J Bionic Eng 16 (2019) 526–539
DOI: https://doi.org/10.1007/s42235-019-0043-2

Journal of Bionic Engineering
http://www.springer.com/journal/42235

Assessment of Cognitive skills via Human-robot Interaction and Cloud Computing

Alessandro Di Nuovo[1*], Simone Varrasi[1], Alexandr Lucas[1], Daniela Conti[1], John McNamara[2], Alessandro Soranzo[1]

1. *Sheffield Hallam University, Sheffield S*1 1*WB, United Kingdom*
2. *IBM Hursley Lab, Winchester SO*21 2*JN, United Kingdom*

Abstract

Technological advances are increasing the range of applications for artificial intelligence, especially through its embodiment within humanoid robotics platforms. This promotes the development of novel systems for automated screening of neurological conditions to assist the clinical practitioners in the detection of early signs of mild cognitive impairments. This article presents the implementation and the experimental validation of the first robotic system for cognitive assessment, based on one of the most popular platforms for social robotics, Softbank “Pepper”, which administers and records a set of multi-modal interactive tasks to engage the user cognitive abilities. The robot intelligence is programmed using the state-of-the-art IBM Watson AI Cloud services, which provide the necessary capabilities for improving the social interaction and scoring the tests. The system has been tested by healthy adults ($N = 35$) and we found a significant correlation between the automated scoring and one of the most widely used Paper-and-Pencil tests. We conclude that the system can be considered as a screening instrument for cognitive assessment.

Keywords: socially assistive robotics, brief cognitive testing, human-robot interaction, neurological screening, cloud computing

1 Introduction

An area of expansion for Artificial Intelligence (AI) and robotics is known as Socially Assistive Robotics[1–3] (SAR). In SAR, robots are a physical representation of modern AI, which is often located in the “Cloud”, i.e. hosted in massively powerful supercomputers that expose computationally intensive state-of-the-art AI algorithms, like the deep learning artificial neural networks, and make them accessible via the internet from low power, energy saving clients like robots[4–6]. Cloud services are offered by the major corporations, for instance, Amazon, Microsoft, Google, and IBM, which provide a standardized and easily reproducible environment for development and testing of Human-Robot Interaction (HRI) applications. Despite being a relatively recent innovation, these Cloud services have been utilised in a number of studies in HRI[7].

The approach of embedding AI into robots widens the scope of potential applications by “physically” engaging and socially interacting with the user, while simplifying the access to the new technologies via a more natural and intuitive interface. This provides a means to overcome the interaction difficulties that can be experienced, particularly by the elderly in engaging with digital technology, thanks also to the multiple interaction multimodalities[8,9], e.g. audio, video, gestures, touch. As part of SAR research, the integration with service robotics and smart environments have been extensively explored too, especially for the elderly who may need complex assistive technology to support healthy ageing[10,11].

SAR have been increasingly studied for health and clinical applications, leading scholars to suggest a stable integration of HRI in healthcare settings[12]. Robots demonstrated potential benefits when employed as therapeutic assistants or caregivers. Indeed, the efficacy of artificial agents has been documented with children[13–15] as well as with elderly people in a variety of neurological and psychiatric conditions[16]. Recently,

***Corresponding author:** Alessandro Di Nuovo
E-mail: a.dinuovo@shu.ac.uk

initial evidence has been collected on the viability of robotic assessments, for instance in the measurement of Patient Reported Outcome[17] and in early diagnosis of Autistic Spectrum Disorder (ASD)[18,19]. Petric proposed a "robotic autism spectrum disorder diagnostic protocol", in order to evaluate child's reaction when called by name, his/her symbolic and functional imitative behaviour, his/her joint attention, and his/her ability to communicate via multiple channels simultaneously, but the results are not clear and definitive[20]. Another diagnostic method for ASD is proposed by Wijayasinghe *et al.*[21], who considered HRI a way to objectively evaluate imitation deficits. In this case, the robot Zeno performs upper body gestures and it encourages the child to imitate them, while the robot automatically assesses the child's behaviour. However, it seems little work has been done regarding other pathologies. Kojima *et al.*[22], for instance, published some speech recognition experiments with elderly people using the robot PaPeRo, aiming at the development of a computerized cognitive assessment system.

It is evident that robots have the potential to be successful assistants in psychological assessment. Robots can be programmed to perform specific, repeatable action, providing the benefit of attainable standardization. Therefore, the robotic implementation of quick screening tests could be promising, because they are often repetitive and easy to take, but time-consuming for human assessors. A robot-led assessment can provide a series of advantages, among others: assessor neutrality, objective measurement of social behaviour, standardization of the interaction, better acceptance of the robotic platform than a non-embodied computer[23,24]. An example is the observation of developmental history and social skills, where clinicians with different specialisations often do not agree when evaluating the same patient[25].

