CHAPTER 1

Embodied Conversational Agents for health

A comparison of Empathic versus Neutral Dialogue

SANA SALMAN; DEBORAH RICHARDS; AND MARK DRAS

ABSTRACT:

Games technology can be used to build Embodied Conversational Agents (ECAs) that tend to have a humanlike appearance and exhibit humanlike verbal and non-verbal behaviours. When it comes to digital health, ECAs can provide vital support to patients by being more reachable through their smart phones or other digital gadgets like IPad/computers. To encourage users to follow the advice given, the use of empathy and development of a working alliance is recommended. While researchers have looked at expression of empathy via non-verbal cues and certain characteristics of the patient, this study focuses on making ECAs more effective by using human-like empathy expressed during conversation through relational cues. Relational cues are the empathic utterances that build a more long term alliance among the patient and the therapist, for example, the choice of words in social dialogues or being polite during a conversation. The goal of this study is to measure the working alliance with a neutral versus an empathic ECA. Furthermore, we evaluate the impact of the relational cues on the change in the behaviours recommended by the agent. The main findings of the study establish that empathic Alex is able to change the behaviours of the users more than the neutral Alex. The likability and

working alliance with the virtual coach is rated high by most of the users. Future work has been derived through the feedback received from the users revolving around having more personalised and advanced content as well as more interactivity.

KEYWORDS:

Game technology, Embodied Conversational Agents, relational cues, working alliance, empathic versus neutral

1. INTRODUCTION

Motivating individuals to follow a health regime can be difficult. In healthcare, effective face to face therapy is frequently underpinned by the development of collaborative rapport between a therapist and a patient, this is often referred to as a working or therapeutic alliance (Abdulrahman et al., 2021). The strength of this working alliance has been found to significantly influence the success of behaviour change interventions (Tremain et al., 2020). However, face to face therapy is costly. We propose the use of games technology to create an embodied conversational agent that provides health advice in a manner that motivates the user to follow the advice. Serious games are a technology that can be used to train and educate people while keeping them engaged through active learning and feedback. Apart from user experience and digital engagement, serious games must have an outcome associated with them that the users need to work through during the game.

Evidence exists that working alliances can be established between users and digital embodied conversational agents (Bickmore et al., 2005; Tremain et al., 2020). By using a conversational style of dialogue, an embodied conversational agent can create a social environment, which provides an opportunity to develop a relationship with the user. Studies have found that relationships with embodied conversation agents can be enhanced through the agent providing empathic and relational cues (Abdulrahman et al., 2021; Moore, 2021). Research on empathic agents has also shown that these agents are perceived to be more likeable, trustworthy and caring

(Brave et al., 2005) and can increase the interaction length and user engagement (Yalcin, 2019).

In our domain of health behaviour change, we are using games technology to create a virtual coach, known as Alex. Alex's dialogues were designed with a goal of establishing a working alliance with players. Alex aims to engage players through their own choice of responses; hence, it contains content for both highly motivated and also non-motivated players and takes them through a journey of self-discovery and routine establishment. It represents an ideal instrument for continuous health education also in terms of costs because it is cheaper than traditional training methods.

This paper reports a study which compares an empathic version of Alex with a neutral version of Alex to determine the impact of these conversational styles on working alliance and intended health behaviour change. Our study asks the following research questions.

Research Question 1 (RQ1): Do ECA's empathic dialogues motivate the player to change their health behaviours more than the neutral dialogues?

Research Question 2 (RQ2): Do ECA's empathic dialogues build a stronger working alliance with the player than the neutral dialogues?

Research Question 3 (RQ3): What is the player's experience of the empathic ECA compared to the neutral ECA?

In the next section we present some background literature. Section 3 presents our methodology followed by results in Section 4. Section 5 discusses our findings. Conclusions and future work are provided in Section 6.

2. BACKGROUND LITERATURE

2.1 MOTIVATIONAL COACHING IN THE HEALTH DOMAIN

There are several key components (Levensky et al., 2007) of motivational interviewing that can be used to successfully conduct clinical practices of counselling patients:

1. Express empathy for the patients.

2. Develop a discrepancy
3. Roll with resistance
4. Support self-efficacy

There are certain therapeutic skills (McCarley, 2009) that need to be applied to act upon the core components which are:

1. Resist the righting reflex-describe this
2. Use reflective listening
3. Ask open ended questions
4. Affirm and summarize

According to (Hall et al., 2012), these four skills are grouped into a motivational counselling method called RULE (Resist the righting reflex; Understand the patient's own motivations; Listen with empathy; and Empower the patient), described further below.

The righting reflex describes the tendency of health professionals to advise patients about the right path for good health. Essentially, most people resist persuasion when they are ambivalent about change and will respond by recalling their reasons for maintaining the behaviour. Motivational interviewing in practice requires clinicians to suppress the initial righting reflex so that they can explore the patient's motivations for change and also build long term working alliance with the patient.

Understand your patient's motivations involves knowing what is the patient's own reasons for change, rather than the practitioner's, that will ultimately result in behaviour change. By approaching a patient's interests, concerns and values with curiosity and openly exploring the patient's motivations for change, the practitioner will begin to get a better understanding of the patient's motivations and potential barriers to change.

Listen with empathy requires effective listening skills that are essential to understand what will motivate the patient, as well as the pros and cons of their situation. A general rule-of-thumb in motivational counselling is

that equal amounts of time in a consultation should be spent listening and talking.

Empower your patient is driven by the recognition that patient outcomes improve when they are an active collaborator in their treatment. Empowering patients involves exploring their own ideas about how they can make changes to improve their health and drawing on the patient's personal knowledge about what has succeeded in the past. A truly collaborative therapeutic relationship is a powerful motivator. Patients benefit from this relationship the most when the practitioner also embodies hope that change is possible.

In summary we can see that the first three involve active listening where you get the player to speak and the final element is providing education and options to the player to choose from.

2.2 SERIOUS GAMES FOR HEALTH COACHING

Drummond, Hadchouel, and Tesnière (2017) describes a research based conceptual process for serious games design upon which our project is built. This idea consists of motivational convergence and evidence based education and was also utilized in other medical related serious games including virtual simulation for emergency medicine departments (McGrath et al., 2018) in which active learning and feedback are the core components used in the situational context of emergency patient treatments. Other serious games for health coaching include: LISSA that is used in nurses training where Cardiopulmonary resuscitation (CPR) is used as a first aid technique for keeping the blood flowing (Boada et al., 2015) and training surgical residents on treating biliary tract disease (Graafland et al., 2014).

