GRANT AGREEMENT N. 871245

Deliverable D1.4

User feedback from the preliminary validation (realistic environments)

Due Date: 31/01/2022
Main Author: ERM
Contributors: INRIA, HWU, APHP
Dissemination: Public Deliverable

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

DOCUMENT FACTSHEET

Deliverable no.	D1.4: User Feedback from the Preliminary Validation (realistic environments)
Responsible Partner	ERM
Work Package	WP1: Experimental Validation
Task	T1.3: Preliminary Experimental Validation
Version & Date	VDraft, 27/01/2022
Dissemination level	[X] PU (public) [ ] CO (confidential)

CONTRIBUTORS AND HISTORY

Version	Editor	Date	Change Log
1	ERM	20/01/2022	First Draft
2	INRIA, HWU	21/01/2022	Second Draft
Final	ERM, INRIA	31/01/2022	Final Draft including all partners comments

APPROVALS

Authors/editors	ERM, INRIA, HWU
Task Leader	ERM
WP Leader	ERM

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

EXECUTIVE SUMMARY

Delivery 1.4 presents the results of a series of experiments conducted by INRIA and HWU in regard to testing out their respective modules, ARI's ability to navigate in order to join people, and the Task Planner and Conversational Manager.

The experiments were held in each partner's laboratory where volunteers were recruited to participate in the experiments.

The document contains:
A. Results of experiments conducted by INRIA to validate ARI's navigation capabilities in order to start a conversation.
B. Results of experiments conducted by HWU to validate the Task Planner and Conversational Manager while making use of the first version of the robot application.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

CONTENTS OF DELIVERABLE

This deliverable aims to group the first results of the preliminary validation of ARI's navigation capabilities, and the Task Planner and Conversational Manager paired up with the first version of the robot application.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

INRIA EXPERIMENTS

INTRODUCTION

The goal of these experiments is to evaluate the full initial architecture working on the ARI robot, and in realistic environments (i.e. in laboratory conditions). This section presents the experiments conducted at INRIA, and some of the conclusions extracted.

An important note is that, since at the time of the experiments ARI's loudspeaker was not functioning correctly, we could not evaluate the dialogue part (ARI could not speak). Luckily this was extensively evaluated in the experiments conducted at HWU.

EXPERIMENT DESCRIPTION

The experiment was conducted in the laboratory at INRIA, and healthy adult volunteers were recruited. The goal is to evaluate how well ARI can navigate to join people to start a conversation. For this purpose, we designed 4 scenarios involving two participants. In each scenario one of the participants will be Person A and the other Person B. Both will be asked to stand in certain areas (area A and B) of the experimental room and wait for ARI to join them. On the path from ARI's start location to area A, there was no obstruction so that ARI could take a direct path. Whereas, on the path from ARI's start location to area B ARI had to navigate around a small obstacle (low table). After the experiment, the participants were asked to fill out a questionnaire about how they felt about ARI's behaviour. The scenarios are:

1. Person A is standing in area A, and Person B in area B. Both facing ARI. ARI will then join Person A to start a conversation.
2. Person A is standing in area A, and Person B in area B. Both facing ARI. ARI will then join Person B to start a conversation.
3. Person A and Person B are standing together in area A. They both face each other and are having a conversation. ARI will then join both of them to start a conversation.
4. Person A and Person B are standing together in area B. They both face each other and are having a conversation. ARI will then join them to start a conversation.

QUESTIONNAIRE

The questionnaire was similar to the one used in the experiments in HWU (without the questions regarding the dialogue, since we could not evaluate that part). Each question could be answered by one of the 5 presented options: strongly disagree, disagree, neutral, agree, or strongly agree:

1. I was afraid I would make mistakes when interacting with the robot.
2. I was afraid I might break the robot by doing the wrong thing.
3. I found the robot scary.
4. I found the robot intimidating.
5. I would be happy to interact with the robot again.
6. I found the robot fascinating.
7. I found the robot boring.
8. I found the robot pleasant to interact with.
9. The robot was friendly.
10. I thought the robot was nice.
11. I thought the robot was too close to me
12. I felt the robot approached me too quickly

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

13. The robot knew where I was

RESULTS AND DISCUSSION

We were able to recruit 6 pairs of participants. Unfortunately, due to last minute COVID-related situations of some of them, only 3 out of these 6 pairs of participants were able to take part in the experiments. Therefore the potential conclusions must be taken with care.