Recently, Desideri *et al.*[26] presented an interesting study on human-robot interaction for Brief Cognitive Testing (BCT), in which they showed that participants' emotional reactions toward the human examiner and the humanoid robot NAO did not differ. This study supports the hypothesis that the use of social robots could increase the ecological validity of computerised cognitive testing and decrease distracting stimuli from the interviewer's face, while at the same time would allow the automatic tracking and recording of the examinee's behaviours. Meanwhile, Rossi *et al.*[27] tested a system for psychometric evaluation with a social robot, which was a preliminary version to the one presented in this article. The test involved 19 older participants (age range = 53–82, mean = 61.16) and investigated the personality factors and technology acceptance showing that the openness to experience factor along with anxiety, trust, and intention to use, correlate with the performance in the psychometric tests.

However, other than the preliminary examples presented above, robotic psychological assessment is almost unexplored, especially with adults. The evidence is limited, the pathologies studied are very few, comparative studies (robotic assessment vs human assessment; robotic assessment vs computerized assessment) are often not available. It is evident that more research and experimental validation is required.

With the aim of increasing the experimental knowledge and providing novel evidence of the feasibility of this field of application for SAR, the present article demonstrates a novel system, named Cognitive level Assessment via Human-robot Interaction (CATHI), that integrates a social robot and a Cloud AI system as a screening tool for Mild Cognitive Impairment (MCI)[14]. MCI was typically associated with cognitive related disorders in older people and considered merely as a prodromal stage of dementia[28], but recent research has recognized MCI as a risk factor to develop more severe cognitive deterioration[29,30]. It is very important to detect the so-called predictors of conversion and to determine if a patient may develop dementia, to provide for efficient planning of rapid intervention with adequate treatment, including non-pharmacological rehabilitation training[31]. For early detection, the markers to be considered can be both biological and psychometric[32]. However, brief psychometric tests are usually preferred for screening, because they are quicker and inexpensive, indeed cognitive deficits are often initially diagnosed in this manner. We would remark that MCI symptoms can be found also the younger population as a consequence of psychiatric disorders like schizophrenia and depression (see section **2.1**) and therefore an MCI test can useful for their detection[33,34].

There are two main goals in presenting this work: 1) to demonstrate the first version of a brief cognitive test for MCI that can be fully administered by an autonomous social robot and scored by a Cloud AI system; 2) to investigate the feasibility of performing a cognitive screening through a social robot with a non-clinical population.

The rest of the paper is organised as follows. Section **2** describes the materials used to build the robotic cognitive assessment, i.e. the psychometric cognitive test, and the underpinning technology, and the methodologies used to experimentally validate the new tool. Section **3** presents and discusses the results of the experimental test of CATHI and the validation of the scores. Section **4** gives our conclusion.

2 Materials and methods

Solutions adopted in this work are inspired by previous preliminary experiments with the SoftBank humanoid social robot Pepper[35]. It was found that the score automatically calculated using the robot software often failed to provide the correct score because of the failures of the embedded speech and object recognition interfaces[23,36]. To tackle this problem, after thorough consideration of what the market can offer, we selected the IBM AI Cloud services, named "Watson", which provide a comprehensive set of easy-to-use tools for speech recognition (speech-to-text) and production (text-to-speech) and object recognition from picture, which definitely improved the robotic system interaction and automated scoring[37].

2.1 The Montreal Cognitive (MoCA) test

An instrument for BCT is the Montreal Cognitive Assessment, widely known as the MoCA[38], which is freely available from the official website[39]. It has been initially validated for 55-85 year olds and translated into 46 languages and used in more than 100 countries for detecting MCI.

Importantly, the MoCA test assesses multiple cognitive domains, therefore, it can be a useful screening tool of MCI in a variety of neurological conditions that affect both younger and older populations, such as dementia[38], brain tumours[40], vascular cognitive impairment[41], schizophrenia[33], early stages of Parkinson's disease[42], Huntington's disease[43], sleep behaviour disorder[44] and depression[34]. In the work described in this article, the robotic test was inspired by the MoCA Full 7.1 English version, which can be found on the MoCA website[39]. The MoCA test is composed of eight subtests with a total of 14 tasks that cover several cognitive domains:

- Visuospatial/Executive, 3 tasks, 5 points: alternating letter/numbers trail making (1 point if correct); copying a cube (1 point if correct); drawing of the clock, including arrows and numbers (1 point for each element up to 3 points);
- Naming, 1 task, 3 points: name the three animals in pictures (1 point for each animal, up to 3 points).
- Attention, 3 tasks, 5 points: digit span – repeat two sequences of digits (1 point for each if correct); vigilance - react to the letter A (1 point if less than 2 errors), and serial 7 subtraction (up to 3 points);
- Language, 2 tasks, 3 points: repeat the two sentences (1 point for each if correct); and fluency – name words with F (1 point if more than 10 words are named);
- Abstraction, 1 task, 2 points: tell how 2 pairs of words are connected, e.g. banana and orange are fruits, (1 point for each pair if correct);
- Memory and Delayed Recall, 1 task, 5 points: the assessor says 5 words that should be recalled after 5 minutes (1 point for each word spontaneously recalled without help);
- Orientation, 1 task, 6 points: tell the full date – date, month, year, day - and the location - place and city – (1 point for each correct element).

When scoring the test, the assessor must compensate for the education level, i.e. one point is added if the person has twelve years of education or less. The maximum total score is 30. It is usually considered normal to score 26 points or above, while a score of 22 or below may indicate a cognitive problem. A recent study[45] set the cut-off score for further examination and MCI diagnosis to 23 to reduce false positives.

The English version of the MoCA test has two alternative versions, 7.2 and 7.3, equivalent to the main Full 7.1 version[46]. The 7.2 version was used for validating the robotic cognitive test.

2.2 Cognitive level Assessment via Human-robot Interaction and IBM Watson AI (CATHI)

The robotic platform used for building a prototype of CATHI and running our experiments is the humanoid "Pepper"[35] manufactured by SoftBank Robotics, the latest commercial product specifically designed for HRI and equipped with state-of-the-art interactive interfaces: a touchscreen, human-like movement, pressure sensors, object recognition, speech production, and comprehension, age, gender, emotion and face detector.

We programmed the Pepper robot to lead the administration of a MoCA-like psychometric test. The robot gives instructions and collects the data via its multimodal interfaces (video, audio, touch) for the subsequent processing and automated scoring by the IBM Watson AI. Indeed, the entire administration session is audio-recorded by the robot's microphones. The robot also takes pictures of the user drawings for the Visuospatial/Executive skills. The robotic cognitive test measures the same areas of the MoCAs, hence the subtests have the same names, as shown in the Table 1.

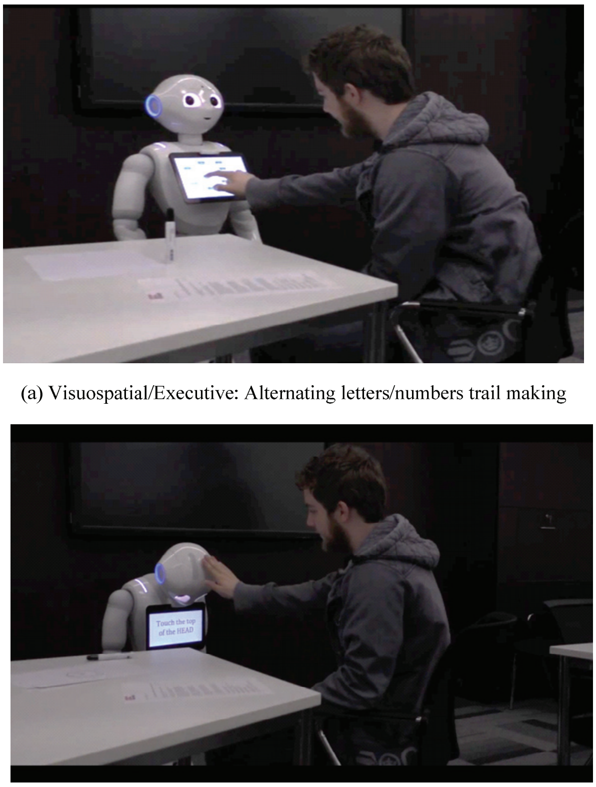
Table 1 presents the technologies used to implement the 7 subtests. These are divided by the Pepper's robot embedded sensors and interfaces, i.e. the touchscreen and the pressure sensor on the head, and the three IBM Watson AI services described in the following section **2.3**. In addition, the robot was also using the non-verbal communication cues to support a more natural HRI, i.e. it was gesturing while speaking, rotating the head to demonstrate attention to the user, and changing the eyes' LEDs to show the internal state and guide the interaction, e.g. blinking blue when listening. The non-verbal communication cues were explained to the user at the beginning of the session.