Three main steps are involved during the development of a serious game like a health coaching application: Motivating effect, learning effectiveness and evaluation. First is building the learning activity on the extrinsic motivation of the user which is the desire to achieve a certain outcome and then intrinsically motivating the user through a series of desirable steps. This phenomenon is called "convergence of motivations". Second

is utilizing the four pillars of learning framework referred to in cognitive science findings as active learning, attention, consolidation and feedback to enhance the learning potential of health applications. Finally comes the evaluation part which is critical to progress towards evidence based education through A/B testing and controlled trials.

3. METHODOLOGY

Our study uses a between subject experimental, pre-post study design to assess whether a socially engaging empathic embodied conversational agent (empathic Alex) is able to build a better working alliance than the neutral one (neutral Alex). The dialogues are built after extensive research on relational cues that are considered vital in conversations led by digital as well as human coaches. We obtained ethics approval from our university Human Research Ethics Committee.

3.1 RECRUITMENT

Recruitment is done through an online research participation portal at the host university. The students are undergraduates enrolled in a first year psychology course who belong to multiple discipline areas including psychology, computing, health sciences, business, arts and others. They receive course credit for their participation after completion of the survey. During a recruitment period of 36 days, 217 students registered on the portal for this activity.

3.2 MATERIALS

Alex was implemented using the Council of Coaches (COUCH) architecture containing the Windows Object Oriented Language (WOOL) dialogue engine. The conversation was displayed on the screen in text (Figure 1) through a web based front end which is built on top of COUCH and handles the dialogue exchange through Application Programming Interface (API) calls built using Python and JavaScript Object Notation (JSON) utilities. Further description of the dialogues, COUCH and WOOL are provided in the following subsections.

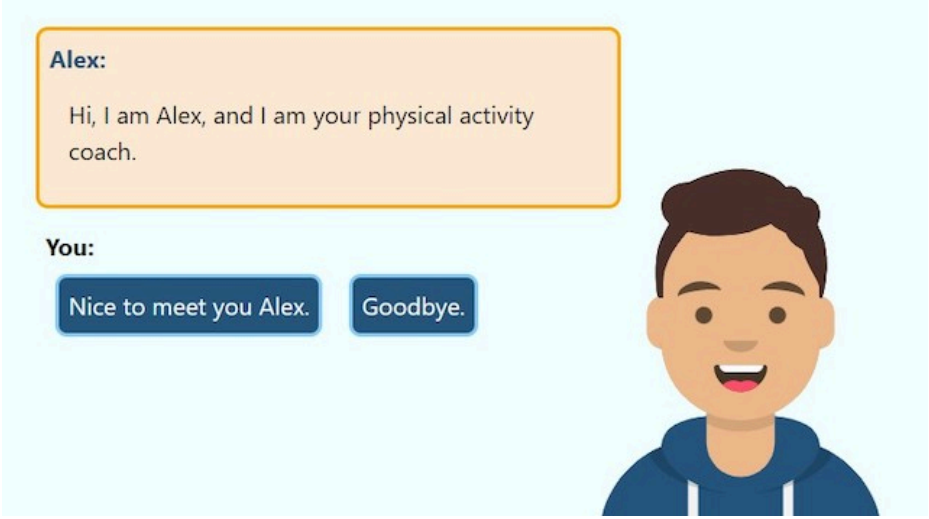

Figure 1: Example of. Alex's Interaction

3.2.1 Coaching strategy- Dialogue creation

We chose to adapt a set of health behaviour dialogues (Beinema et al., 2021) in which persuasion has been the key in building motivational dialogues. This persuasion first measures the motivation level in users and based on that either motivates them more or moves forward with healthy goal setting.

In our implementation of Alex, the persuasion dialogues are further enriched with relational cues that have been identified in the dialogue of skilled human empathic therapists who build a strong working alliance with their patients for improving patients intention to change their behaviour change (Levensky et al., 2007). There are three main motivational strategies that are used. Firstly the dialogue aims to build a working alliance through politeness, affirmation, empathic and encouragement cues. Secondly, the cues aim to empower the user by giving them options or clarifying consequences of each choice. Thirdly, the cues aim to build stronger social ties through adding non-task based social dialogues, particularly during the greeting and farewell phase, or through self-disclosure by the ECA. An

example of Alex's dialogue and the different types of relational cues can be found in Table 1. The example shows both the neutral dialogue version and the empathic version of the dialogue.

Empathic Alex	Neutral Alex
Nice to meet you too!	Let's continue!
I can help you…	I think you should…
I love spending time in nature	Spend time in nature
If you prefer to go outside…	Go outside

Table 1: Sample dialogue of empathic Alex and neutral Alex.

3.2.2 COUCH Features and Architecture

The council of coaches' baseline prototype consists of three components: Unity, ASAP (Articulated social agents platform) Realizer and Dialogue Manager. Unity renders the animation and Dialogue Manager waits for input from the user and also response from the ASAP Realizer. Dialogue Manager has multiple scenarios encoded as BML (Behavior Markup Language). After ASAP Realizer finishes execution of behaviours, the dialogue manager puts responses on top left for the user to choose from. After the user makes a selection, the dialogue manager will send the corresponding BML blocks to the ASAP Realizer to execute the behaviours. The scenarios are encoded in BML which is also stored as assets in Unity. In short, the dialogue manager interacts with the ASAP Realizer and Unity to select the next move in the dialogue.

3.2.3 Message/ Agent Dialogue through WOOL Platform

The WOOL dialogue framework allows domain experts to write and test WOOL dialogue scripts without extensive coding. The five stages of dialogue creation start with domain expert discussion with a technical expert on how to write the dialogues using the framework. In the second stage the domain expert actually writes down the dialogues in the WOOL framework. In the third stage, the technical expert corrects and modifies the content and flow of the dialogues according to the WOOL standards.

The fourth is the final validation of the flow and the content. Finally, the WOOL generated output can be translated into other languages.

