Designing an Adaptive Embodied Conversational Agent for Health Literacy: a User Study

Joy Egede
joy.egede@nottingham.ac.uk
School of Computer Science
University of Nottingham
Nottingham, UK

Maria J. Galvez Trigo
maria.galveztrigo@nottingham.ac.uk
School of Computer Science
University of Nottingham
Nottingham, UK

Adrian Hazzard
adrian.hazzard@nottingham.ac.uk
School of Computer Science
University of Nottingham
Nottingham, UK

Martin Porcheron
m.a.w.porcheron@swansea.ac.uk
Computational Foundry
Swansea University
Swansea, Wales, UK

Edgar Bodiaj
edgar.bodiaj@nottingham.ac.uk
School of Computer Science
University of Nottingham
Nottingham, UK

Joel E. Fischer
joel.fischer@nottingham.ac.uk
School of Computer Science
University of Nottingham
Nottingham, UK

Chris Greenhalgh
chris.greenhalgh@nottingham.ac.uk
School of Computer Science
University of Nottingham
Nottingham, UK

Michel Valstar
michel.valstar@nottingham.ac.uk
School of Computer Science
University of Nottingham
Nottingham, UK

ABSTRACT

Access to healthcare advice is crucial to promote healthy societies. Many factors shape how access might be constrained, such as economic status, education or, as the COVID-19 pandemic has shown, remote consultations with health practitioners. Our work focuses on providing pre/post-natal advice to maternal women. A salient factor of our work concerns the design and deployment of embodied conversation agents (ECAs) which can sense the (health) literacy of users and adapt to scaffold user engagement in this setting. We present an account of a Wizard of Oz user study of 'ALTCAI', an ECA with three modes of interaction (i.e., adaptive speech and text, adaptive ECA, and non-adaptive ECA). We compare reported engagement with these modes from 44 maternal women who have differing levels of literacy. The study shows that a combination of embodiment and adaptivity scaffolds reported engagement, but matters of health-literacy and language introduce nuanced considerations for the design of ECAs.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Natural language interfaces**; **Empirical studies in HCI**.

IVA '21, September 14–17, 2021, Virtual Event, Japan

ACM ISBN 978-1-4503-8619-7/21/09…$15.00
https://doi.org/10.1145/3472306.3478350

KEYWORDS

Embodied conversational agents, virtual human, design, health and well-being, literacy, user study

ACM Reference Format:
Joy Egede, Maria J. Galvez Trigo, Adrian Hazzard, Martin Porcheron, Edgar Bodiaj, Joel E. Fischer, Chris Greenhalgh, and Michel Valstar. 2021. Designing an Adaptive Embodied Conversational Agent for Health Literacy: a User Study. In *21th ACM International Conference on Intelligent Virtual Agents (IVA '21), September 14–17, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472306.3478350

1 INTRODUCTION

We chart the work to design embodied conversational agents (ECA) to promote better health outcomes for maternal women with low levels of literacy. In particular, we focus on the context of designing for women in sub-Saharan Africa, where level of education is known to impact the outcomes of pregnancies [1, 23]. We collaborated with two Nigerian partner organisations who provide support to pregnant and nursing mothers and a company which supplies goods and training to Patent and Proprietary Medicine Vendors (PPMVs) in Nigeria, who are the first point of call for medicine and healthcare information for low-income communities [4].

Our long term aim is to deliver an ECA as a touch screen kiosk suitable for deployment in PPMVs in Nigeria to enable hard-to-reach communities access to information. The ECA would converse with users, respond to questions, and employ face and voice analysis to monitor user engagement, from which it can respond with clarifying questions or adaptations to the language. In this paper, we report on a Wizard of Oz study using a prototype ECA, 'ALTCAI', implemented to mimic our future interactive 'AI' ECA. The COVID-19 pandemic impeded our plans for deployment in Nigeria, thus we reconfigured the study for online delivery engaging with maternal women based in Nigeria and the United Kingdom.

We implemented three communication 'modes' into our system: A) a conversationally adaptive speech and text-based version (without ECA); B) an adaptive ECA, and C) a non-adaptive ECA. Our aim was to capture measures of user engagement.

First we outline pertinent related work to position our study, after which we detail the design of ALTCAI. The study protocol and participant group is described, after which we present the study findings which are then discussed alongside the study limitations.

2 RELATED WORK

Face-to-face engagements with health professionals are preferred to remote interactions due to non-verbal cues and the communication of empathy. Conversational agents aim to replicate these encounters, however, limitations have been revealed regarding the efficacy of text-based systems [22]. Similarly challenging is the communication of health advice to individuals with low health-literacy, with its application of scientific terminology [12, 25]. Sherwani et al.'s study [32] noted workers in Pakistan with limited reading competency successfully used a voice-based system [32]. Medhi et al. [22] examined a number of interaction modes including text, speech and graphical interfaces, and human operators, observing that graphical and speech-based interfaces supported low-literacy users better than text-only interfaces.

