Usability Evaluation of a Robotic System for Cognitive Testing

Alessandro Di Nuovo, Simone Varrasi, Daniela Conti, Joshua Bamsforth, Alexandr Lucas, Alessandro Soranzo
Centre for Automation and Robotics Research
Sheffield Hallam University
Sheffield, United Kingdom
a.dinuovo@shu.ac.uk

John McNamara
Hursley Laboratory
IBM UK Limited
Winchester, United Kingdom
jon_mcnam@uk.ibm.com

***Abstract*—This abstract presents a preliminary evaluation of the usability of a novel system for cognitive testing, which is based on the multimodal interfaces of the social robot "Pepper" and the IBM cloud AI "Watson". Thirty-six participants experienced the system without assistance and filled the System Usability Scale questionnaire. Results show that the usability of the system is highly reliable.**

Keywords—Social Robot, Cognitive Assessment, IBM Watson

I. INTRODUCTION

A possible application of social robots is in the assessment of psychological abilities [1]. Robots can be programmed to perform specific actions in a very standardized way, which is very desirable in psychological assessments.

The robotic implementation of cognitive screening tests could be effective, because they are often repetitive and easy to take, but time-consuming for human assessors. Furthermore, administering tests in a truly standardized way may be problematic for a human assessor, for instance, in the case of the observation of developmental history and social skills, clinicians with different specialisations often disagree when evaluating the same patient [2]. A robot-led assessment can provide a series of advantages, among others: neutrality, objectivity, standardization of the interactions; and better acceptance and willingness to use of the robotic platform than non-embodied avatars [3].

We programmed the SoftBank Robotics "Pepper" to lead the administration of a cognitive test (Figure 1). The robot was able of giving instructions and automatically collecting users' answers via its multimodal interface, consisting of video, audio, and touch interface. Collected data was then processed and preliminary scored by the IBM Watson Cloud AI services. A complete description and an evaluation in terms of efficiency and reliability of the scoring for cognitive assessment via human-robot interaction can be found in [4].

In this abstract we focus on the usability of a prototype of the system in a simulated application scenario, where the participants interacted with the robot without any assistance or any further instructions than those provided by the robot itself.

II. MATERIAL AND METHODS

A. The Participants

A total of 36 healthy adults volunteered, 22 males and 14 females; the age range was 19-61, average 26.74 years, and standard deviation of 9.85. All of them completed high school, and 26 obtained a university degree, with the average number of years in education equal to 19.5, standard deviation is 4.16. All participants provided informed consent to use their data, video/audio recordings and pictures for scientific research.

B. The Robotic test for Cognitive Assessment

The Robotic test is inspired by the Montreal Cognitive Assessment (MoCA) [5], which is freely available from the official website. This is a brief cognitive test, widely adopted to detect mild cognitive impairments. Similarly, to the MoCA, our robotic test is composed of eight subtests with a total of 14 tasks that cover several cognitive domains:

- Visuospatial/executive (3 tasks): alternating letter/numbers trail making on touchscreen; drawing a cube and of a clock, including arrows and numbers (robot takes pictures);
- Naming: say the name of the three animals in pictures;
- Memory and Delayed Recall: the robot says 5 words that should be recalled after 5 minutes while the test goes on;
- Attention (3 tasks): digit span – repeat two sequences of digits; vigilance - react to the letter 'A' by touching the robot head, and say the serial 7 subtraction from 100;
- Language (2 tasks): repeat the two sentences said by the robot; and fluency – name words with F;
- Abstraction: tell why 2 pairs of words are connected, e.g. robot says "banana" and "orange" answer should be "fruits";
- Orientation: tell the full date – date, month, year, day - and the location – place, and city.

Further details on the implementation can be found in [4], [6].

C. The robot behaviours and IBM Watson services

We used IBM Watson Assistant (formerly Conversation) to organise the workflow of the test and the Watson Text-to-Speech service for generating the robot's voice. This was used for giving instructions and the other dialogue in the test. The interaction for assessing the visuospatial/executive skills includes also taking pictures of the user drawings, which are analysed by Watson Visual Recognition. Watson Speech-to-Text to perform speech recognition.

The speech was generated in advance and stored on the robot's internal memory. We opted for this solution to avoid accessing the cloud and minimize the latency. For the same reason, Speech-to-text and object recognition were performed after the administration of the test. One exception to this was the orientation subtest, in which the system had to recognise

The authors gratefully acknowledge the receipt of the IBM Shared Research Award in support of this work.

responses in real-time and ask for any missing details, such as the day of the week or the year. Another exception was occasionally addressing the participant by name (asked at the beginning). We opted for the British English female voice called 'Kate' to replace Pepper's default voice, which in preliminary tests was considered as childish and inappropriate for this type of task [6]. Figure 1 shows some examples of the interaction.

