMIR Rehabil. Assist. Technol. 4, e12 https://doi.org/10.2196/rehab.6294.
Chew, L.D., Bradley, K.A., Boyko, E.J., 2004. Brief questions to identify patients with inadequate health literacy. Fam. Med. 36, 588–594.
Cole-Lewis, H., Ezeanochie, N., Turgiss, J., 2019. Understanding health behavior technology engagement: pathway to measuring digital behavior change interventions. JMIR Form. Res. 3, 1–10. https://doi.org/10.2196/14052.
Couper, M.P., Alexander, G.L., Zhang, N., Little, R.J.A., Maddy, N., Nowak, M.A., McClure, J.B., Calvi, J.J., Rolnick, S.J., Stopponi, M.A., Johnson, C.C., 2010. Engagement and retention: measuring breadth and depth of participant use of an online intervention. J. Med. Internet Res. 12, e52 https://doi.org/10.2196/jmir.1430.
Crutzen, R., Cyr, D., de Vries, N.K., 2011. Bringing loyalty to e-health: theory validation using three internet-delivered interventions. J. Med. Internet Res. 13, e73 https://doi.org/10.2196/jmir.1837.
Das, K.S.J., Beinema, T., op den Akker, H., Hermens, H.J., 2019. Generation of multi-party dialogues among embodied conversational agents to promote active living and healthy diet for subjects suffering from type 2 diabetes. In: ICT4AWE '19: Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health, pp. 297–304. https://doi.org/10.5220/0007750602970304.
de Vries, R.A.J., Truong, K.P., Kwint, S., Drossaert, C.H.C., Evers, V., 2016. Crowd-designed motivation: Motivational messages for exercise adherence based on behavior change theory. In: CHI '16: Proceedings of the 2016 ACM Conference on Human Factors in Computing Systems, ACM, pp. 297–308. https://doi.org/10.1145/2858036.2858229.
Ekeland, A.G., Bowes, A., Flottorp, S., 2010. Effectiveness of telemedicine: a systematic review of reviews. Int. J. Med. Inform. 79, 736–771. https://doi.org/10.1016/j.ijmedinf.2010.08.006.
Ekeland, A.G., Bowes, A., Flottorp, S., 2012. Methodologies for assessing telemedicine: a systematic review of reviews. Int. J. Med. Inform. 81, 1–11. https://doi.org/10.1016/j.ijmedinf.2011.10.009.
Fadhil, A., Wang, Y., Reiterer, H., 2019. Assistive conversational agent for health coaching: a validation study. Methods Inf. Med. 58, 9–23. https://doi.org/10.1055/s-0039-1688757.
Gardiner, P.M., McCue, K.D., Negash, L.M., Cheng, T., White, L.F., Yinusa-Nyahkoon, L., Jack, B.W., Bickmore, T.W., 2017. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: a feasibility randomized control trial. Patient Educ. Couns. 100, 1720–1729. https://doi.org/10.1016/j.pec.2017.04.015.
Hamari, J., Shernoff, D.J., Rowe, E., Coller, B., Asbell-Clarke, J., Edwards, T., 2016. Challenging games help students learn: an empirical study on engagement, flow and immersion in game-based learning. Comput. Hum. Behav. 54, 170–179. https://doi.org/10.1016/j.chb.2015.07.045.
Hardiker, N.R., Grant, M.J., 2011. Factors that influence public engagement with eHealth: a literature review. Int. J. Med. Inform. 80, 1–12. https://doi.org/10.1016/j.ijmedinf.2010.10.017.
Hayashi, Y., Ogawa, H., 2012. Facilitating creative interpretations on collaboration with multiple conversational agents. In: APCHI '12: Proceedings of the 10th Asia Pacific Conference on Computer Human Interaction, pp. 443–449.
Huber, M., van Vliet, M., Giezenberg, M., Winkens, B., Heerkens, Y., Dagnelie, P.C., Knottnerus, J.A., 2016. Towards a 'patient-centred' operationalisation of the new dynamic concept of health: a mixed methods study. BMJ Open 6, 1–11. https://doi.org/10.1136/bmjopen-2015-010091.
Hurmuz, M.Z.M., 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 Res. Protoc. 9, e16641 https://doi.org/10.2196/16641.
Joseph-Williams, N., Elwyn, G., Edwards, A., 2014. Knowledge is not power for patients: a systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. Patient Educ. Couns. 94, 291–309. https://doi.org/10.1016/j.pec.2013.10.031.
Kairy, D., Lehoux, P., Vincent, C., Visintin, M., 2009. A systematic review of clinical outcomes, clinical process, healthcare utilization and costs associated with telerehabilitation. Disabil. Rehabil. 31, 427–447. https://doi.org/10.1080/09638280802062553.
Kamphorst, B.A., 2017. E-coaching systems: what they are, and what they aren't. Pers. Ubiquit. Comput. 21, 625–632. https://doi.org/10.1007/s00779-017-1020-6.
Kantharaju, R.B., De Franco, D., Pease, A., Pelachaud, C., 2018. Is two better than one? effects of multiple agents on user persuasion. In: IVA '18: Proceedings of the 18th ACM International Conference on Intelligent Virtual Agents, pp. 255–262.
Kantharaju, R.B., Pease, A., Reidsma, D., Pelachaud, C., Snaith, M., Bruijnes, M., Klaassen, R., Beinema, T., Huizing, G., Simonetti, D., Heylen, D., op den Akker, H., 2019. Integrating argumentation with social conversation between multiple virtual coaches. In: IVA '19: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, pp. 203–205.