ECAs are increasingly employed in healthcare settings (e.g., [17, 19, 33]) across a number of specialisms, such as healthy eating [9], guiding treatment choices [30], and post-traumatic stress [34]. Moreover, ECAs are shown to be effective at changing patient behaviour (e.g.,[6, 8, 15, 21]). Prior studies consider user perceptions, acceptance, usability and trust. Bickmore et al. [5] tested attitudes and usability observing few differences in relation to health literacy, but noted that users with lower literacy tended to anthropomorphise ECAs more. In [6] ECA nurses advised patients on hospital discharge procedures, revealing that low literacy patients *"preferred receiving the discharge information from the agent over their doctor or nurse"*. Other work has highlighted advantages of ECAs for engagements with low literacy users, such as [7], whose study on consent form explanation revealed that participants preferred the ECA as it provided more time, ability to repeat questions and a perceived lack of bias compared to the human. This latter example speaks to Katz et al.'s [18] study of doctor patient consultations who noted patients with lower literacy asked fewer questions and used fewer medical terms, and also requested information to be repeated more, potentially indicating a lower-level of comprehension that required extra explanation. Our work is focused on the design of healthcare ECAs, with a focus on how we can present pregnancy information in an adaptive way to support users with different levels of literacy.

3 RESEARCH QUESTIONS

Our study focuses on evaluating user engagement with different modes of conversational agent (i.e., adaptive ECA, ECA, and speech and text). Engagement is measured primarily using an adapted version of the short form User Engagement Scale (UES-SF) questionnaire. Our research questions are:

RQ1 Is there a difference in user engagement between (i) adaptive and non-adaptive ECAs?; (ii) adaptive ECA and adaptive speech and text?; and (iii) non-adaptive ECA and adaptive speech and text?

RQ2 Do the levels of health literacy, education or prior attitudes to ECAs affect engagement with our ECAs?

4 DESIGN OF THE ALTCAI ECA

4.1 Overview

ALTCAI was developed in Unity 3D[1] as a multi-platform system for running Wizard of Oz studies [29]. ALTCAI is based on the ARIA agent[2], which uses the Greta ECA, and can deliver computer generated speech through the Text-To-Speech engine CereVoice[3], text and a visual embodiment of a human with synchronised facial and bodily gestures. Speech and text is primarily driven by a text-based script delivered in a sequence of statements triggered by the operator. There is the option to cue predefined statements and custom text. These latter two options are used to mimic real-time conversational responses, i.e., *adaptive* mode. The design also has an *advice mode*, which switches between a low complexity script and an high complexity script, enabling the wizard to respond to participant's perceived understanding.

4.2 ALTCAI System Interfaces

The system comprises a participant's interface which displays the ECA, and the wizard's interface to coordinate system interactions. Communication between interfaces uses queues of messages delivered through an ActiveMQ[4] channel. The participant's interface has two visual modalities: (a) an ECA with text-to-speech (TTS) and text; and (b) a non-embodied speech and text-based version (Figure 1). The interface displays three optional buttons for user navigation. We drew motivation for the visual design of our ECA from Parmar et al.'s [28] who observed that ECAs in professional attire and environments were perceived as *"more credible, trustworthy, likeable, and easy to talk to"*, and users were more likely to act on the ECA's advice (Figure 1).

The wizard's interface consists of a configuration screen to select the country in which it is being used (see ALTCAI Conditions section), the modality (i.e., ECA or speech and text), and the desired script. A second screen controls participant's interface interactions, with controls to manage progression through the scripted content, cueing predefined responses, and entering custom text.

5 STUDY DESIGN AND PROTOCOL

We designed a within-subjects counter-balanced study using the three modes of interaction with the ALTCAI system (i.e., conditions), alongside three content packages of pregnancy information (so that each participant would receive different content in each mode). To address potential order effects, the three interaction modes were rotated (i.e., independent variable), as were the scripted content packages (i.e., extraneous variable), which in turn generated a total of 18 unique mode/package configurations, assigned in turn to our participants. Participants were not informed in advance that this was a Wizard of Oz study.

[1] https://unity.com/
[2] https://aria-agent.eu
[3] https://www.cereproc.com/en/products/sdk
[4] https://activemq.apache.org/components/classic/

Figure 1: ALTCAI's participant's interface. Text based to the left with ECA based overlapping to the right.

5.1 Participants

We recruited 44 participants, 24 Nigerian nationals–recruited by our partner organisation–and 20 United Kingdom (UK) residents, recruited via social media and the host university's student networks. The inclusion of UK participants was a response to COVID-19 forcing us to switch to an online study and limiting our capacity to work in Nigeria, but also allowed us to assess local differences. 6 participants were aged between 18-24 (1 UK & 5 Nigeria), 35 between 15-34 (17 UK & 18 Nigeria), 3 between 35-44 (2 UK & 1 Nigeria). For 29 participants it was their first pregnancy (14 UK & 15 Nigeria). Formal level of education saw 15 report secondary school (2 UK & 13 Nigerian), 5 further/tertiary (1 UK & 4 Nigerian), 15 undergraduate (UK & 7 Nigerian) and 9 post-graduate (9 UK).

The UK based participants took part using their own networked device, in a space of their choosing (e.g., at home). In contrast, Nigerian participants were hosted at our partner organisation with a member of the partner organisation to attend to any technical difficulties and–on occasion–the participant's midwife.

5.2 Protocol

5.2.1 ***NARS Questionnaire.*** We captured participants prior experiences and attitudes to ECAs to help contextualise our analysis. We employed the Negative Attitudes towards Robots Scale (NARS) [26]. We adapted the questionnaire, replacing instances of the word *robot* with *virtual human* and simplified the language in places after the first six participants asked for clarifications.