3.3 PROCEDURE AND DATA COLLECTION

The procedure for data collection is divided into three modules using the research survey software Qualtrics (qualtrics.com). First is a pre-interaction survey in which participants provide information regarding gender, age and personality types. Participants answer ten questions to measure their personality using Ten-Item Personality Inventory abbreviated as TIPI (Gosling et al., 2003). Then participants answer eight questions about their health behaviours to provide baseline data on quantifying the user's intention to be healthy.

The second module of data collection involves interaction with Alex whereby the user goes through the whole coaching session which is either empathic or neutral, randomly selected by the Qualtrics research survey system. This interaction stores the duration, interaction screen information and the attempts made on accessing the app.

The third module is the post-survey where the user completes the same health behaviour questionnaire that was completed at baseline so as to see how much is the user willing to change the behaviour now. Additional standardised and customised measures which explored users' satisfaction with Alex and their interaction were captured. Each of the measures is described briefly below.

3.3.1 Behaviour change intention analysis

To establish the change in motivation towards health behaviours that comprise of the following 8 behaviours which were asked pre-interaction with Alex and then post-interaction. These are based on a Likert scale of 1-5 for each behaviour, ranging from Never to Always:

1. Keep track of your progress by using automatic step counter
2. Log your food choices in a food diary

3. Define a healthy activity goal for yourself
4. Tweak your daily activities to get closer to the goal
5. Decide and act on drinking additional glasses of water everyday
6. Set an exercise goal
7. Set a healthy eating goal
8. Link your healthy activities to specific moments in time

Van Velsen et al. (2019) talks about persuasive features that correlate with the motivation level of a user and are important in bringing a behaviour change. The above mentioned behaviours are examples of these persuasive features and can be categorized as follows:

1. Self-goal setting feature includes defining a goal, set an exercise goal or a healthy eating goal.
2. Showing progress feature includes the use of step counter, log food choices or tweak a goal.
3. Implementation intention feature include linking health activities to specific moments in time.

In our study, the goal setting included the health behaviours “Define a healthy activity goal for yourself.”,” Set an exercise goal” and “Set a healthy eating goal”, the showing progress included the behaviours “Keep track of your progress by automatic step counter”,” Log your food choices in a food diary.”,” Tweak your daily activities to get closer to the goal.” and “Decide and act on drinking additional glasses of water everyday”. Lastly, implementation intention had only one behaviour “Link your healthy activities to specific moments in time”. These groupings were already found in the original dialogue extracts of health coaching application (Beinema et al., 2021).

3.3.2 Working alliance analysis

Working alliance between Alex and users was measured through the Session Rating Scale Survey (SRSS). The SRSS consists of four 10-point scales. In our implementation in Qualtrics, participants could use a slider

from 0-10, where 0-4 indicated a negative sentiment and above 4 indicated a positive sentiment.. The software would click to the nearest number and record that integer. First, a relationship scale rates the session from "I did not feel heard, understood, and respected" to "I felt heard, understood, and respected." Second is a goals and topics scale that rates the session from "We did not work or talk about what I wanted to work on or talk about" to "We worked on or talked about what I wanted to work on or talk about." Third is an approach or method scale requiring the client to rate the session from "The therapist's approach is not a good fit for me" to "The therapist's approach is a good fit for me?" Finally, the fourth scale looks at how the client perceives the session in total: "There was something missing in the session today" to "Overall, today's session was right for me."

The SRSS is scored by summing the marks for each of the four lines. Based on a total possible score of 40, any score lower than 36 overall, or 9 on any scale, could be a source of concern and therefore prudent to invite the user to comment.

Psychometric testing of the measure has identified a Cronbach's alpha of 0.88 and reliability of 0.64 (Duncan et al., 2003). The measure has also been found to have a moderately strong correlation with the Working Alliance Inventory r=0.63 (Campbell & Hemsley, 2009).

3.3.3 System Usability Scale (SUS)

The SUS (Brooke, 1986) is a 10 item questionnaire with a 5 point Likert scale rating from strongly disagree to strongly agree. Recent confirmatory factor analysis identified that the total sum score of the SUS appears to be a valid and interpretable measure to assess the usability of internet-based interventions (Mol et al., 2020).

3.4 DATA ANALYSIS

The quantitative data generated from the study including the ratings on the health behaviours change intention and likability of the coaching application were analysed using kurtosis and skewness analysis of the health behaviour scores before and after interaction. The data was found

to be normally distributed as thus parametric tests have been chosen. The parametric test chosen to find any significant difference in health behaviours change data before and after using Alex is the two-tailed paired t-test. Comparisons between groups (empathic and neutral) used independent t-test. In our analysis, the p- value is considered significant if $p<0.05$. Bonferroni adjusted p-values have also been calculated and seen from the perspective of avoiding Type-I error but since the analysis is being done to determine future options for research the un-adjusted p values are given more preference. This also helps to reduce the downside of adjusted values (Nakagawa, 2004) and hence avoiding an increase in Type-II error which restricts the option for further exploration. The session rating scales survey (SRSS) and the system usability scale (SUS) along with the likability questionnaire will be examined to understand the user experience and to determine whether participants felt that they had established a therapeutic alliance with the conversational agent. For the users who rated less than 5 for the question "Alex's approach is a good fit for me" regarding the SSRS, feedback was collected which were summarized into themes using thematic analysis (Braun & Clarke, 2006; Joffe & Yardley, 2004) by the first author and reviewed by the second author.

4. RESULTS

The recruited users were first filtered on the condition if they had successfully interacted with Alex. The total users were 217, out of which 206 users were able to interact with Alex and hence were eligible for further analysis. Users are 64% psychology students, 2% a computing major, 18% are studying health sciences, 5% are business, 7% are arts and remaining 5% are enrolled in other subjects. These students were randomly distributed to both groups: 103 of these students were presented with empathic Alex and 103 were presented with neutral Alex.

4.1 INDIVIDUAL FACTOR ANALYSIS

Table 2 shows the descriptive statistics of gender and age with respect to empathic and neutral groups. Overall, 66% of the users are females and the mean age is 21.2.

Group	N	Female	Male	Age	
				mean	std
Empathic	103	73	30	21.5	7.7
Neutral	103	62	41	21.09	6.8
Total	206	135	71	21.2	7.2

Table 2: Gender and Age distribution across all groups

Cultural group based classification for both empathic and neutral Alex is given in the Table 3 which divides the users into nine cultural groups. We observe similar percentages represented in both treatment groups.