Fig. 1 presents two examples of the non-verbal HRI: in Fig. 1a the participant uses the touchscreen to connect letters/numbers in the trail making task; in Fig. 1b the participant touches the sensor on the head to demonstrate attention in the vigilance task.

In summary, Watson text-to-speech was used for both providing test instructions and test dialogue. The digital speech was generated in advance and stored locally in the robot to avoid the latency associated with accessing the cloud. Watson speech recognition was used to score: Naming, Attention, Language, Abstraction, Delayed Recall, and Orientation. The Watson object recognition system was used to assess Visuospatial/Executive skills by having the user draw a cube and a clock on two sheets of paper, which was also used to assess the executive functioning, including manual skills and motor coordination. Speech-to-text and object recognition are performed after the administration, the only exception is the orientation subtest, in which the robot should recognise the full date, including the day of the week, in real-time and ask for any missing piece of information.

Table 1 Technologies used to implement the 7 Subtest of the different cognitive domains

Subtest	Native Pepper robot's sensors/interfaces	IBM Watson AI services		
		Text-to-speech	Speech-to-text	Object recognition
Visuospatial/ Executive	Task 1: Touchscreen shows characters and numbers to connect. Task 2: Touchscreen shows the image to copy on the tablet	Task instructions		User's drawings
Naming	The touchscreen shows images to recognize	Task instructions	Name of animals	
Attention	Head sensor, for the Vigilance task	Task instructions, sequences of numbers to recall	Numbers	
Language		Task instructions, sentences to repeat	Sentences & words starting with "F"	
Abstraction		Task instructions, items to categorize	Connections between words	
Delayed Recall		Task instructions, words to recall	Words	
Orientation		Instructions	Full date including the day of the week	



(a) Visuospatial/Executive: Alternating letters/numbers trail making

(b) Attention subtest – Vigilance task

Fig. 1 Examples of the human-robot interaction during the robotic administration.

2.3 Integration of the HRI with IBM Watson Cloud AI services

IBM Watson is a subset of the IBM Cloud services that focus on providing enhanced artificial intelligence solutions, providing what is referred to by IBM as a cognitive capability which can be embedded into technology applications. For the autonomous scoring of the CATHI system, we used the IBM Watson solutions for post-processing the data collected from the HRI sessions. We used Watson Assistant (formerly Conversation) to organise the workflow of the test, Watson Text to Speech for generating the robot's voice during the administration, Watson Speech to Text to perform speech recognition and Watson Visual Recognition to perform the analysis of the hand-drawn images. We used NaoQi framework's Python SDK for Pepper robot to write Python scripts that ran on the robot and handled its behaviour during the administration. The communication with the IBM Cloud was implemented as HTTP POST requests, generated by Watson Developer Cloud module for Python. Further implementation details are reported in the following subsections.

2.3.1 Watson Assistant and text to speech

Watson Assistant was used to organising the workflow of the test and storing the robot's instructions for each subtest, with some basic analysis of the participant's response. For example, during the Orientation Date subtest, based on the response it received, Assistant would ask for missing details, such as day of the week or year. Another example would be occasionally addressing the participant by name if provided. We opted for using Watson's British English female voice called 'Kate', which is provided as an option in Watson text to speech service, to replace Pepper's default voice, which could be described as childlike and so inappropriate for administering a cognitive assessment test. To reduce the latency, the speech was often pre-generated as most of the dialogue was fixed because of the standardisation requirement.

2.3.2 Watson speech to text

Pepper's default voice recognition system is designed for recognizing individual words, which was not suitable for the parts of the test, e.g. the Language assessment, where long sentences must be recognized. Moreover, Pepper's recording system would often pick up constant noise coming from the cooling fans located inside the robots' head just below the microphones. This noise had a negative impact on the quality of the audio source, which consequently degrades the speech recognition performance. Watson speech-to-text service was able to overcome these limitations by supporting the analysis of longer sentences, including context analysis, and the customisation of the model with additional samples to embed the noise.