5.2.2 ***Demographic Questionnaire.*** A demographics questionnaire captured age, education level, whether English was their first or second language, number of pregnancies/births, typical sources for pregnancy information, and prior use and experience of digital devices including ECAs. These demographics were captured to support analysis of participants' perceptions and engagement.

5.2.3 ***REALM Test.*** We administered the Rapid Estimate of Adult Literacy in Medicine (REALM) test [24], a widely used read-aloud test to measure health literacy. Two researchers on the video call made notes of mispronounced terms. Post-study, they compared notes and agreed on a definitive grading.

5.2.4 ***ALTCAI Conditions.*** Three conditions were administered, interleaved with the Post-Condition Questionnaire. Each condition presented a different interaction mode, as detailed below:

Adaptive Speech and Text mode (app A): The interface presents speech and text only. 'Adaptive' mode requires the wizard to mimic an interactive AI by observing the participant for perceived signs of disengagement (e.g., looking away, yawning) or confusion (e.g., frowning, head shake, verbal markers). When observed, they trigger pre-scripted responses, such as *'You seem distracted, do you want me to repeat?'* or *'Does that makes sense?'*. Pre-scripted responses to questions posed from participants, include, *'That is a good question, but I cannot help with it. I suggest you ask your midwife or doctor'*. Custom text responses were used less frequently, only when other pre-scripted responses were deemed inappropriate. Finally, the wizard switched from the 'Low' literacy script to the 'High' mode (i.e., advice mode) after 4-5 sentences if the participant appeared engaged, changing back to 'Low' if deemed necessary.

Adaptive Virtual Human mode (app B): The participant's interface displays our ECA with speech and text (Figure 1). Adaptations are conducted as detailed in app A above.

Non-adaptive Virtual Human mode (app C): The participant's interface displays our ECA as app B above. As a non-adaptive mode, the wizard responds only to 'repeat' and 'next' statements from participants, the advice mode is set to 'Low' throughout.

We had three different packages of advice, namely, 'Staying Healthy'*(SH)*, 'What to Expect' *(WE)* and 'Labour and Birth' *(LB)*. To assure validity, information was drawn from the UK's National Heath Service (NHS) website[5]. This information was used to create two versions of each package, using low complexity language and higher complexity language. We ran each script through a Flesch Kincaid readability test to assure appropriateness for all participants: *SH* advanced 67.4 reading ease; *SH* simple 74 reading ease; *WE* advanced 74.1 reading ease; *WE* simple 81.1 reading ease; *LB* advanced 72.1 reading ease; *LB* simple 75.6 reading ease. Primarily, the high complexity script used more medical terminology. Alternative scripts were then prepared for our Nigerian participants to address any issues of colloquialism and terminology not applicable. The wizard selects which language package to load on the configuration screen (i.e., UK or Nigerian based).

5.2.5 ***Post-Condition Questionnaire.*** Participants completed the short form of the User Engagement Scale (UES-SF) questionnaire following each condition to evaluate user engagement. The UES-SF questionnaire is used in a range of Human Computer Interaction (HCI) settings [27] and covers four categories: focused attention (FA); perceived usability (PU); aesthetic appeal (AE); and reward (RW). Each statement is ranked using a five-point Likert scale.

After the first six participants asked for clarifications, or answered questions inconsistently, we took the decision to adapt the wording of the UES-SF questionnaire to simplify their language. Following O'Brien et al.'s guidance for adapting the UES-SF questionnaire [27], we carefully reworded the nine items that were relevant to our study focusing on not changing their meaning. The resulting statements can be seen in Table 1.

[5] https://www.nhs.uk/pregnancy/

Table 1: Adaptation of the UES-SF for our study

Category	Statement
FA	When I was using the app I felt that time passed quickly.
FA	I wanted to give the app all my attention.
PU	I felt frustrated while using the app.
PU	I was confused when using the app.
PU	I had to work or think hard to use the app.
AE	The app was nice to look at.
AE	I liked how the app sounded.
RW	Apps like this will help me in the future.
RW	I thought the app was interesting.

5.2.6 Practicalities of Study Delivery. We administered the study using Microsoft Teams[6] and Zoom[7]. Teams was used for the UK based participants, whereas Zoom was employed for the majority of the Nigerian participants due to local familiarity with the system. All study materials were shared from researchers' computers to the participant's device via screen sharing. Researcher 1 administered the participant paperwork and all questionnaires–entering the participant's responses into the forms–and Researcher 2 acted as the wizard, sharing ALTCAI's participant's interface with the participant. Researcher 1 also recorded the video conference call.

6 FINDINGS

6.1 Data Analysis

We analysed the the adapted UES-SF questionnaire results to understand if the different conditions affected participant engagement. We also explored if participants' health literacy (REALM), attitudes towards ECAs (NARS), or demographic data had an effect on engagement. We disregarded the first six participants that completed the original UES-SF and NARS questionnaires, as these measures were subsequently altered. Thus, our analysis focused on the remaining 38 participants who completed the adapted questionnaires. Facial behaviour analysis of our participant video was conducted using the OpenFace 2.0 Toolkit[8]. This focuses on changes in facial expressions and head movements to identify instances and intensities of upper and lower facial action units [2]. Two of the 44 recruited participants video data was corrupted, thus the facial behaviour data set is based on the remaining 42 participants.