Group	Total	Empathic	Neutral
Oceania (Including Australia)	44%	37%	50%
North-Western European	3%	3%	4%
Southern-Eastern European	5%	6%	5%
North African and Middle Eastern	8%	11%	6%
South East Asian	12%	13%	12%
North East Asian	1%	1%	2%
Southern and Central Asian	6%	10%	3%
People of the Americas	2%	1%	3%
Sub-Saharan African	0%	0%	0%
I don't identify with any cultural	2%	3%	2%
I prefer not to answer	1%	2%	1%

Table 3: Cultural group based percentages

Using TIPI score calculations, five personality types were analysed in the context of users of empathic Alex versus neutral Alex. Tests for normality through kurtosis and skewness analysis between the two groups revealed that the data was normally distributed, hence parametric statistical tests were used. Table 4 provides the descriptive statistics and the t-test scores. Three personality types were found to be statistically significant different between the two cohorts (Extroversion, Emotional stability and Openness)

and the remaining two were not statistically significant (Agreeableness and Conscientiousness).

	Empathic		Neutral		
Personality Types	*mean*	*std*	*mean*	*std*	**p-value**
Extroversion	3.27	1.20	2.79	1.28	**0.01**
Agreeableness	2.68	1.25	2.78	1.13	0.56
Conscientiousness	3.77	1.06	3.52	1.09	0.10
Emotional Stability	2.33	1.14	2.78	1.12	**0.01**
Openness	2.36	1.13	2.78	1.12	**0.01**

Table 4: Analyzing variations in personality types through descriptive scores and t-test

4.2 BEHAVIOUR CHANGE INTENTION ANALYSIS

Before selecting a test to measure the change in behaviours the skewness and kurtosis of the user rating distribution was analysed and all of them were in a safe normal range which gave the confidence to go with the parametric test. Most of the distributions are either not skewed or slightly skewed both for neutral and empathic Alex users hence parametric tests are used for further analysis. Table 5 contains the before and after interaction analysis for individual behaviours. P-values shown in bold indicate significant differences. The paired T-test compares means before and after interaction and was performed for the empathic and neutral groups separately and also on the total responses that includes both groups. Table 5 also provides the p-value comparing the **change** (after minus before) in both groups.

Behaviour	Group	Before Interaction		After Interaction		Paired t-test
		mean	*std*	*mean*	*std*	*p*
Show Progress						
Keep track of your progress by automatic step counter	Empathic	2.57	1.41	2.97	1.29	**<.0001***
	Neutral	2.6	1.36	3.04	1.33	**<.0001***
	Total	2.58	1.39	3.00	1.31	**0.001***
Independent t-test p-value		0.88		0.67	change	0.71
Log your food choices in a food diary.	Empathic	2.03	1.09	2.49	1.21	**<.0001***
	Neutral	1.97	1.14	2.33	1.19	**.0007***
	Total	2.00	1.12	2.41	1.19	**.0003***
Independent t-test p-value		0.66		0.32	change	0.49
Tweak your daily activities to get closer to the goal.	Empathic 3.09		1.09	3.39	1.05	**0.003***
	Neutral	3.15	1.09	3.48	1.02	**0.0003***
	Total	3.11	1.09	3.43	1.03	**0.002***
Independent t-test p-value		0.61		0.46	change	0.82
Goal Setting						
Define a healthy activity goal for yourself.	Empathic	3.3	1.09	3.59	0.89	**.003***
	Neutral	3.53	1.03	3.58	1.08	0.64
	Total	3.41	1.06	3.58	0.98	0.103
Independent t-test p-value		0.11		1.0	change	0.11

Set an exercise goal	Empathic	3.44	0.98	3.61	0.96	**.03**
	Neutral	3.47	1.03	3.74	1.00	**.002***
	Total	3.46	1.07	3.67	0.98	**0.03**
Independent t-test p-value		0.84		0.32	change	0.39
Set a healthy eating goal	Empathic	3.31	1.08	3.54	0.96	**0.01**
	Neutral	3.28	1.19	3.63	1.05	**<.0001***
	Total	3.29	1.00	3.58	0.98	**0.003***
Independent t-test p-value		0.83		0.52	change	0.36
Drink an additional glasses of water everyday	Empathic	3.62	1.11	3.91	1.02	**.0001***
	Neutral	3.67	1.10	3.79	1.02	0.202
	Total	3.64	1.11	3.84	1.02	**0.04**
Independent t-test p-value		0.61		0.45	change	0.13
Implementation Intentions						
Link your healthy activities to specific moments in time.	Empathic	3.18	1.07	3.31	0.96	.210
	Neutral	3.18	1.19	3.41	1.05	**.03**
	Total	3.18	1.13	3.36	1.00	0.089
Independent t-test p-value		1		0.44	change	0.48

Table 5: Health behaviour change analysis
**Significant with adjusted p-value=p-value/total behaviours=0.05/8 = 0.006*

Table 6 combines the eight behaviours in Table 5 into three main categories of behaviours according to the three persuasive strategies for pre and post Alex interaction in both empathic and neutral

implementations and then the test is performed on the three categories' derived data.

Behaviour		Before interaction		After Interaction			
		mean	*std*	*mean*	*std*		*p*
Goal setting	Empathic	3.35	0.81	3.57	0.89		**.001***
	Neutral	3.43	0.91	3.63	1.00		**.03**
	Total	3.39	0.9	3.62	0.85		**<.001***
Independent t-test p-value		0.53		0.68		change	**0.82**
Showing progress	Empathic	2.82	0.82	3.18	0.83		**<.001***
	Neutral	2.85	0.81	3.41	1.05		**<.001***
	Total	2.84	0.82	3.18	0.81		**<.001***
Independent t-test p-value		0.79		0.08		change	**0.12**
Implementation intentions	Empathic	3.18	1.07	3.31	0.96		.210
	Neutral	3.18	1.19	3.41	1.05		**.03**
	Total	3.18	1.13	3.36	1.01		**0.01***
Independent t-test p-value		1		0.44		change	0.48

Table 6: Health behaviours strategies change analysis by behaviour category

A total of 35 comments were received from those who scored 4 or below to one of the SRSS question "Why do you feel Alex's approach was/ was not a good fit for you?". Thematic analysis on these led to Table 8.