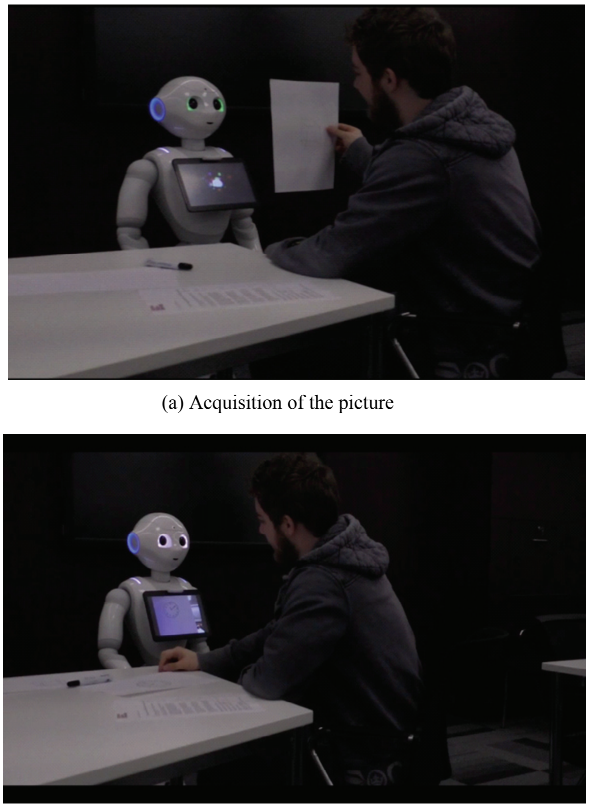
Considering the language background of participants and the characteristics of audios, we used as a base for the customisation the British English model, named "*en-GB_BroadbandModel*". Furthermore, to better explore the characteristics of the Watson service and set-up the parameters, we did not set a confidence threshold under which the transcriptions were discarded, and we considered the 10 best alternatives of recognition transcription according to the confidence level.

Utilising these features, the Watson score was calculated via a customised and flexible approach which drastically improved the system performance as reported in our preliminary study[37]. It was customized because we used the language customization service, to refine the speech-to-text model for the specific recognition requests of our cognitive test. Indeed, we created three customized models by adding specific corpora. One model was trained for the recognition of numbers, months and weekdays (Attention and Orientation tasks), one for the recognition of all the words starting with *B*, *F* and *S* (Fluency task), and the last one for all the remaining tasks. The customization weight was set to 0.9.

The scoring was also flexible because the calculation accepted a certain percentage of error so that the points were given if one of the transcriptions fitted for at least the 70% the target word or string.

2.3.3 Watson visual recognition

As part of the administration procedure, the robot asked the participants to produce hand-drawing of a cube and a clock showing a specific time. The procedure is exemplified in Fig. 2. Participants presented these pictures so that the robot could take a photo (Fig. 2a). The interaction included the robot showing the picture taken on the tablet and asking the participant to check if the drawing was captured correctly and, if not, offering to retake the picture (Fig. 2b).



(a) Acquisition of the picture

(b) Confirmation that the drawing has been captured correctly

Fig. 2 Interaction in the drawing of the clock task.

The pictures were sent to Watson visual recognition for analysis. The Watson service generated a general class and individual subclass of the object/setting the recognition system deemed as central on the picture. We reprocessed the drawings of cubes and clocks made by participants with the default visual recognition. The drawings were considered correct and the points were awarded when the cube was recognized as polyhedron and the clock as clock or wall clock. For the hours, a point was given when numbers from 1 to 12 were recognized. The same applies for the clock hands/arrows.

2.4 Experimental procedure

The experimental procedure included two sessions for each participant: the traditional “Paper and Pencil” administration and the robotic administration. Each session was carried out in the same room, within the premises of Sheffield Hallam University. The participants were invited to enter the room one by one and asked, first of all, to read and sign the forms regarding the processing of personal data.

The traditional “Paper and Pencil” session was entirely run by the experimenter and timed as well. In order to reduce any learning effect between the two administrations, for the Paper and Pencil, the 7.2 alternative version of the MoCA test was used. Moreover, the two sessions were spaced by at least five days, and the modality – robotic or paper – experienced first was alternated.

The session was entirely led by Pepper: the robot gave the instructions and audio-registered the answers. An experimenter was present in the room, but he did not interfere with the interaction and maintained a marginal position. The session was externally video-recorded and timed. After each session, the participants provided feedback and comments about the experience.

The robotic administration instructions for the experimental procedure are derived from the English

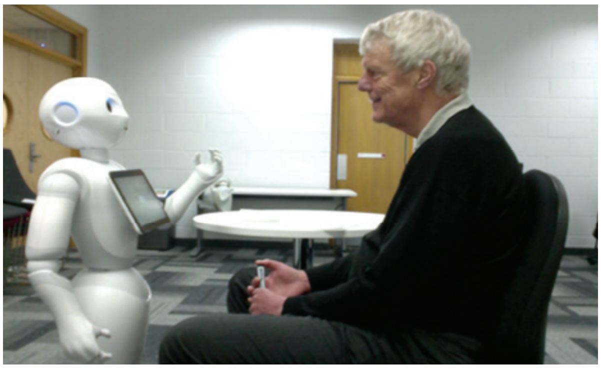


Fig. 3 Example of welcome task that precedes the administration of the cognitive test.