By giving numerical scores to the above answers (1-Strongly disagree to 5-Strongly agree), we are able to compute the mean and standard deviation of the answers. These are reported in the following Table 1:

	Question	Mean	Std dev
1	I was afraid I would make mistakes when interacting with the robot.	1.67	0.52
2	I was afraid I might break the robot by doing the wrong thing.	1.67	0.52
3	I found the robot scary.	3.50	1.22
4	I found the robot intimidating.	3.50	1.05
5	I would be happy to interact with the robot again.	4.00	0.63
6	I found the robot fascinating.	4.00	1.10
7	I found the robot boring.	3.00	1.10
8	I found the robot pleasant to interact with.	3.17	0.75
9	The robot was friendly.	2.83	0.75
10	I thought the robot was nice.	3.00	0.63
11	I thought the robot was too close to me	1.83	0.41
12	I felt the robot approached me too quickly	1.67	0.52
13	The robot knew where I was	3.00	0.63

Table 1 User Questionnaire Results

The color corresponds to different levels of standard deviation, and therefore of agreement among participants. Green, yellow and red correspond to moderate, mild and weak agreement respectively.

We can see, for instance, that the participants agreed in the fact that they are not afraid to make mistakes when interacting with the robot, and that they didn't feel that they could break the robot while interacting with it. Importantly, the participants also agreed that the distance and speed of the robot were comfortable. Less agreement is found on the interest and pleasantness of interacting with the robot, or on the robot's behavior (friendly, nice). We believe this is due to the fact that the robot could not verbally interact with the participants. Finally, the less agreement is found on the impression that the robot makes in terms of being scary, intimidating, fascinating or boring. Also, the corresponding average scores are fairly neutral, and it is therefore difficult to extract sound conclusions.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

CONCLUSIONS

These results are the outcome of a very long process that started with the delivery of the ARI units, continued with intensive efforts from several partners into developing software that was compatible with the ARI platform. It was followed up by a series of discussions and joint work to integrate all these modules, and concluded with the experiments conducted in the respective laboratories. These experiments are therefore the practical evidence that we managed to develop an initial software architecture for ARI, and to accomplish certain basic tasks. Our next steps are to conduct experiments in all laboratories, to guarantee the reproducibility of our research and development, on one side, and to deploy the robotic platform in the waiting room of the day-care hospital to start evaluating the platform in the relevant environment.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

HWU EXPERIMENTS

VALIDATION OF THE TASK PLANNER AND CONVERSATIONAL MANAGER

For the preliminary validation of the Conversational Manager and Task Planner these modules, together with the Robot Application, were integrated with the robot as ROS(Stanford Artificial Intelligence Laboratory et al.) nodes, and run on an external PC. Its navigation around the room was remote-controlled by a researcher in a separate room, who viewed the scene through the robot's on-board cameras (as seen in Figure 4).

The robot was provided with a female Text-to-Speech (TTS) voice (Acapela's UK English voice 'Rachel') set at 50% of the maximum volume. Its autonomous behaviors, which are designed to increase its "lifelikeness" - random eye blinks and slight movements of the arms - were enabled throughout.

Following feedback from the project partners on the initial conversational system Deliverable D5.1, the dialogue capabilities of the system were adjusted to focus more on task-based assistance. This meant disabling some bots and enhancing the capabilities of others. Using data from partner interactions with the system, the Reception, Directions and Persona Bot NLU and NLG were expanded with additional examples of user enquiries. For entertainment purposes the Quiz Bot was retained, however the Covid-19 information and advice bot was disabled.