Theme	Count	Example
Preference for Human Coaches	6	"well. I don't believe he knew enough about me to start and I would rather speak to an actual person this could be my age but interacting with a computer programme is at the forefront of my mind so if I care to lie to it I could. "
Limited answering options	6	"I like the idea of an interactive avatar, however the questions and the limited answering possibilities felt quite contrived. The questions didn't feel like they were linked to the available responses. It felt a bit prescribed rather than personal"
Not personalised	12	"Personal goals are very different to auto-mated generated response from Alex. I understand I need to eat healthy and drink more water, however what can be done for my BMI and other more specific related health related problems."
Features need enrichment	4	"Disadvantage of talking to an online bot with only given questions you can ask, and not being able to back track to ask a different question out of the two offered."
Need for Voice Assistance	1	"I think I understand better when things are verbally communicated to me."
No new information	6	"I believe I already have great knowledge about what Alex was trying to teach."

Table 8: Thematic analysis on empathic Alex's users

Theme	Count	Example
Preference for Human Coaches	6	"well. I don't believe he knew enough about me to start and I would rather speak to an actual person this could be my age but interacting with a computer programme is at the forefront of my mind so if I care to lie to it I could. "
Limited answering options	6	"I like the idea of an interactive avatar, however the questions and the limited answering possibilities felt quite contrived. The questions didn't feel like they were linked to the available responses. It felt a bit prescribed rather than personal"
Not personalised	12	"Personal goals are very different to auto-mated generated response from Alex. I understand I need to eat healthy and drink more water, however what can be done for my BMI and other more specific related health related problems."
Features need enrichment	4	"Disadvantage of talking to an online bot with only given questions you can ask, and not being able to back track to ask a different question out of the two offered."
Need for Voice Assistance	1	"I think I understand better when things are verbally communicated to me."
No new information	6	"I believe I already have great knowledge about what Alex was trying to teach."

Table 9: Responses to questions about Alex

Table 9 reports responses to the three questions concerning the players experience or attitude towards Alex on a Likert scale of 1 to 5 where 5 represents the best experience. The p-values for independent t-tests for all questions are also provided in the tables to determine any significant differences between the empathic and neutral groups.

Questions	Empathic		Neutral		p-value
	mean	*std*	*mean*	*std*	
I liked Alex	3.77	1.01	3.79	1.04	0.89
I found Alex empathic	3.53	1.25	3.38	1.15	0.35
I would recommend Alex to a friend or family member	3.26	1.28	3.38	1.28	0.51

4.3 USABILITY

The SUS statistical scores for empathic and neutral Alex are shown in Table 10. The average SUS for empathic Alex is 2.94 and neutral Alex is 2.98 which means that the application is easy to use and does not require technical support.

Question	Empathic		Neutral		
	Mean	Std	Mean	Std	p-value
I think that I would like to use this system frequently	2.77	1.17	2.77	1.28	1.00
I found the system unnecessarily complex	2.28	1.02	1.93	1.03	**0.008**
I thought the system was easy to use	4.26	0.78	4.46	0.78	**0.02**
I think that I would need the support of a technical person to be able to use this system	1.75	1.15	1.60	1.07	0.32
I found the various functions within this system were well integrated	3.56	1.07	3.60	1.06	0.77
I thought there was too much inconsistency in the system	2.35	1.02	2.28	1.04	0.62
I imagine that most people would learn to use this system very quickly	4.27	0.87	4.37	0.94	0.35
I found the system very cumbersome to use	2.40	1.19	2.50	1.20	0.56
I felt very confident using the system	3.99	0.99	4.36	0.77	**0.001**
I needed to learn a lot of things before I could get going with this system	1.81	1.01	1.96	1.19	0.33
Overall Average Mean	2.94	1.03	2.98	1.04	

Table 10: System Usability Scale Results

5. DISCUSSION

The main objective of our study was to determine the efficacy of an empathic coach to change health behaviours in comparison to a neutral one. The dialogue of the neutral coach was modified to include relational cues (Bickmore et al., 2005) that were extracted from previous work done in health coaching and support strategies for building working alliance, intention to change and likability of the coach (Abdulrahman & Richards,

2019). The study was designed to factor in gender, age and personality traits so as to determine which relational cues are liked more by the individual's attributes (Bickmore et al., 2005). This will establish traits that need to be built into the future versions of the coach along with the customization according to the most liked cues. This work is novel in its design because it implements an application that uses the COUCH architecture and relational cues found in theory but now to be seen in action in the serious gaming world. The previous applications of COUCH (Beinema et al., 2021) did not consider the enrichment of verbal dialogues through the empathic cues.

To measure the impact of the relational cues, our between-subject design exposed participants to a neutral dialogue or an empathic dialogue delivered by Alex. Random allocation resulted in equal numbers in both groups. Analysis of the individual factors of these two groups revealed similar distribution across gender and age. However, t-test comparison of the personality of the two groups revealed that the neutral group were more introverted, emotionally stable and open to new ideas. We might expect that individuals who are open to new ideas might be more willing to explore (Gosling et al., 2003; Heinstrom, 2004) and change their behaviours. Inversely, this potentially means it would be harder to change the behaviours of the empathic group. However, our analysis reports the opposite, adding further weight to our results.

The individual analysis of each health behaviour change led us to interesting insights that can help plan the future extensions of focus with empathic and neutral Alex. We can see from the independent t-tests before the interaction, that there are no significant differences between the neutral and empathic groups. Random allocation by Qualtrics to groups has thus successfully distributed participants so that both groups represent a similar spread of behaviours, perhaps with the exception of the behaviour "Define a healthy activity goal for yourself' where the mean for the empathic group was 3.3 compared to 3.53 for the neutral group. We expected randomisation to ensure no significant differences between the two groups at baseline. For all behaviours, we saw an increase in the mean after the intervention in both groups, indicating that interacting with Alex to chat about their health was beneficial to both groups. To identify

whether neutral or empathic Alex was more effective, we conducted independent t-tests on the change in behaviour intention from before and after the interventions between the two groups. We found some significant differences between the two groups:seven behaviours for empathic Alex and six behaviours for neutral Alex. Thus, it is unclear that either version of Alex delivers better behaviour change intention.