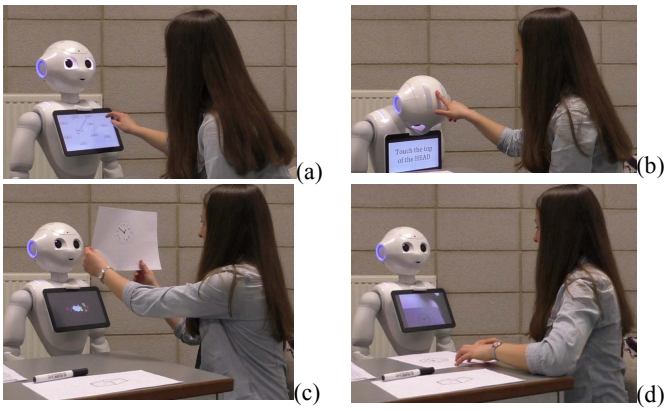

Figure 1. (a) Visuospatial/Executive: Alternating letters/numbers trail making (b) Attention subtest – Vigilance task; (c) drawing of the clock task: acquisition of the picture; (d) confirmation that is correct.

The robot program included two interactive tasks: *a welcome task*, before the test, and *a thank-you task* at the end. In the welcome task, Pepper introduced itself and asked the participants age, gender and years of education. This aimed both to collect information about the person, as well as to let the participant familiarise with the interaction modalities.

The administration was temporized in such a way that Pepper always performed in the same way. The robot was impassable, and it did not react to any of the participant's responses. Instructions were repeated only when this was allowed by the MoCA manual for the corresponding task. If the participant did not complete a task, the session continued until the internal timer expired. The timing of each task was set empirically as the maximum time taken in preliminary tests.

D. The System Usability Scale (SUS)

At the end of the test, participants were requested to fill a usability questionnaire. We used the System Usability Scale (SUS) which is a reliable and widely adopted usability scale that can be used for assessments of technological systems usability [7]. It consists of ten-items with a five-point (1-5) attitude Likert scale, providing a global view of subjective assessments of usability. SUS score is processed to be on a scale of 0-100. According to normative data [7], sufficiently usable products have SUS scores above 68, with better products scoring in the high 70s to upper 80s.

The questions asked in our experimentation were:

1. I think that I would like to use this system frequently.
2. I found this system unnecessarily complex.
3. I thought this system was easy to use.
4. I think that I would need assistance to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found this system very cumbersome/awkward to use.
9. I felt very confident using this system.
10. I needed to learn a lot of things before I could get going with this system.

In the analysis, negative questions (2,4,6,8,10) scores are inverted, thus high scores (4,5) identify positive answers for all.

III. SUS Scores

The average SUS score was 76.4, the median 80, the standard deviation 14.7, the maximum 92.5, and the minimum 35.

79% and 57% of the participants totalled a score equal to or higher than 70 and 80 respectively. 14% of the participants scored 90 or above. Table I shows the scores to each question.

TABLE I. SUS Questionnaire Scores*

#	*Average*	*Median*	*Mode*	*Negative*	*Positive*
1	3.1	Neutral	Neutral	21%	36%
2	4.5	Positive	Positive	0%	86%
3	3.9	Positive	Positive	7%	79%
4	4.2	Positive	Positive	7%	93%
5	4.0	Positive	Positive	0%	93%
6	4.4	Positive	Positive	0%	93%
7	4.2	Positive	Positive	7%	93%
8	4.1	Positive	Positive	14%	71%
9	3.4	Positive	Positive	21%	57%
10	4.7	Positive	Positive	0%	93%

*Negative questions (2,4,6,8,10) are inverted: Score=(6-actual Score)

IV. Discussion and Conclusion

The SUS scores suggest the prototype can be ranked among the most usable software products available on the market.

This positive result is particularly encouraging for continuing the research and development of a system that can assist clinicians in the screening of cognitive skills and in the early detection of neurological impairments like dementia.

References

[1] S. M. Rabbitt, A. E. Kazdin, and B. Scassellati, "Integrating Socially Assistive Robotics into Mental Healthcare Interventions: Applications and Recommendations for Expanded Use," *Clin. Psychol. Rev.*, vol. 35, pp. 35–46, Jul. 2015.

[2] B. Scassellati, H. Admoni, and M. Matarić, "Robots for Use in Autism Research," *Annu. Rev. Biomed. Eng.*, vol. 14, pp. 275–294, Jan. 2012.

[3] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *Int. J. Hum. Comput. Stud.*, vol. 77, pp. 23–37, 2015.

[4] A. Di Nuovo, S. Varrasi, A. Lucas, D. Conti, J. McNamara, and A. Soranzo, "Assessment of Cognitive skills via Human-robot Interaction and Cloud Computing," *J. Bionic Eng.*, vol. to appear, 2019.

[5] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment," *J. Am. Geriatr. Soc.*, vol. 53, no. 4, pp. 695–699, 2005.

[6] S. Varrasi, A. Lucas, A. Soranzo, J. McNamara, and A. Di Nuovo, "IBM Cloud Services Enhance Automatic Cognitive Assessment via Human-Robot Interaction," in *New Trends in Medical and Service Robotics*, 2019, pp. 169–176.

[7] A. Bangor, P. T. Kortum, and J. T. Miller, "An Empirical Evaluation of the System Usability Scale," *Int. J. Hum. Comput. Interact.*, vol. 24, no. 6, pp. 574–594, 2008.