Visual dialogue was not available in this implementation. The conversational drivers that 'advertise' the system's capabilities were adjusted to reflect these changes and highlight more often the task-based assistance available.

Figure 1 shows a close-up of the robot's display screen. The text on the lower line shows the output from the Automatic Speech Recognition (ASR) for the user's utterance. Above it is the robot's response as uttered by the TTS engine.

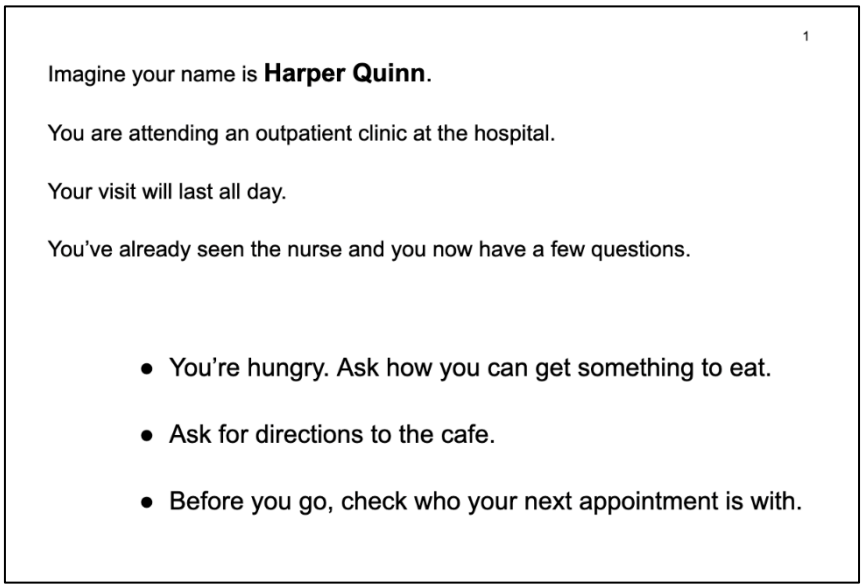
Figure 1 ARI Display Screen - TTS and ASR outputs

EXPIREMENT DESIGN

A formative user evaluation was carried out, in which participants were asked to interact with the SPRING system deployed on the ARI robot using a set of predetermined tasks, and to give their opinions of it.

EXPERIMENT PROCEDURE

The detailed procedure was as follows. On arrival, participants received a participant information sheet and consent form. Following consent, participants completed a preliminary questionnaire to collect demographic information. They were then supplied with a fictitious persona to use and asked to carry out three tasks designed to provide them with experience of the system's main capabilities: asking for directions (D), for information on catering arrangements (C) and for schedule information (S). The details of each task were varied to provide a variety of inputs with which to test the ASR and NLU components of the system. An example task scenario is shown below in Figure 2. A total of six task scenarios were used. The order in which the different enquiry types were presented to After interacting with the robot, participants completed a user attitude questionnaire about the experience. Finally, a de-briefing interview was carried out with the researcher, who was present throughout and followed a closely scripted procedure. Both researcher and participants wore a mask throughout the session to comply with local Covid-19 prevention measures.


1

Imagine your name is **Harper Quinn**.

You are attending an outpatient clinic at the hospital.

Your visit will last all day.

You've already seen the nurse and you now have a few questions.

- You're hungry. Ask how you can get something to eat.
- Ask for directions to the cafe.
- Before you go, check who your next appointment is with.