6.2 Influence of conditions on user engagement

To understand if the mode used influenced the results of the adapted UES-SF, we used a within-subjects one factor ANOVA with the mode type (adaptive speech and text, adaptive ECA, and non-adaptive ECA) as a three-level factor. We did this for the total questionnaire score as well as for the score of each of its categories. We found that the mode used had a statistically significant effect on the total score, $F(1.776, 67.740) = 8.940, p < 0.001$. Post-hoc pairwise comparisons (see Figure 2) adjusted using the Bonferroni method revealed that, for the total score, adaptive ECA (app B) had a significantly higher score than the adaptive speech and text (app A) (4.13 ± 0.46 vs 3.92 ± 0.51, respectively), and that this difference was also significant between the adaptive ECA (app B) and the non-adaptive ECA (app C) (4.13 ± 0.46 vs 4.02 ± 0.48, respectively). However, this difference was not statistically significant between the adaptive speech and text (app A) and the non-adaptive ECA (app C). The tests also showed a statistically significant difference between apps A and B in the categories of focused attention (3.63 ± 0.70 vs 3.85 ± 0.68, respectively), aesthetic appeal (3.71 ± 0.83 vs 3.95 ± 0.71, respectively) and reward (4.14 ± 0.67 vs 4.35 ± 0.61, respectively), with this difference being also significant between apps B and C in the reward category (4.35 ± 0.61 vs 4.14 ± 0.66, respectively).

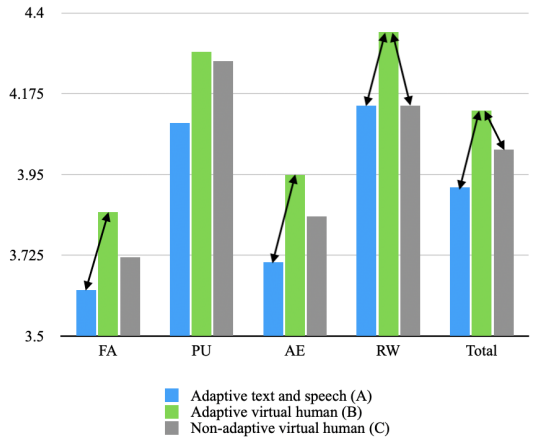

Figure 2: Scores on the adapted UES-SF questionnaire. Arrows indicate statistically significant differences.

6.3 Influence of NARS responses on user engagement

To investigate the relation between the negative attitudes towards ECAs and the user engagement as measured by the adapted UES-SF questionnaire, we carried out a one-tailed Pearson's correlation test for the total results on the NARS questionnaire and for the total results on each subscale (S1 - Negative Attitudes towards Situations and Interactions with ECAs, S2 - Negative Attitudes towards Social Influence of ECAs, and S3 - Negative Attitudes towards Emotions in Interaction with ECAs), along with the total results on he UES-SF questionnaire for each of the apps.

As can be seen in the last row of Table 2, we didn't find a statistically significant correlation between the total NARS scores and the engagement scores for any of the modes. However, we found a statistically significant negative correlation of moderate strength between the results of the subscale 1 of the NARS and the engagement scores for apps B and C, meaning that those participants that demonstrated a less negative attitude towards situations and interactions with virtual humans obtained a higher engagement score for the two modes that feature a virtual human.

[6] https://www.microsoft.com/en-gb/microsoft-teams/group-chat-software
[7] https://zoom.us/
[8] https://github.com/TadasBaltrusaitis/OpenFace

Table 2: Pearson's correlation coefficients (r) between NARS results and UES-SF scores for each app. Significant correlations are marked with *.

NARS/UES-SF		App A	App B	App C
S1	r	-0.265	-0.321*	-0.304*
	p	0.054	0.025	0.032
	N	38	38	38
S2	r	0.027	-0.078	-0.069
	p	0.437	0.32	0.339
	N	38	38	38
S3	r	-0.1	-0.115	-0.146
	p	0.276	0.246	0.191
	N	38	38	38
Total	r	-0.137	-0.217	-0.214
	p	0.206	0.095	0.098
	N	38	38	38

6.4 Influence of health literacy and native language on user engagement

To understand if the health literacy of the participants influenced their engagement scores for each mode, a Pearson's correlation test between the REALM results and the engagement scores was conducted for each mode. We found a statistically significant negative correlation of moderate strength between the results on the REALM test and the engagement scores for app A, $r(38) = -0.374, p = 0.021$; and app C, $r(38) = -0.353, p = 0.03$, but not for app B. This suggests that those participants with a higher health literacy score, had a tendency to a lower engagement score for the adaptive speech and text and the non-adaptive ECA modes than those with a lower health literacy score.

Participants were allocated to a group depending on their REALM scores: those classed as *"able to read most patient education materials"* (scores above 60); and those classed as *"with trouble reading most patient education materials"* (scores of 60 or below). A point-biserial correlation showed a statistically significant negative correlation between those who scored above 60 and the engagement scores for the aesthetic appeal of app A, $r_{pb} = -0.524, p < 0.001$. This suggests that those with lower health literacy score (60 or below) had a tendency to rate the aesthetic appeal of the adaptive speech and text mode higher than participants with higher health literacy. We also observed a significant positive correlation between participants with a literacy score above 60 and the perceived usability of app B, $r_{pb} = 0.380, p = 0.019$, meaning that those classed as *"with trouble reading most patient education materials"* had a tendency to rate the perceived usability of the adaptive ECA mode lower than those with a higher health literacy score.