Comparison of the impact of neutral versus empathic Alex is aided when the behaviours are grouped together into three persuasive feature categories: goal setting, showing progress and implementation intentions extracted from the work done by van Velsen et al. (2019), in which goal setting, showing progress and implementation intention were found to be the most important persuasive strategies to motivate.

The goal setting category of behaviours consists of three behaviours relevant to setting or defining a goal in which for “setting an exercise goal”, the rating before interaction was with a mean value of 3.44 for empathic Alex which became 3.61 after interaction. Although the t-tests between empathic and neutral before or after interaction were not significant, but for each group separately the t-tests conducted for before and after interaction were significantly different which means that there was a change in health behaviour intention after the interaction. The same was observed for “setting a healthy eating goal” with mean of rating before interaction being 3.3 and after interaction increased to 3.6. For the statement “Define a healthy activity goal for yourself.” there was a significant change for empathic Alex after interaction with p-value of 0.003 but for neutral the p-value was not significant. Research on motivational strategies that help patients follow advice in the long term (Pereira et al., 2021) also use empathy during goal setting so that patients can understand triggers better and reciprocate to the treatment advised. One reason could be that emphasizing on personal goal setting in the cues as in the cue which uses “for yourself” is more empowering for the users. This refers to de Vries et al. (2017) where the process of change in behaviours is highly affected by personalised texts especially in raising consciousness.

Another behaviour category is “showing progress” in which there are four quantifiable incremental goals that help in building healthier habits in a gradual manner. The first one is “Keep track of your progress by automatic

step counter". Both empathic and neutral Alex before and after interaction had a higher intention to change (p-value for both tests <0.0001). One reason could be that use of digital technology in self-tracking is considered more effective with the user being tagged as a quantified self (a person who feels empowered enough to measure his/her own progress) who shows more intention to change (Didžiokaitė et al., 2018) The second health behaviour is "Log your food choices in a food diary" both interaction groups had significant p-values when comparing before and after interaction (p-values: empathic -E=<0.0001, neutral-N=0.0007), with empathy reporting a higher significance value. Empathy through politeness, empowerment and decision-making coaching strategies (Lelorain, 2021) maintains hope and helps show continuous progress. The third health behaviour in this category is "Tweak your daily activities to get closer to the goal". For this behaviour the neutral users rated more intention to change before and after interaction (before interaction mean: E=3.09, N=3.15 and after interaction mean: E=3.39, N=3.48). One reason that has been quoted by users in improvement is that more clarity and guidance is required for certain coaching cues and this could be one of them For the last behaviour "Drink additional glasses of water everyday" which is also a quantifiable and gradual change, the empathic Alex's change in intention was significant (p-value=0.0001) but not significant for neutral (p-value=0.26). In a nutshell, two behaviours showed a higher intention to change after interaction with empathic Alex. These two behaviours are considered as popular health maintenance goals and can be easily achieved as compared to the other two of which one required an automatic health counter and the other requires more explanation and guidance as to what to change in terms of daily activities. This links to users' feedback in which having more personalised explanation and giving more alternatives can be a viable solution. Some suggestions from the users are "the options given are not for me" or "I need more explanation to my specific health status".

The last behaviour group is implementation intention which has only one health behaviour in it which is "Link your healthy activities to specific moments in time". In terms of significance, neutral Alex users showed intention to change after interaction (p-value=0.03) whereas empathic users did not show much intention to change (p-value=0.21). Both groups had no significant variation in rating when compared before interaction or

after interaction (p-value: before=1.0, after=0.44). This group needs more health behaviours to analyse it further. One behaviour was not found sufficient enough to conclude. This behaviour also has a different interpretation for different types of users and is tightly bound to self-determination and intrinsic motivation (Ryan & Deci, 2000). The more users' psychological needs of being competent, autonomous and related are satisfied, the more the user feels enhanced self-motivation and mental health and hence can make better decisions.

In a nutshell, after interaction with the empathic health coach Alex 7 behaviours showed significant intention to change and for neutral Alex 6 behaviours showed significant intention to change. This provides marginally positive, but inconclusive, support, for RQ1: "Do ECA's empathic dialogues motivate the player to change their health behaviours more than the neutral dialogues?" where the impact of empathic cues in terms of intention to change eight health behaviours needs to be analysed. It can be seen that 87.5% of the health behaviours were rated higher after interaction with empathic Alex. The interaction with neutral Alex also resulted in higher intention to change in 62% of the behaviours. In empathic Alex, the most impactful persuasive strategies were found to be goal setting and showing progress (100% behaviours showed higher change in intention) in which the users were given quantitative goals or monitoring techniques based on their current health status. The neutral Alex had the same motivational strategies but without relational cues and 60% behaviours showed a higher intention to change after interaction. The intention to change was also found higher for neutral Alex users especially in implementation intention strategy which show that relational cues have more impact in goal setting and showing progress than in implementation intention strategies. This also relates to the research work of Beinema et al. (2022), where implementation intentions relevancy is greater for the dialogues associated with outlining the concrete steps to achieve a task than to the non-task based relational cues but this needs to be explored further since there was only one behaviour in implementation intention feature group.

The second research question RQ2: "Do ECA's empathic dialogues build a stronger working alliance with the player than the neutral dialogues?'

analyses the efficacy of the relational cues in building better working alliance which is determined by the working alliance questionnaire (Brooke, 1986). It consists of four questions whose individual contribution to the analysis determines multiple aspects of the impact of relational cues. The first SRSS question is "I felt heard, understood and respected" that caters to dialogues that according to Cameron (2015), builds mutual understanding and shared goal's planning through politeness, inclusivity and affirmation (Cameron,et al., 2015). Empathic Alex has a higher mean rating for this question (6.50) than neutral Alex (6.47) and in both groups 70% users rated Alex above 5.

For the second SRSS question: "We worked on and talked about what I wanted to work on and talk about" is about consideration of keeping the content of coaching to be as helpful as possible for all users. Empathic Alex's average rating was higher (6.19) as compared to neutral Alex (5.98). The content was designed to accommodate the motivational and persuasion features required for all users whereby the original dialogues were extracts of initial implementation of health coaching applications using virtual coaches (Beinema et al., 2021).