Figure 2 Example Task Scenario

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Figure 3 shows the experiment layout. A mock waiting room was created with rows of seats. Participants were only shown this after they had completed the initial questionnaires and had their tasks explained to them. Prior to entering the 'waiting room' they were advised by the researcher that on doing so they were free to take a seat, stand anywhere or approach the robot as they saw fit. The robot's approach behaviour was then adapted depending on what they chose to do, using a predetermined protocol. In cases where the participant did not themselves immediately move towards the robot, the researcher controlling the robot initiated an approach. If the user remained seated, or stayed within the waiting area defined as 50cm in front of the seats, the approach was halted 1m from the edge of the waiting area (this was required for ethical approval). Floor markings were used to indicate the correct stopping point (the longer line seen in Figure 3). If the user moved towards the robot as it approached, the robot was halted once the user themselves crossed this threshold, as seen via the robot cameras on the front of the torso. An example view is shown in Figure 4. On completion of the approach phase, the greeting message was played and the ASR was switched on.

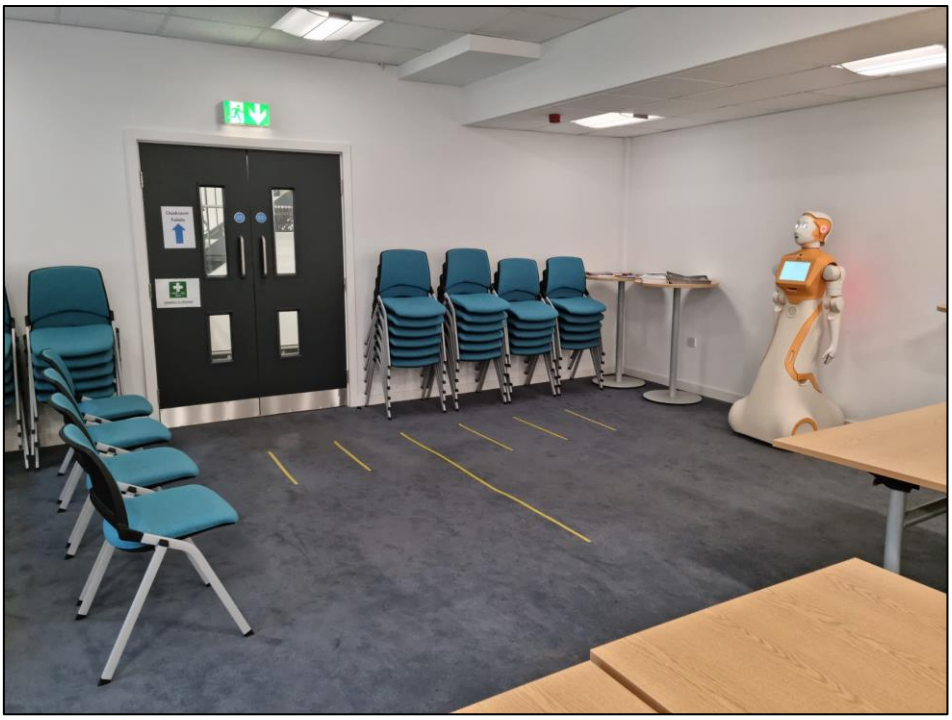


Figure 3 Experiment Layout - "In the Waiting Room"