Regarding the effect of native language on participant engagement scores, after running a point-biserial correlation, we did not find a correlation between the participants with English as a second language and the engagement results for apps B and C. However, we found a moderate positive correlation, $r_{pb} = 0.372, p = 0.02$ with the engagement scores for app A, indicating that those participants whose native language was not English had a tendency to a higher engagement score for the adaptive speech and text mode than those whose native language was English.

6.5 Influence of previous device use and ownership on attitudes

Participants were asked about their use and ownership of smartphone assistants, smartspeakers or ECA's to understand how prior attitudes and experiences affected engagement scores. For this purpose we run point-biserial correlations with the relevant variables.

Regarding their NARS scores, we only observed a statistically significant correlation between a prior use of smartspeakers and the results of the NARS subscale 3, $r_{pb} = 0.330, p = 0.03$. This suggests that there was a tendency for participants who had previously used a smartspeaker to present a more negative attitude towards emotions in the interaction with virtual humans than those participants that had not previously used a smartspeaker.

Regarding the engagement scores, we observed a statistically significant negative correlation between the engagement scores obtained for app A and the ownership of smartphone assistants, $r_{pb} = -0.334, p = 0.02$, and the ownership of smartspeakers, $r_{pb} = -0.325, p = 0.044$. Meaning that owners of these devices showed a tendency for lower engagement scores for the adaptive speech and text mode than non-owners of these devices. Moreover, there was a negative correlation between the engagement scores across all three modes and the prior use of smart speakers: app A, $r_{pb} = -0.499, p = 0.001$; app B, $r_{pb} = -0.339, p = 0.035$; app C, $r_{pb} = -0.432, p = 0.006$. This indicates a tendency for those with prior experience of using smartspeakers to obtain lower engagement scores for all modes than for those that had never used them.

6.6 Differences between the UK and Nigeria

To better understand if the negative attitudes towards ECAs or the engagement scores were affected by country of residence, we also ran a point-biserial correlation taking into account the country and the NARS scores, and the country and the adapted UES-SF scores. We did not find any statistically significant correlation between being from Nigeria and the responses to the NARS questionnaire or any of its subscales. However, we found a statistically significant positive correlation between being from Nigeria and some of the engagement scores obtained, with participants from Nigeria having a tendency to higher total engagement for apps A and C than UK participants, including the focused attention (FA) and the reward (RW) subscales; and higher scores for the aesthetic appeal (AE) subscale for all three apps, as can be seen in Table 3.

6.7 Video Analysis

Videos of participants' interaction were analysed to check for differences in their behaviour across app modes. The analysis focused on changes in facial expressions and head movements. OpenFace 2.0 [3] was used to extract facial actions, head movements and eye gaze information. Data extracted included (i) 18 upper and lower facial action unit (FAUs) intensities and occurrences (ii) head pose as pitch, roll and yaw displacements, and (iii) vertical and horizontal eye movements. FAU intensities are scored on a 6-point scale from zero (not active) to 5 (highest activation), while the occurrences are scored as either absent (0) or present (1). OpenFace extracts framewise measurements of each behavioural data. Since the videos are of variable length, direct comparison on the extracted measurements was not possible, hence various statistical metrics including

Table 3: Point-biserial correlation coefficients between the country of the participant and UES-SF (correlations with p<0.05 marked with *, p<0.01 marked with **). A positive correlation indicates a higher UES-SF score by participants in Nigeria.

	App A total	App B total	App C total	App A FA	App B FA	App C FA	App A PU	App B PU	App C PU	App A AE	App B AE	App C AE	App A RW	App B RW	App C RW
r_{pb}	0.553**	0.195	0.392*	0.417**	0.172	0.362*	0.193	−0.203	−0.128	0.710**	0.428**	0.507**	0.369*	0.236	0.415**
p	0	0.233	0.014	0.008	0.295	0.023	0.24	0.216	0.437	0	0.007	0.001	0.021	0.148	0.009

Table 4: Within-Subject ANOVA for differences in observed participant behaviour across modes (sig. diff. marked with *)

Group	Behaviour observed	Descriptor	App A	App B	App C	p (α = .05)
Nigeria (*N=24*)	*AU15*	P_{active}	31.03	33.39*	25.42*	.013
	AU17	P_{active}	30.79	35.57*	27.21*	0.039
UK (*N=18*)	*AU6*	*Mean*	0.35	0.41*	0.30*	0.031

measures of central tendency and dispersion were used to generate video descriptors. An additional custom descriptor was computed for the facial data as the percentage of frames in which a given facial action unit was active in a video as follows: $P_{active} = \frac{n}{N}$. Where n is the number of active frames and N is the total number of frames. Other custom facial data descriptors such as the length of contiguous frames with AU activations were explored, however, some network dropouts made this metric unreliable.

A within-subject design analysis of variance (ANOVA) test was applied to check for differences in the observed behavioural expressions in each mode. The sphericity condition was first checked with the Mauchly's test; where this condition was not met, the Greenhouse-Geisser univariate test was applied. Where a statistically significant difference was found between behavioural expressions in the app modes, a post-hoc pairwise comparison determined where this difference occurred.