The third SRSS question "Alex's approach is a good fit for me" is about how Alex personalises and caters to individual users' more progressive needs in the session. Although 70% users rated high for this question in both groups but mean of neutral Alex was found to be on a higher end (5.97) as compared to empathic Alex (5.62). The feedback received from the users on this question was thematically analysed to identify enhancements and future directions for the coaching approach. The main limitations were found in the overall application's need to be more personalised (12 out of 35 users who rated it less than 5 also mentioned Alex being not personalised). One reason could be that a working alliance usually needs more personalisation that, in future implementations, needs to be structured in multiple sessions with history and response choices of the user creating more personalised and engaging scenarios (Busseri & Tyler, 2003).

The fourth question "Overall, today's session was right for me" which measures the satisfaction level of the user with the session itself was also rated higher by 70% of the users in both groups with neutral being rated

at a higher end (N=5.76, E=5.65). Overall, the working alliance rating was on a higher end for both groups and the first two questions that depict more direct association with relational cues to build working alliance have a higher alliance score for empathic Alex.

Some suggestions that have been given by users to make Alex a better fit include certain feature enhancements like having a navigational menu to re-route to other topics, having more options to embed the personalised context of the user's current health status and make recommendations accordingly, have more check and balance on whether the user wants to know advanced content or is happy knowing the basics and having the voice feature added to Alex.

The third research question (RQ3): "What is the player's experience of the empathic ECA compared to the neutral ECA?" is analysed by the user experience questionnaire that consists of three questions. The first question "I liked Alex" was rated slightly higher by the neutral Alex users although both groups rated it highly in terms of likability. One reason could be that neutral Alex was not impolite or rude ad since users of neutral Alex did not experience empathic Alex, the rating is for their own limited experience with neutral Alex. The second question is "I felt Alex was empathic". This had a higher average rating for empathic users although both groups were not significantly different. The third question was "I would recommend Alex to a friend or family member" was rated slightly higher by the neutral users. The reason could be that both groups were unaware of the dialogues variation in each other. In summary the overall experience of both user groups is rated high irrespective of empathic versus neutral.

Lastly, the SUS score also contributes towards answering RQ3 as it allows us to compare the user experience for empathic Alex and neutral Alex groups. In the SUS analysis, three usability experiences were found significantly different for both groups which are "I found the system unnecessarily complex", "I thought the system was easy to use" and "I felt very confident using the system". The empathic Alex average rating for these three was inclined towards a more negative experience which means that empathic Alex was found to be more complex and raised more concerns for users. One reason could be that empathic Alex's dialogue

paths involved longer conversations and involved Alex referring to his own life and background. This attempt to make the discussion more humanlike, rather than a set of survey questions may have added to the perceived complexity and possible frustration. Users may have just wanted to get to the point and may have found the conversations deviated too much from the goal of the conversation. There were two SUS questions which favoured empathic Alex, although for two groups they were not significantly different. These were "I found the system very cumbersome to use" and "I needed to learn a lot of things before I could get going with this system". Empathic Alex users found the system less cumbersome and also the learning curve was considered shorter by the empathic Alex users. One reason could be that, although the journey was longer for empathic coaching, that extended duration of usage also made users comfortable with the system.

5.1 LIMITATIONS

Although the goal of this study was to analyse the impact of health coaching gaming technology with a flavour of empathy and relational cues utilization in comparison to the neutral one with no relational cues, our results did not confirm better outcomes in terms of behaviour change, working alliance or experience. We note that motivating an individual to change their behaviour is likely to require a long term working alliance (Bickmore et al., 2010). Thus, to see changes may require more sessions with the same users and every session needs to adapt to the user history and personalised goal plan.

Secondly, the gaming applications are built on a plethora of modes of interaction that range from providing multiple response options through to a variety of channels. Currently, in our application the user can only interact using fixed text responses that are provided and the user selects one of them. This can be enriched with more interactive options including user menus and re-routing options to find more personalised journeys within the app.

Thirdly, the coaching application can be made more generalisable by testing on a wider range of individuals including more age variations and

users with more specific health issues so that the value addition of the virtual application can be enriched. Currently it has been tested only on students who are mostly health aware and active.

5.2 FUTURE WORK

This study is a baseline for empathic health coaches' development and can be extended further to bring in multiple coaches to provide coaching expertise that covers more than one domain area (e.g. diet, physiotherapy, diabetes, etc) in the application. This will increase the breadth of learning when the game is used by coaches who are in training and also provides also more tailored support to patients.

In the future, adding personalisation could be achieved through automatic generation of relational cues according to the preferences of the user to contribute towards building a context aware bot. Natural language processing techniques for context categorisation and generation of bags of words that are most suitable to health domains can be explored. The third idea revolves around bringing in more personality aspects into the dialogues involving showing empathy and adapting social dialogues based on individual user factors. For example, there can be a strong liking towards an assertive coach or vice versa. Hence personality based diversity can bring in more likability.

6. CONCLUSION

The main motivation behind this study was to analyse the impact of an empathic virtual coach when it comes to health coaching. The use of the right motivational techniques as well as the choice of words has been shown to be vital in many domains whereby human interaction has been explored, analysed and studied. The human embodiment of a conversational agent is a growing area of research and this study has sought to understand how empathy and relational cues could impact on bringing a change in current behaviour.

REFERENCES

Abdulrahman, A., & Richards, D. (2019). Modelling Working Alliance Using User-aware Explainable Embodied Conversational Agent for Behaviour Change: Framework and Empirical Evaluation. doi:https://aisel.aisnet.org/icis2019/human_computer_interact/human_computer_interact/12/

Beinema, T., Davison, D., Reidsma, D., Banos, O., Bruijnes, M., Donval, B., . . . Huizing, G. (2021). *Agents United: An Open Platform for Multi-Agent Conversational Systems.* Paper presented at the Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents.

Beinema, T., Op den Akker, H., Hermens, H. J., & van Velsen, L. (2022). What to Discuss?—A Blueprint Topic Model for Health Coaching Dialogues With Conversational Agents. *International Journal of Human–Computer Interaction*, 1-19. doi:https://doi.org/10.1080/10447318.2022.2041884

Beinema, T., op den Akker, H., van Velsen, L., & Hermens, H. (2021). Tailoring coaching strategies to users' motivation in a multi-agent health coaching application. *Computers in Human Behavior, 121*, 106787. doi:https://doi.org/10.1016/j.chb.2021.106787

Bickmore, T., Gruber, A., & Picard, R. (2005). Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling, 59*(1), 21-30. doi:https://doi.org/10.1016/j.pec.2004.09.008

Bickmore, T., & Picard, R. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI), 12*(2), 293-327.