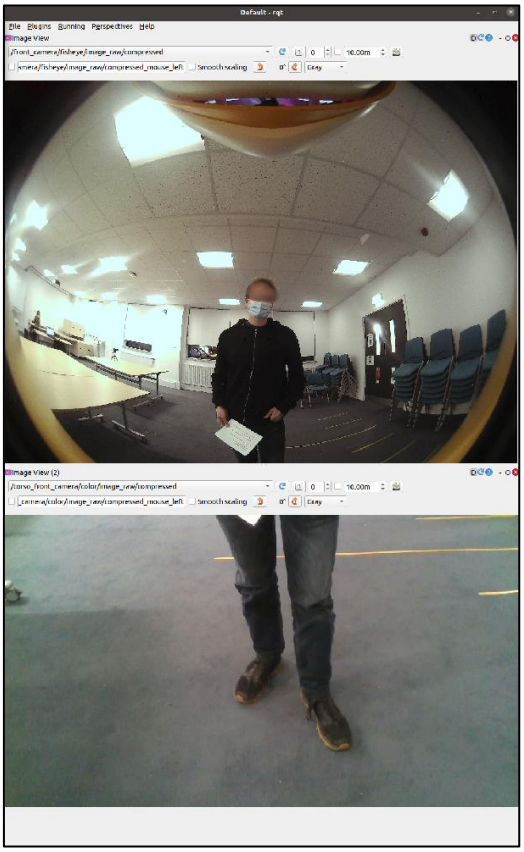


Figure 4 View Through Robot Cameras

DATA COLLECTION

A range of subjective and objective measures were collected. In order to assess system performance, objective data regarding participants' interaction with the robot was gathered via the system logfiles, together with audio-visual recordings of each session. User attitude was assessed through a combination of a post-use questionnaire and a de-briefing interview carried out at the end of the session. The questionnaire was in Likert format (Likert [1932]), which means participants are presented with a series of proposal statements e.g. "*I found talking to the robot enjoyable*" and are asked to indicate their level of agreement, in this case on a five-point scale from "strongly disagree" to "strongly agree". The questionnaire employed in this research comprised 33 statements and included constructs from the ALMERE questionnaire, which was designed to assess the acceptance of assistive social agent technology (Heerink et al. [2010]): namely, *Anxiety, Perceived Enjoyment, Perceived Sociability and Intention to Use*, together with the construct *Competence* taken from the Robotics Social Attributes Scale (RoSAS) questionnaire (Carpinella et al. [2017]). Importantly, a set of items specifically developed to measure the conversational aspects of the interaction was included.

PARTICIPANTS

A total of seven participants took part in the evaluation, all of whom were students at Heriot-Watt University. All seven were male; five were in the age group 18-25 years, two were aged 26-34 years. Two were non-native speakers.

RESULTS

TASK COMPLETION

Mean task completion across participants was encouragingly high at 71.4%, and was similar for each of the different task types (see Table 2). Few participants, however, completed the task on their first attempt, for any task type, as shown in Table 2.

	Catering	Directions	Schedule	Overall
Task Completion	71.4%	71.4%	71.4%	**71.4%**
Completed on first attempt	28.6%	14.3%	28.6%	**23.8%**
Mean attempts (success)	2.00	3.00	2.00	**2.33**
Max attempts (success)	3	4	4	**4**

Table 2 Task Performance

Closer inspection showed that the majority of failed attempts were the result of problems with the ASR. This predominately took the form of an ASR result that contained only part of the user's utterance, mis-recognition of critical words or a combination of the two (to the extent that the user's original intention could not be discerned). Failure to detect any speech at all was also an issue, particularly when users were positioned more than 1m away from the robot. Finally, a small proportion (5.4%) of failed attempts were due to users speaking to the robot during the approach phase, before the ASR was activated.

Reason for Failure	% Failures
Partial recognition / mis-recognition	62.2%
Speech not detected - user > 1m from robot	24.3%
Speech not detected - user ≤ 1m from robot	8.1%
Speech too early (before ASR activated)	5.4%

Table 3 Reasons for Task Failure

USER QUESTIONNAIRE

Scores on the five-point Likert scale were normalized according to the polarity of the question, with scores above three reflecting a positive attitude towards the system on the current issue, those below three indicating a negative attitude, with three the neutral point. Figure 5 shows the mean attitude scores for the questionnaire items specifically about the conversation with the robot.