Table 4 shows the result of these ANOVA tests. For the Nigeria group, statistically significant differences were found in activations of FAU15 (lip corner depressor) and FAU17 (Chin raiser) across the modes. FAU15 is an activation of the *Depressor anguli oris (Triangularis)* muscle typically associated with the inverted-U mouth shape found in sad or disgust expressions [3]. FAU17 is an activation of the *Mentalis* muscle which can occur in sad or disgust expression but is not exclusive to these.

Table 4 illustrates that the mean FAU15 activations across the modes differed significantly (F (2,46) = 4.77, p = .013). A posthoc pairwise comparison using the Bonferroni correction showed increased FAU15 activations in app B (33.39) compared to app C (27.21) which was statistically significant (p=.038). However, difference observed between app A and C (p=.674), and that between app A and B was not statistically significant (p=.479).

Similarly, for FAU17, statistically significant difference in activations across modes were found (F (2,36.08)=3.896). Further posthoc pairwise comparison revealed higher activations for app B (35.57) compared to app C (27.21) significant at p=0.042, while the difference observed between App A and B, and App A and C did not reach significance (p=.311 and p= .164 respectively).

In the UK group, the only significant difference found was for FAU6 (cheek raiser), an activation of the *Orbicularis oculi, pars orbitalis*, typically associated with a happy expression. ANOVA results showed that the mean FAU6 activations differed significantly across the three app modes (F(2,28) = 3.952, p = .031). The posthoc pairwise comparison showed increased FAU6 activations in app B (0.34) compared to app C (0.30) which was statistically significant (p=.018), whereas the differences observed between the other app pairs were not statistically significant (p (A&B)= .479; p(A&C)=.674).

7 DISCUSSION

7.1 Social Embodiment

Our data analysis highlights an overall significantly higher engagement for the adaptive ECA mode over the two other modes. This broadly relates to prior work that finds 'relational' ECAs demonstrate a better user experience [20], with higher likeability and trustworthiness [9, 11, 20]. If we look into each engagement category separately, we find that this difference is only significant as compared with the other two modes for the *reward* subscale. For the other subscales, the engagement was significantly higher only for the adaptive ECA mode compared with the adaptive speech and text, but not compared with the non-adaptive ECA. The adaptive ECA mode has no single feature to which this finding might be attributed. One interpretation is that the *combination* of embodiment (ECA) and conversational adaptivity is important for shaping participants' perceived sense of 'reward', whereas either of these characteristics in isolation do not. Looking at the other UES-SF subscales, we can interpret that this same combination creates a higher anthropomorphism which contributes positively to *focused attention* and *aesthetic appeal*, in contrast to the similar but disembodied conversational behaviours of the speech and text mode. We conclude that human social characteristics and behaviours were considered more engaging when visually and audibly connected.

7.2 Usability

All system modes were perceived as usable, which is unsurprising given the simplicity of our system (i.e., 'repeat' and 'next' commands) and pre-study instructions. The adaptive modes did not appear to unduly confuse or disrupt usability, i.e., those unexpected

questions or utterances from the agent that in turn solicit participant responses beyond the regular uttering of commands. While these instances of solicitation require participants to comprehend and contextualise the agent's utterance and formulate a response, it appears our interaction design was largely intuitive and in the case of the adaptive ECA mode, "expectations" were met [31]. However, those with English as a second language had a tendency to score the speech and text mode more positively than native English speakers. And participants with a health literacy classed as *"[...] trouble reading most patient education materials"* [24] or below found the speech and text mode more aesthetically pleasing and the adaptive ECA harder to use than those with a health literacy classed as *"able to read most patient education materials"* [24]. This may indicate that the advantages of the adaptive ECA are partially offset for users who may be working harder to understanding the content either because of a low health literacy or the content being delivered in a second language, perhaps by its greater visual complexity. This contrasts with Bickmore et al. observations of a tendency for low literacy users to prefer ECAs more than those with higher levels of health literacy [5]. Further work is needed to explore the degree of extraneous cognitive load the ECA modes may demand (e.g., [10]) and how this may affect user engagement.

This interpretation may be further illuminated by the findings from the video analysis. However, we note that further work is required to contextualise these measures against specific events. Our observations of participants–while not validated or presented here in detail–hint at some possible interpretations. For example, when asked questions by the adaptive ECA, some Nigerian participants in the health literacy category of *"[...] trouble reading most patient education materials"* ostensibly displayed expressions interpreted as 'uncertainty' or 'thinking', which were followed by a re-reading of the displayed text, suggesting processing effort. These instances may explain those moments of 'negative' facial expression activity observed in the adaptive modes, as an indication of increased mental activity rather than dislike or disengagement from the adaptive ECA. This also points to the potential for participants to demonstrate a form of negative engagement that arises from extraneous cognitive load. Further work is needed to understand the additional challenges experienced by ECA users of a second language or with low (health) literacy and how these can be mitigated.

7.3 User Attitudes and Prior Experience

The NARS results suggest that those with a more positive attitude towards virtual humans showed a higher overall engagement whilst using the two ECA modes than those with a more negative attitude. While this bodes well for continued acceptance of future ECAs, it also foregrounds a challenge of how best to on-board late adopters or reluctant users, which speaks to some elements of the explainable AI challenge [16]. We also found that users of smartspeakers presented a more negative attitude towards emotions in the interaction with ECA's than those who had not previously used a smartspeaker, and, although those that had used a smartspeaker showed lower engagement than those that had never used them for all modes, owners of these devices only had a tendency to lower engagement for the adaptive speech and text mode. These differences could be due to the level of experience using these devices (i.e., owners would presumably have a higher experience of interacting with them). However, future work is needed to understand how this affects the engagement and perceptions using ECAs.