MoCA manual. Following the manual, the robot was programmed to perform two interactive tasks: the welcome task, at the beginning of the test, and the thank you task at the end. In the welcome task (Fig. 3), Pepper introduced itself and asked the participant to provide his/her age, gender and years of education. This was meant both to collect important information about the person, as well as to train the participant on the interaction modalities. For a more effective interaction[47], the robot engages with the participant by recognizing the face and rotating the head to follow. This is a standard HRI procedure that includes also to move the robot arms and hands as suggested in the literature for more effective and natural interaction with adults[48].

The administration was temporized in such a way that Pepper always performed the same way, and it did not react to the participant's responses. Therefore, if the participant did not complete a task, the session continued when those internal timers expired. Furthermore, even if the participant asked, instructions were not repeated where this is not allowed by the MoCA manual for the corresponding task. The timing of the administration was regulated by internal timers that were set empirically as the maximum time taken in preliminary tests.

Pepper audio-recorded the whole session and took photos of the second and third tasks' drawings; moreover, it produced a Dialog file with the transcription of the verbal conversation with the participant and a Log file containing information about any technical failures that occurred, the automatic score achieved, any wrong answers received, and any tasks ignored. This way, a clinical psychologist could fully review the administration and re-evaluate for validation.

Following the procedure described above, the following scores were obtained:

(1) the Watson CATHI score, which is autonomously calculated by the system using the IBM Watson AI Cloud services to process the audio and video recorded by the robot;

(2) a Supervised CATHI score, which is the score of the robotic test calculated by a clinical psychologist, who re-scored the performance through the recordings of the HRI;

(3) the Paper and Pencil MoCA score, which was calculated via the administration of an alternative version of the MoCA test by the clinical psychologist, who calculated the score in order to provide a reference score for testing external validity.

2.5 Statistical methods

The Cronbach's alpha coefficient and the Spearman-Brown (SB) coefficient were used to analyse the reliability of the overall Watson score. The alpha is a measure of internal consistency, i.e. how closely related the items (subtests in our case) are as a single score. We consider acceptable an alpha value greater than 0.600, with a stronger reliability above 0.700, for details see[49].

The SB coefficient is another index of internal consistency, known also as split-half reliability because it takes the correlation between the two halves of the test for estimating the level of reliability for the full-length test. The SB coefficient is used to predict reliability when the number of items changes, e.g. doubles. Here we use the SB coefficient because it relates the reliability with the test length, which should not be excessively long to avoid a cognitive overload that can bias the overall result. In practice, a test with SB coefficient equal to 1 can be halved without losing efficacy, while below 0.5 can indicate an inconsistency between the two halves and may suggest extending the test.

Spearman correlations were then calculated to explore the relationship among the Watson, the Supervised and the Paper and Pencil MoCA scores, the latter represents the external validity for the test. The Spearman correlation was chosen for the shape of the distribution and the typology of data. A correlation above 0.5 is considered a high effect-size according to Cohen's criteria[50].

Following the positive results of the correlations, a multiple linear regression analysis was applied to confirm the relations and derive the predictive function which allows deriving MoCA score from the robotic test scores. The existence of the linear function that directly relates the subtests scores to the MoCA confirms the validity of the robotic test.

The multiple linear regression was performed within the scoring modalities to find the proportion of the contribution of each subtest to the overall score. The regression coefficients can give an indication of the amount of variance that can be attributed to each test. A similar distribution of the coefficients is expected as the subtest covers the same cognitive domains.

Finally, we conducted a Paired Samples *t*-test, which is used to determine whether the mean difference between two sets of observations is zero. In practice, a new set is generated by calculating the paired differences, then, the *t*-test is applied to test the null hypothesis on the new set.

The statistical analyses presented in this section were performed with the SPSS software (version 24). A detailed description of the procedures and detailed formulation can be found in Ref. [51, 52].

2.6 Participants

A total of 35 healthy adults voluntarily participated in our experiments, 22 males and 14 females, the age range was 19 – 61, average 26.74 years, and standard deviation of 9.85. Participants were all proficient in British English, but only 14 of them were native speakers, i.e. did all their education in the UK. Most foreign participants had a distinctive accent according to their native language. All of them completed high school, and 26 obtained a university degree, with the average number of years in education equal to 19.5, the standard deviation of 4.16. The study received ethical approval from the committee of the Faculty of Arts, Computing, Engineering and Science, Sheffield Hallam University. All participants gave informed consent to use their data, video/audio recordings and pictures for scientific research purposes.