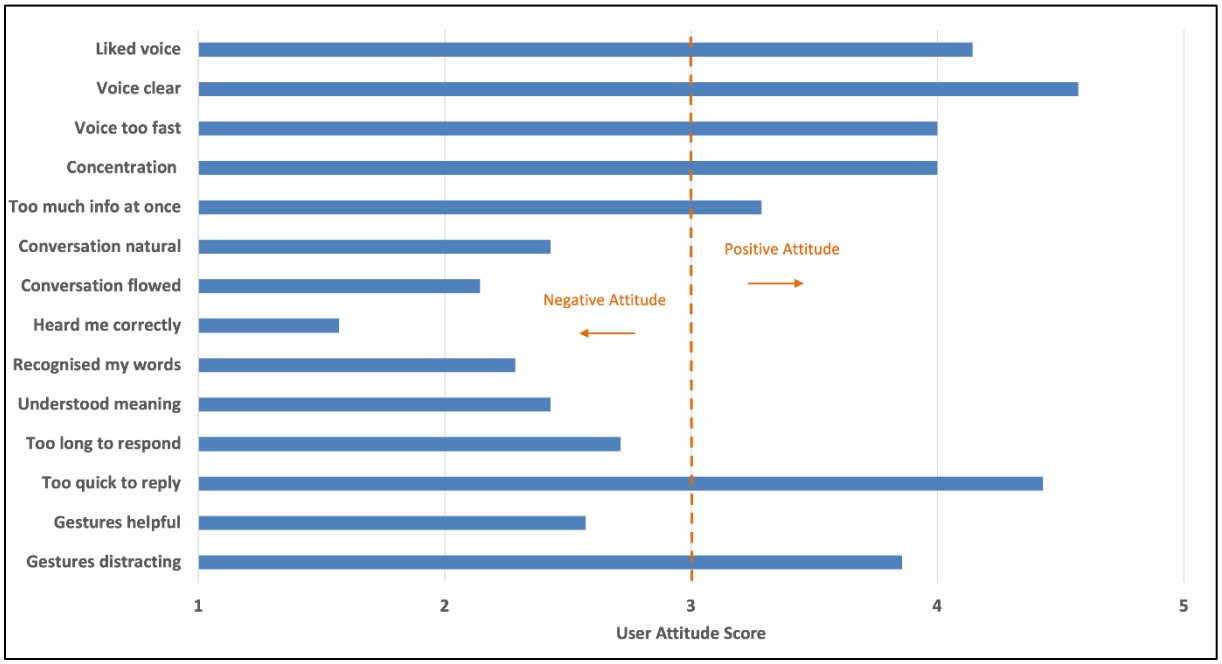


Figure 5 User Questionnaire - Conversational Items

Results showed that participants *liked the voice*, found it *clear* and did *not* find it *too fast*. Moreover, they did not feel they had to *concentrate* hard to understand what the robot was saying although they were more neutral towards the issue of being given *too much information at once*.
The conversation was rated negatively (below three on the 5-point scale) on a number of issues. Participants did not feel it was *natural* or that it *flowed* in the way they expected. The lowest score (just 1.57) was received for the statement "*The robot heard me correctly most of the time*", reflecting the issues with ASR described earlier. Related to this, "*The robot recognized the words I said most of the time*" also scored poorly (2.29), together with "*I felt confident the robot understood the meaning of my words*" (2.43). Clearly, participants felt they had difficulty making themselves understood.

Another area for improvement clearly indicated by the questionnaire is the system's speed of

response. Participants felt that it took *too long to respond*, meanwhile indicating very strongly that it was **not** *too quick to reply*. On the issue of the robot's gestures, participants did not find these *helpful*, rating this issue negatively on the 5-point scale (2.57). On average, however, they did not find them *distracting*. These results support findings from the interview, discussed below. Figure 6 shows the mean user scores for the constructs taken from the Almere and RoSAS questionnaires. It indicates broadly positive attitudes.