7.4 Limitations

ALTCAI was initially designed to be deployed in PPMV shops in Nigeria. Our online study resulted in the UK-based participants using their own devices devices in uncontrolled spaces of their choosing. Conversely, the Nigeria-based participants were in the offices of our partner, which mirrors some characteristics of our intended deployment setting. However, our Nigerian partner is a NGO that provides professional support to the community and typically there were practitioners observing participants during the studies. Consequently, there may have been unintended influence on participants, biasing their study engagement (i.e., this may contribute to the tendency for Nigerian participants to score engagement higher across all app modes). We experienced some network issues and artefacts with Nigerian participants such as 'freezing' and latency, which may have impacted participant engagement.

We adapted two of our questionnaire-based measures (i.e., NARS and UES-SF) to create simple language versions accessible to our participants. This is a potential tension of working with standardised questionnaires and users with low literacy, as acknowledged by [7] and explored by (e.g., [13]). The validation of these measures represents a target for future work. Similarly, we highlight Dowse et al.'s [14] study that questioned the applicability of administering the REALM test with second language or low literacy users. Whilst REALM is intended as a quickfire measure, we nonetheless need to consider its appropriateness going forward.

Finally, we acknowledge that our study focused on user engagement, which does not address matters of user comprehension, a crucial factor to achieve an effective system design. This is to be addressed in the next iteration of our work.

8 CONCLUSION

We conducted a Wizard of Oz user study with a mixed group of UK and Nigerian-based participants who interacted with our ALTCAI ECA, designed to deliver maternal health advice. We explored methods to make such advice adaptive to the literacy skills of users. Participants engaged with three interaction modes: an adaptive speech and text mode; an adaptive ECA mode; and a non-adaptive ECA mode. Our findings reveal the adaptive ECA mode scored significantly higher on engagement across the participant group and suggest that the combination of embodiment and adaptivity is important in creating a system considered to be rewarding, attentive and aesthetically pleasing. However, health-literacy, English as a second language, and prior attitudes to ECAs also affected engagement. Further work is needed to understand how any additional demands due to adaptive ECAs can be mitigated for users with lower health literacy or English as a second language.

ACKNOWLEDGMENTS

This work was supported by the UK Department for International Development and the Engineering and Physical Sciences Research Council [grant number EP/V00784X/1]. To access the research data supporting this publication, see https://doi.org/10.17639/nott.7137.

REFERENCES

[1] Romi O Abidoye and C I Tomin-West. 1999. The relationship of poverty on malnourished children in Lagos, Nigeria. *Nutrition Research* 19, 10 (oct 1999), 1485–1495. https://doi.org/10.1016/s0271-5317(99)00105-0

[2] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, New York, NY, USA, 1–6.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, New York, NY, USA, 59–66.

[4] Naomi Beyeler, Jenny Liu, and Maia Sieverding. 2015. A systematic review of the role of proprietary and patent medicine vendors in healthcare provision in Nigeria. *PLoS One* 10, 1 (2015), e0117165.

[5] Timothy W Bickmore, Laura M Pfeifer, Donna Byron, Shaula Forsythe, Lori E Henault, Brian W Jack, Rebecca Silliman, and Michael K Paasche-Orlow. 2010. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. *Journal of health communication* 15, S2 (2010), 197–210.

[6] Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the Time to Care: Empowering Low Health Literacy Hospital Patients with Virtual Nurse Agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, New York, NY, USA, 1265–1274. https://doi.org/10.1145/1518701.1518891

[7] Timothy W Bickmore, Laura M Pfeifer, and Michael K Paasche-Orlow. 2009. Using computer agents to explain medical documents to patients with low health literacy. *Patient education and counseling* 75, 3 (2009), 315–320.

[8] Timothy W. Bickmore, Daniel Schulman, and Candace Sidner. 2013. Automated interventions for multiple health behaviors using conversational agents. *Patient Education and Counseling* 92, 2 (2013), 142–148. https://doi.org/10.1016/j.pec.2013.05.011

[9] Chris Creed, Russell Beale, and Benjamin Cowan. 2015. The impact of an embodied agent's emotional expressions over multiple interactions. *Interacting with Computers* 27, 2 (2015), 172–188.

[10] Ton De Jong. 2010. Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional science* 38, 2 (2010), 105–134.

[11] Ewart J. de Visser, Samuel S. Monfort, Kimberly Goodyear, Li Lu, Martin O'Hara, Mary R. Lee, Raja Parasuraman, and Frank Krueger. 2017. A Little Anthropomorphism Goes a Long Way: Effects of Oxytocin on Trust, Compliance, and Team Performance With Automated Agents. *Human Factors* 59, 1 (2017), 116–133. https://doi.org/10.1177/0018720816687205 arXiv:https://doi.org/10.1177/0018720816687205 PMID: 28146673.