3 Results

Descriptive statistics are presented in section **3.1** to summarise the experimental results. Section **3.2** reports the Alpha and SB coefficients along with the Spearman correlations to evaluate the reliability of the robotic test. Section **3.3** presents the correlations between the Supervised and the Watson CATHI scores for each subtask to validate the results of the multiple regression. Finally, section **3.4** simulates a use case study to show the practical application of the proposed CATHI system as a screening tool.

3.1 Descriptive statistics

Table 2 presents the descriptive statistics calculated for the Paper-and-Pencil MoCA, the Supervised (benchmark) and the Watson's scores: mean, median, standard deviation (Std. Dev.), minimum (Min) and maximum (Max).

Table 2 shows that the median score calculated by Watson is the same as the Supervised score while both are lower than that of the MoCA. The standard deviation of the CATHI scores is higher than that of the MoCA because of some non-native participants, who scored very low because difficulties in understanding instructions and some of their answers were misunderstood by the Watson speech recognition system.

To investigate in the detail the cognitive domains, Table 3 reports the descriptive statistics for each subtest according to the scoring modality.

The Supervised CATHI scores are significantly lower (>1 point) for the Visuospatial/Executive and the Delayed Recall. The first is mainly due to the draw a clock task: in the case of the Paper and Pencil, the participant was able to see the scoring table and indirectly suggested to draw “contour”, “numbers” and “hands”, whereas in the case of the robot this information is not given and many missed these details and lost up to 2 points. In the case of the Delayed Recall we haven't identified a main cause, however, some participants

Table 2 Descriptive statistics

	Paper and Pencil MoCA	Supervised CATHI score	Watson CATHI score
Mean	25.0	21.2	19.4
Median	25.5	21.0	21.0
Std. Dev.	2.07	5.22	5.64
Min	21	10	9
Max	28	28	27

Table 3 Descriptive statistics for subtest scores

	Visuospatial/Executive	Naming	Attention	Language	Abstraction	Delayed Recall	Orientation
Paper and Pencil MoCA score							
Mean	4.50	2.81	4.56	1.75	1.13	4.31	5.94
Median	5.00	3.00	5.50	1.50	1.00	5.00	6.00
Std. Dev.	0.73	0.40	1.63	0.86	0.72	1.08	0.25
Min	3.00	2.00	2.00	1.00	0.00	1.00	5.00
Max	5.00	3.00	6.00	3.00	2.00	5.00	6.00
Supervised score							
Mean	3.03	2.77	4.09	1.45	0.72	3.44	5.38
Median	4.00	3.00	5.00	1.00	1.00	4.00	6.00
Std. Dev.	1.72	0.49	1.85	0.97	0.68	1.46	1.18
Min	0.00	1.00	0.00	0.00	0.00	0.00	1.00
Max	5.00	3.00	6.00	3.00	2.00	5.00	6.00
Watson score							
Mean	3.40	2.40	3.37	1.21	0.63	2.91	3.94
Median	4.00	3.00	4.00	1.00	1.00	3.00	4.00
Std. Dev.	1.75	0.85	1.80	1.02	0.66	1.44	1.61
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max	5.00	3.00	6.00	3.00	2.00	5.00	6.00

Table 4 Measures of reliability

		Watson score	Supervised score
Cronbach alpha coefficient		0.737	0.687
Spearman-brown coefficient		0.769	0.604
Spearman correlations $^{**}p < .01$; $^{*}p < .05$	Supervised score	**0.819****	–
	Paper and Pencil MoCA score	**0.515***	**0.637****

Table 5 Results of the multiple linear regression of the robot subtest scores to the MoCA overall score

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
Supervised subtest scores	0.996	0.993	0.987	2.80
Watson subtest scores	0.984	0.969	0.944	5.93

reported that the robot was too fast and they failed to memorise the words for this reason.

Lower Watson scores are mainly related to speech recognition failures with difficult accents, e.g. the Orientation task, and the technical solutions, e.g. in the case of the Attention task some participants didn't touch long enough the robot head and their answer was not recorded.

3.2 Reliability of the overall scores and their external validation

First, to verify the internal consistency and reliability of the overall Supervised and Watson scores, the alpha coefficient and the SB coefficient were calculated and reported in Table 4. The alpha coefficient shows an acceptable internal consistency for the Supervise score, while it is good for Watson. According to the SB coefficient, both Supervised and Watson show a moderate split-half reliability, moreover, the SB coefficient confirms that the length of the test is correct.

Table 4 presents also the Spearman correlations of the Supervised and Watson scores with respect to the Supervised and the standard "Paper and Pencil" MoCA scores. In both cases, we found a strong significant correlation that confirms the reliability of the robotic test for the cognitive skills assessment.