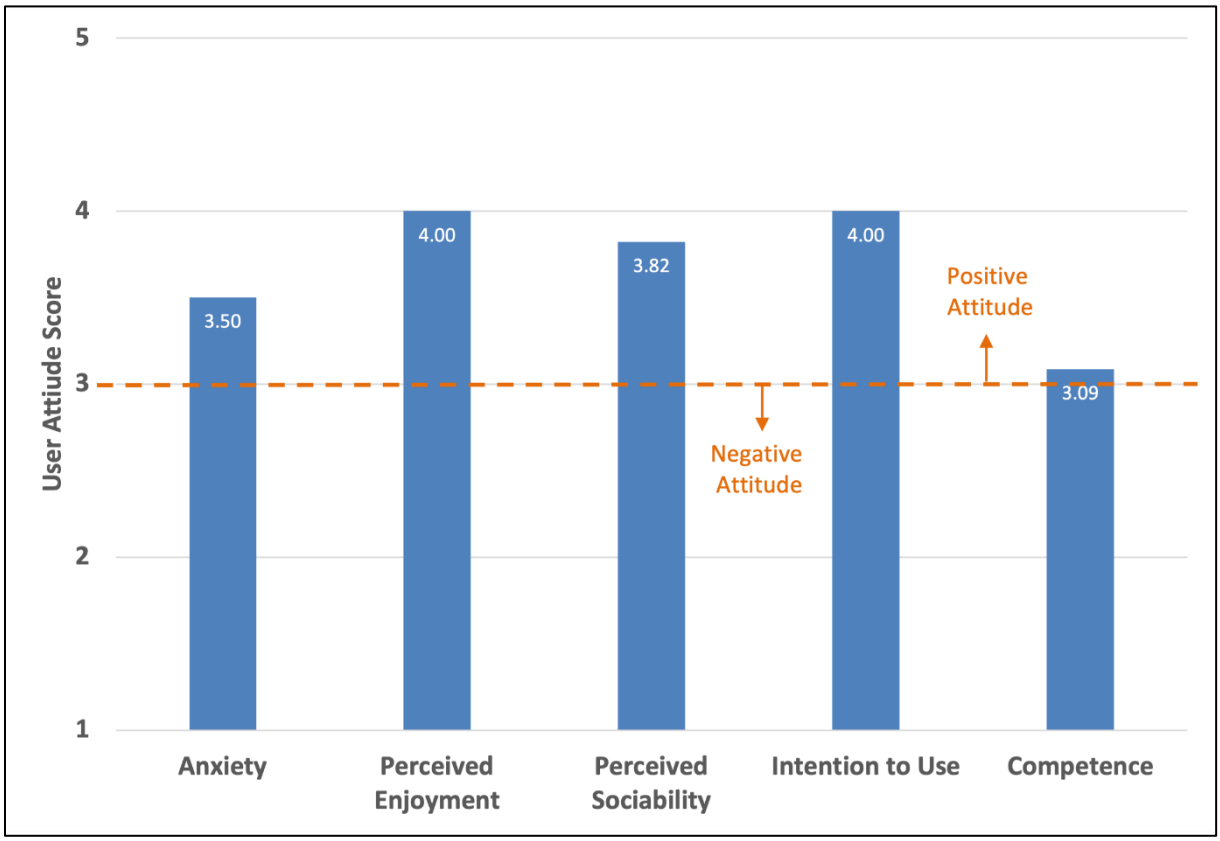


Figure 6 User Questionnaire - Almere and RoSAS Constructs

On the issue of user *anxiety*, the score was above three indicating participants did **not** feel anxious when interacting with the robot. Closer inspection of the individual items involved showed that participants did not find the robot *scary* (mean 4.00) or *intimidating* (3.86) but were more concerned they might make *mistakes* (3.00) or *break it* (3.14) in some way when talking to it. Encouragingly, participants *enjoyed* interacting with the robot, felt it was *sociable* (a pleasant conversational partner, friendly etc.) and with respect to intention to use, would *be happy to talk to the robot again*. The mean score for the robot's perceived *competence* was lower, at just above neutral. This is perhaps unsurprising given the issues around users being understood described earlier.

INTERVIEW

The de-briefing interview was designed to yield both quantitative and qualitative data on topics such as general likes and dislikes about interacting with the robot, its appearance, voice and display screen.