Bickmore, T., Schulman, D., & Yin, L. (2010). Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence, 24*(6), 648-666.

Bickmore, T. W., Caruso, L., & Clough-Gorr, K. (2005). *Acceptance and usability of a relational agent interface by urban older adults*. Paper presented at the CHI'05 extended abstracts on Human factors in computing systems.

Boada, I., Rodriguez-Benitez, A., Garcia-Gonzalez, J. M., Olivet, J., Carreras, V., & Sbert, M. (2015). Using a serious game to complement CPR instruction in a nurse faculty. *Computer methods and programs in biomedicine, 122*(2), 282-291. doi:https://doi.org/10.1016/j.cmpb.2015.08.006

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101.

Brooke, J. (1986). System usability scale (SUS): a quick-and-dirty method of system evaluation user information. *Reading, UK: Digital equipment co ltd, 43*, 1-7.

Busseri, M. A., & Tyler, J. D. (2003). Interchangeability of the working alliance inventory and working alliance inventory, short form. *Psychological assessment, 15*(2), 193. doi:https://psycnet.apa.org/doi/10.1037/1040-3590.15.2.193

Cameron, R. A., Mazer, B. L., DeLuca, J. M., Mohile, S. G., & Epstein, R. M. (2015). In search of compassion: a new taxonomy of compassionate physician behaviours. *Journal of Health Expectations, 18*(5), 1672-1685. doi:https://doi.org/10.1111/hex.12160

Campbell, A., & Hemsley, S. (2009). Outcome Rating Scale and Session Rating Scale in psychological practice: Clinical utility of ultra‐brief measures. *Clinical Psychologist, 13*(1), 1-9. doi:https://doi.org/10.1080/13284200802676391

de Vries, R. A., Truong, K. P., Zaga, C., Li, J., & Evers, V. (2017). A word of advice: how to tailor motivational text messages based on behavior change theory to personality and gender. *Personal and Ubiquitous Computing, 21*(4), 675-687. doi:https://link.springer.com/article/10.1007/s00779-017-1025-1#citeas

Didžiokaitė, G., Saukko, P., & Greiffenhagen, C. (2018). The mundane experience of everyday calorie trackers: Beyond the metaphor of Quantified Self. *New Media & Society, 20*(4), 1470-1487. doi:https://doi.org/10.1177/1461444817698478

Drummond, D., Hadchouel, A., & Tesnière, A. (2017). Serious games for

health: three steps forwards. *Advances in Simulation, 2*(1), 1-8. doi:https://advancesinsimulation.biomedcentral.com/articles/10.1186/s41077-017-0036-3#citeas

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of brief Therapy, 3*(1), 3-12.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in personality, 37*(6), 504-528. doi:https://doi.org/10.1016/S0092-6566(03)00046-1

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in personality, 37*(6), 504-528.

Graafland, M., Vollebergh, M. F., Lagarde, S. M., Van Haperen, M., Bemelman, W. A., & Schijven, M. P. (2014). A serious game can be a valid method to train clinical decision-making in surgery. *World journal of surgery, 38*(12), 3056-3062. doi:https://link.springer.com/article/10.1007/s00268-014-2743-4#citeas

Hall, K., Gibbie, T., & Lubman, D. I. (2012). Motivational interviewing techniques: facilitating behaviour change in the general practice setting. *Australian family physician, 41*(9), 660-667. doi:https://search.informit.org/doi/10.3316/informit.737035419857450

HEINSTROM, J. (2004). FIVE PERSONALITY DIMENSIONS AND THEIR INFLUENCE ON INFORMATION BEHAVIOR.

Joffe, H., & Yardley, L. (2004). Content and thematic analysis. *Research methods for clinical and health psychology, 56*, 68.

Lelorain, S. (2021). Discussing Prognosis with Empathy to Cancer Patients. *Current Oncology Reports, 23*(4), 42-42. doi:https://link.springer.com/article/10.1007/s11912-021-01027-9#citeas

Levensky, E. R., Forcehimes, A., O'Donohue, W. T., & Beitz, K. (2007).

Motivational interviewing: an evidence-based approach to counseling helps patients follow treatment recommendations. *AJN The American Journal of Nursing, 107*(10), 50-58. doi:10.1097/01.NAJ.0000292202.06571.24

McCarley, P. (2009). Patient empowerment and motivational interviewing: engaging patients to self-manage their own care. *Nephrology nursing journal, 36*(4), 409.

McGrath, J. L., Taekman, J. M., Dev, P., Danforth, D. R., Mohan, D., Kman, N., . . . Lemheney, A. (2018). Using virtual reality simulation environments to assess competence for emergency medicine learners. *Academic Emergency Medicine, 25*(2), 186-195. doi:https://doi.org/10.1111/acem.13308

Mol, M., van Schaik, A., Dozeman, E., Ruwaard, J., Vis, C., Ebert, D. D., . . . Mora, T. (2020). Dimensionality of the system usability scale among professionals using internet-based interventions for depression: a confirmatory factor analysis. *BMC psychiatry, 20*(1), 1-10. doi:https://doi.org/10.1186/s12888-020-02627-8

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral ecology, 15*(6), 1044-1045. doi:https://doi.org/10.1093/beheco/arh107

Pereira, R. A., Alvarenga, M. S., Avesani, C. M., & Cuppari, L. (2021). Strategies designed to increase the motivation for and adherence to dietary recommendations in patients with chronic kidney disease. *Nephrology Dialysis Transplantation, 36*(12), 2173-2181. doi:https://doi.org/10.1093/ndt/gfaa177

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist, 55*(1), 68. doi:https://doi.org/10.1037/0003-066X.55.1.68

van Velsen, L., Broekhuis, M., Jansen-Kosterink, S., & op den Akker, H. (2019). Tailoring persuasive electronic health strategies for older adults on the basis of personal motivation: Web-based survey study. *Journal of Medical Internet Research, 21*(9), e11759. doi:https://doi.org/10.2196/11759