King, A.C., Campero, I., Sheats, J.L., Castro Sweet, C.M., Garcia, D., Chazaro, A., Blanco, G., Hauser, M., Fierros, F., Ahn, D.K., Diaz, J., Done, M., Fernandez, J., Bickmore, T.W., 2017. Testing the comparative effects of physical activity advice by humans vs. computers in underserved populations: the COMPASS trial design, methods, and baseline characteristics. Contemp. Clin. Trials 61, 115–125. https://doi.org/10.1016/j.cct.2017.07.020.
Klaassen, R., Bul, K.C.M., op den Akker, R., van der Burg, G.J., Kato, P.M., Di Bitonto, P., 2018. Design and evaluation of a pervasive coaching and gamification platform for young diabetes patients. Sensors 18, 1–27. https://doi.org/10.3390/s18020402.
Klasnja, P., Hekler, E.B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., Murphy, S.A., 2015. Micro-randomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol. 34, 1220–1228. https://doi.org/10.1016/j.physbeh.2017.03.040.
Kohl, L.F.M., Crutzen, R., de Vries, N.K., 2013. Online prevention aimed at lifestyle behaviors: a systematic review of reviews. J. Med. Internet Res. 15, e146 https://doi.org/10.2196/jmir.2665.
Krämer, N.C., Hoffmann, L., Kopp, S., 2010. Know your users! empirical results for tailoring an agent's nonverbal behavior to different user groups. In: IVA '10: Proceedings of the 10th ACM International Conference on Intelligent Virtual Agents, pp. 468–474. https://doi.org/10.1007/978-3-642-15892-6_50.
Kramer, L.L., ter Stal, S., Mulder, B.C., de Vet, E., van Velsen, L., 2020. Developing embodied conversational agents for coaching people in a healthy lifestyle: scoping review. J. Med. Internet Res. 22, e14058 https://doi.org/10.2196/14058.
Kramer, L.L., Mulder, B.C., van Velsen, L., de Vet, E., 2021. Use and effect of web-based embodied conversational agents for improving eating behavior and decreasing loneliness among community-dwelling older adults: protocol for a randomized controlled trial. JMIR Res. Protoc. 10, e22186 https://doi.org/10.2196/22186.
Krebs, P., Prochaska, J.O., Rossi, J.S., 2010. A meta-analysis of computer-tailored interventions for health behavior change. Prev. Med. 51, 214–221. https://doi.org/10.1016/j.ypmed.2010.06.004.
LaPlante, C., Peng, W., 2011. A systematic review of e-health interventions for physical activity: an analysis of study design, intervention characteristics, and outcomes. Telemed e-Health 17, 509–523. https://doi.org/10.1089/tmj.2011.0013.
Lee, M.K., Kiesler, S., Forlizzi, J., 2010. Receptionist or information kiosk: how do people talk with a robot?. In: CSCW '10: Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 31–40. https://doi.org/10.1145/1718918.1718927.
Ma, T., Sharifi, H., Chattopadhyay, D., 2019. Virtual humans in health-related interventions: a meta-analysis. In: CHI '19: Proceedings of the 2019 ACM Conference on Human Factors in Computing Systems, pp. 1–6.
Nakamura, J., Csikszentmihalyi, M., 2002. The concept of flow. In: Handbook of Positive Psychology, pp. 89–105. https://doi.org/10.1002/9780470172698.ch19.
Nijland, N., 2011. Grounding eHealth - Towards a Holistic Framework for Sustainable eHealth Technologies. University of Twente. https://doi.org/10.3990/1.9789036531337. Ph.D. thesis.
O'Brien, H.L., Toms, E.G., 2013. What is user engagement? A conceptual framework for defining user engagement with technology. J. Am. Soc. Inf. Sci. Technol. 59, 938–955. https://doi.org/10.1002/asi.
Olafsson, S., O'Leary, T., Bickmore, T.W., 2019. Coerced change-talk with conversational agents promotes confidence in behavior change. In: PervasiveHealth '19: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 31–40. https://doi.org/10.1145/3329189.3329202.
op den Akker, H., Jones, V.M., Hermens, H.J., 2014. Tailoring real-time physical activity coaching systems: a literature survey and model. User Model. User-Adap. Inter. 24, 351–392. https://doi.org/10.1007/s11257-014-9146-y.
op den Akker, H., op den Akker, R., Beinema, T., Banos, O., Heylen, D., Pease, A., Pelachaud, C., Traver Salcedo, V., Kyriazakos, S., Hermens, H.J., 2018. Council of coaches: a novel holistic behavior change coaching approach. In: ICT4AWE '18: Proceedings of the 4th International Conference on Information and Communication Technologies for Ageing Well and e-Health, pp. 978–989. https://doi.org/10.5220/0006787702190226.
Payne, J., Szymkowiak, A., Robertson, P., Johnson, G., 2013. Gendering the machine: preferred virtual assistant gender and realism in self-service. In: IVA '13: Proceedings of the 13th ACM International Workshop on Intelligent Virtual Agents, pp. 106–115. https://doi.org/10.1007/978-3-642-40415-3_9.