ROBOT APPEARANCE AND NONVERBAL BEHAVIOURS

When asked, the majority of participants (71.4%) said they liked the appearance of the robot. One commented "*Think it looks quite medical. The off-white/cream color fits well in hospitals, so would definitely fit the role.*".

Other positive comments related to the eyes moving and the presence of the screen to aid conversational understanding. Of the two participants (26.6%) who did not like the robot's appearance, one commented "*It looks like something out of a horror film ...that gazing stare...bit intimidating.*". The other felt the robot was too big and the design of the face could be nicer.

Participants were asked what gender they felt the robot was. The group was roughly split in opinion between female and gender-neutral (see Table 4). All participants mentioned the "*female*" or "*more feminine*" voice in their responses, but four also mentioned the design itself as "*neutral*" or "*androgynous*". One participant thought the robot had a feminine design. Participants were also asked how they felt about where the robot stopped in relation to them when initiating conversation. This was not applicable in two cases, since these participants walked straight up to the robot themselves, but amongst the five where the robot made an approach four (80%) felt the distance was about right. Just one felt the robot could have come a little closer (mentioning 15cm).

Responses were more divided on the issue of the robot's other nonverbal behaviors. These were random, slight movements of the arms and blinking of the eyes to give an impression of "aliveness". Slightly more participants disliked them (four or 57.1%) than liked them (three or 42.9%), two using the terms "*distracting*" and "*loud/noisy*"

Perceived Gender	Num Participants
Gender-neutral	3 (42.9%)
Female	4 (57.1%)
Male	0 (0.0%)

Table 4 Perceived Gender of Robot

in their reasons. Positive remarks included that they were "*pleasant*", "*realistic*". Regardless of preference, two participants also mentioned that the movements could relate more to the dialogue than they do currently.

ROBOT VOICE (TEXT-TO-SPEECH)

Participants were asked questions on the pace and volume of the TTS voice. All of the participants thought the speed of the TTS voice was about right. Two (28.6%) volunteered, however, that they thought the quantity of spoken information was too much. Both results support the findings from the Likert questionnaire.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Most participants (85.7%) thought the volume of the robot's voice (set at 50%) was about the right level, although three noted they might feel differently in a busy waiting room. One thought it would be too loud in those circumstances, making their conversation too public, the others thought it would be more difficult to hear.

DISPLAY TEXT

The majority of the group said the text on-screen was *too small* (5 or 71.4%). Comments included having to lean in closer to read it (two participants), with one noting "*it drew me away from looking at the face*". Despite this, almost all participants (85.7%) felt the text was helpful. The most frequent reason given (by five of the six in this group) was being able to read the system *output*. Three mentioned the fact that it showed the ASR output : "*what it thought I was saying*", so they could either "*correct it*", "*helped me assist in future questions - adjusting what I said*" or "*so at least if you get the wrong response you know why*". When asked how the display text could be improved, the most common suggestion (by four participants) was that the different roles of the text on-screen should be made clearer, i.e., a clearer distinction made between what was the robot's speech and the user's. Four participants also suggested enlarging or otherwise improving the font, two said better use should be made of the space, two thought the output text should be broken down into shorter, multiple lines on a single page rather than a single line at the top broken across different pages, while two also mentioned there should be some indicator when the system was '*listening/processing*'.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

REFERENCES

Stanford Artificial Intelligence Laboratory et al. Robotic operating system. URL https://www.ros.org.

Rensis Likert. A technique for the measurement of attitudes. Archives of psychology, 140, 1932.

Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. Assessing acceptance of assistive social agent technology by older adults: The almere model. *International Journal of Social Robotics*, 2(4):361–375, 2010.
ISSN 18754805. doi: 10.1007/s12369-010-0068-5.

Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. The Robotic Social Attributes Scale (RoSAS): Development and Validation. *In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, volume* Part F1271, pages 254–262, New York, NY, USA, mar 2017. ACM. ISBN 9781450343367. doi: 10.1145/2909824.3020208. URL https://dl.acm.org/doi/10.1145/2909824.3020208.