[12] Nicola Diviani, Bas van den Putte, Stefano Giani, and Julia CM van Weert. 2015. Low Health Literacy and Evaluation of Online Health Information: A Systematic Review of the Literature. *J Med Internet Res* 17, 5 (07 May 2015), e112. https://doi.org/10.2196/jmir.4018

[13] Elizabeth Dormandy, Elaine Y.L. Tsui, and Theresa M. Marteau. 2007. Development of a measure of informed choice suitable for use in low literacy populations. *Patient Education and Counseling* 66, 3 (2007), 278–295. https://doi.org/10.1016/j.pec.2007.01.001

[14] Ros Dowse, Lebo Lecoko, and Martina S Ehlers. 2010. Applicability of the REALM health literacy test to an English second-language South African population. *Pharmacy world & science* 32, 4 (2010), 464–471.

[15] Roger A Edwards, Timothy Bickmore, Lucia Jenkins, Mary Foley, and Justin Manjourides. 2013. Use of an interactive computer agent to support breastfeeding. *Maternal and child health journal* 17, 10 (2013), 1961–1968.

[16] Hani Hagras. 2018. Toward Human-Understandable, Explainable AI. *Computer* 51, 9 (2018), 28–36. https://doi.org/10.1109/MC.2018.3620965

[17] Shashank Jaiswal, Michel Valstar, Keerthy Kusumam, and Chris Greenhalgh. 2019. Virtual Human Questionnaire for Analysis of Depression, Anxiety and Personality. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. ACM, New York, NY, USA, 81–87. https://doi.org/10.1145/3308532.3329469

[18] Marra G Katz, Terry A Jacobson, Emir Veledar, and Sunil Kripalani. 2007. Patient literacy and question-asking behavior during the medical encounter: a mixed-methods analysis. *Journal of general internal medicine* 22, 6 (2007), 782–786.

[19] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[20] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)* 4, 4 (2013), 1–28.

[21] Martin H. Luerssen and Tim Hawke. 2018. Virtual Agents as a Service: Applications in Healthcare. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. ACM, New York, NY, USA, 107–112. https://doi.org/10.1145/3267851.3267858

[22] Indrani Medhi, Somani Patnaik, Emma Brunskill, S.N. Nagasena Gautama, William Thies, and Kentaro Toyama. 2011. Designing Mobile Interfaces for Novice and Low-Literacy Users. *ACM Trans. Comput.-Hum. Interact.* 18, 1, Article 2 (May 2011), 28 pages. https://doi.org/10.1145/1959022.1959024

[23] J K Mojoyinola. 2011. Influence of maternal health literacy on healthy pregnancy and pregnancy outcomes of women attending public hospitals in Ibadan, Oyo State, Nigeria. *African Research Review* 5, 3 (jun 2011), 12 pages. https://doi.org/10.4314/afrrev.v5i3.67336

[24] Peggy W Murphy, Terry C Davis, Sandra W Long, Robert H Jackson, and Barbara C Decker. 1993. Rapid estimate of adult literacy in medicine (REALM): a quick reading test for patients. *Journal of reading* 37, 2 (1993), 124–130.

[25] Lynn Nielsen-Bohlman, Allison M. Panzer, and David A. Kindig (Eds.). 2004. *Health Literacy: A Prescription to End Confusion*. The National Academies Press, Washington, DC. https://doi.org/10.17226/10883

[26] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kennsuke Kato. 2004. Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*. IEEE, New York, NY, USA, 35–40.

[27] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[28] Dhaval Parmar, Stefan Olafsson, Dina Utami, and Timothy Bickmore. 2018. Looking the Part: The Effect of Attire and Setting on Perceptions of a Virtual Health Counselor. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. ACM, New York, NY, USA, 301–306. https://doi.org/10.1145/3267851.3267915

[29] Martin Porcheron, Joel E Fischer, and Stuart Reeves. 2021. Pulling Back the Curtain on the Wizards of Oz. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 243 (jan 2021), 22 pages. https://doi.org/10.1145/3432942

[30] Scott Robertson, Rob Solomon, Mark Riedl, Theresa Wicklin Gillespie, Toni Chociemski, Viraj Master, and Arun Mohan. 2015. The visual design and implementation of an embodied conversational agent in a shared decision-making context (eCoach). In *International Conference on Learning and Collaboration Technologies*. Springer, Cham, Switzerland, 427–437.

[31] Zsófia Ruttkay and Catherine Pelachaud. 2006. *From brows to trust: Evaluating embodied conversational agents*. Vol. 7. Kluwer Academic Publishers, New York, NY, USA.

[32] Jahanzeb Sherwani, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld. 2007. HealthLine: Speech-based access to health information by low-literate users. In *Proceedings of the 2007 International Conference on Information and Communication Technologies and Development* (Bangalore, India) *(ICTD '07)*. IEEE, New York, NY, USA, 1–9. https://doi.org/10.1109/ICTD.2007.4937399

[33] Silke ter Stal, Lean Leonie Kramer, Monique Tabak, Harm op den Akker, and Hermie Hermens. 2020. Design Features of Embodied Conversational Agents in eHealth: a Literature Review. *International Journal of Human-Computer Studies* 138 (2020), 102409. https://doi.org/10.1016/j.ijhcs.2020.102409

[34] Myrthe L Tielman, Mark A Neerincx, Marieke Van Meggelen, Ingmar Franken, and Willem-Paul Brinkman. 2017. How should a virtual agent present psychoeducation? Influence of verbal and textual presentation on adherence. *Technology and Health Care* 25, 6 (2017), 1081–1096.