Perski, O., Blandford, A., West, R., Michie, S., 2017. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl. Behav. Med. 7, 254–267. https://doi.org/10.1007/s13142-016-0453-1.

Pezzullo, L.G., Wiggins, J.B., Frankosky, M.H., Min, W., Boyer, K.E., Mott, B.W., Wiebe, E.N., Lester, J.C., 2017. "Thanks alisha, keep in touch": gender effects and engagement with virtual learning companions. In: AIED '17: Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2017), pp. 299–310. https://doi.org/10.1007/978-3-319-61425-0_25.

Ring, L., Barry, B., Totzke, K., Bickmore, T.W., 2013. Addressing loneliness and isolation in older adults: proactive affective agents provide better support. In: ACII '13: Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, pp. 61–66. https://doi.org/10.1109/ACII.2013.17.

Ruttkay, Z., Dormann, C., Noot, H., 2004. Embodied conversational agents on a common ground: a framework for design and evaluation. In: From Brows to Trust: Evaluating Embodied Conversational Agents. Kluwer, pp. 27–66.

Ryan, K., Dockray, S., Linehan, C., 2019. A systematic review of tailored eHealth interventions for weight loss. Dig. Health 5, 1–29. https://doi.org/10.1177/2055207619826685.

Sebastian, J., Richards, D., 2017. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. Comput. Hum. Behav. 73, 479–488. https://doi.org/10.1016/j.chb.2017.03.071.

Trinh, H., Shamekhi, A., Kimani, E., Bickmore, T.W., 2018. Predicting user engagement in longitudinal interventions with virtual agents. In: IVA '18: Proceedings of the 18th ACM International Conference on Intelligent Virtual Agents, pp. 9–16. https://doi.org/10.1145/3267851.3267909.

van Velsen, L., Broekhuis, M., Jansen-Kosterink, S.M., op den Akker, H., 2019. Tailoring persuasive eHealth strategies for older adults on the basis of personal motivation: an online survey. J. Med. Internet Res. 21, e11759 https://doi.org/10.2196/11759.

van Velsen, L., Cabrita, M., op den Akker, H., Brandl, L., Isaac, J., Suárez, M., Gouveia, A., de Sousa, R.D., Rodrigues, A.M., Canhão, H., Evans, N., Blok, M., Alcobia, C., Brodbeck, J., 2020. LEAVES (optimizing the mentaL health and resiliencE of older Adults that haVe lost thEir spouSe via blended, online therapy): proposal for an online service development and evaluation. JMIR Res. Protoc. 9, e19344 https://doi.org/10.2196/19344.

Wangberg, S.C., Bergmo, Trine, S., Johnson, J.A.K., 2008. Adherence in internet-based interventions. Patient Prefer. Adherence 2, 57–66.

Watson, A., Bickmore, T.W., Cange, A., Kulshreshtha, A., Kvedar, J., 2012. An internet-based virtual coach to promote physical activity adherence in overweight adults: randomized controlled trial. J. Med. Internet Res. 14, e1 https://doi.org/10.2196/jmir.1629.

World Health Organization, 1946. Constitution of the World Health Organization. Technical Report.

Xiao, J., Stasko, J., Catrambone, R., 2002. Embodied conversational agents as a UI paradigm: a framework for evaluation. In: Proceedings of the AAMAS02 Workshop on Embodied Conversational Agents - Let's Specify and Evaluate Them!.

Yardley, L., Spring, B.J., Riper, H., Morrison, L.G., Crane, D.H., Curtis, K., Merchant, G. C., Naughton, F., Blandford, A., 2016. Understanding and promoting effective engagement with digital behavior change interventions. Am. J. Prev. Med. 51, 833–842. https://doi.org/10.1016/j.amepre.2016.06.015.

Zhang, Z., Bickmore, T.W., 2018. Medical shared decision making with a virtual agent. In: IVA '18: Proceedings of the 18th ACM International Conference on Intelligent Virtual Agents, pp. 113–118. https://doi.org/10.1145/3267851.3267883.