The multiple linear regression analysis was used to build models to confirm the relation between the robot-administered subtest (predictive variables) and the standard MoCA (dependent variable). Results are in Table 5. The models explain almost entirely the dependent variable's variance, indeed R^2 is equal 0.993 and 0.969 respectively for the Supervised and the Watson, confirming that the tests are directly related as there is a linear function that can derive the MoCA from the robotic test. However, while the standard error of the estimate is just 2.80 for the Supervised scores, in the case of scores calculated by Watson this is more than double at 5.93. However, the standard error for the estimate from the Watson scores was reduced to 1.99 when the analysis considered native speakers only.

Multiple regression was also applied within the scoring modalities to build models that relate the subtests' scores and the overall score. Table 6 reports the regression coefficients, which show a similar order with Orientation and Attention always among the strongest contributors and Language and Abstraction being the weakest. The regression equation is as follows Eq. (1), where x_i are the variables (subtests) and coeffients β_i are reported in Table 6.

$$MoCA\ score = \sum_{i=1}^{7} \beta_i x_i \qquad (1)$$

3.3 Comparison between automated Watson score and Supervised score for the robotic test

The similarities shown by the multiple regression are confirmed by the correlations among the subtests, which are reported in Table 7. All correlations with the same subtests are statistically significant ($p < 0.01$) and above the high-effect threshold. This confirms the positive reliability of the Watson scoring also for the subtests of the single cognitive domains.

To further test the equivalence between the Watson and the Supervised scores, we performed an equivalence test following the procedure suggested by Lakens *et al.*[53]. We set the upper boundary equivalence of smallest effect size of interest (SESOI) to 0 and the lower boundary to −3.40. The upper bound is zero because we would avoid that higher scores can induce false negatives of MCI. Meanwhile we can accept some false positives that can be corrected by the human supervisor; therefore we set the lower bound as the 90% of the confidence interval of our benchmark (the Paper MoCA) as suggested for equivalence tests in Ref. [53]. The two one-sided tests (TOST) procedure indicated that the observed effect size (dz = 0.58) was significantly within the equivalent bounds of −3.4 and 0 scale points, (or in Cohen's dz: −1.1 and 0), $t(34) = -3.07$, p = 0.002.

3.4 A simulated use case scenario

This section presents a simulated use case scenario, where CATHI is used as a screening tool to exclude those who are very likely to not have MCI from being actually tested by a medical practitioner. To this end, in absence of normative data and considering Watson lower scores than MoCA, we used the Watson standard error in estimate (5.93, Table 5), to derive a cut-off threshold for the Watson score at 20, which is 6 points lower than the level considered normal for the MoCA. Applying this threshold, we can potentially exclude 45.71% of participants (16/35) from being evaluated by a professional. Table 8 presents the descriptive statistics of these participants potentially excluded from a human evaluation. The statistics demonstrate that none would have actually required a professional screening, because the minimum score in the Paper MoCA was 23, which represents the lower bound cut-off for this test[45].

The next step is the human supervision, which could have excluded two more participants from being evaluated in person by conservatively applying the same cut-off score of MoCA (23 points). In total 18 (51.43%)

Table 6 Regression variables and coefficients

i	Variables (x_i)	Coefficients (β_i)		
		MoCA	Watson	Supervised
1	Visuospatial/Executive	0.182	0.213	0.163
2	Naming	0.113	0.174	0.129
3	Attention	0.193	0.192	0.213
4	Language	0.077	0.081	0.079
5	Abstraction	0.053	0.046	0.045
6	Delayed recall	0.177	0.157	0.171
7	Orientation	0.237	0.202	0.252

Table 7 Spearman correlations between Supervised and Watson scores (direct correlations in bold)

	$^{**}p < .01$; $^{*}p < .05$	Supervised scores							
		Visuospatial/Executive	Naming	Attention	Language	Abstraction	Delayed recall	Orientation	Overall
Watson scores	Visuospatial/Executive	**0.775****	0.061	0.225	**0.559****	**0.495****	0.320	0.312	**0.677****
	Naming	0.315	**0.576****	**0.430****	0.286	**0.537****	**0.406***	0.195	**0.541****
	Attention	0.188	0.180	**0.847****	0.321	0.259	0.080	0.165	**0.510****
	Language	**0.413***	**0.493****	**0.489****	**0.566****	**0.453****	**0.603****	**0.390***	**0.734****
	Abstraction	**0.507****	**0.489****	0.316	0.458**	0.810**	0.271	0.190	**0.616****
	Delayed recall	0.345	0.247	0.259	**0.352***	**0.421***	**0.641****	0.317	**0.607****
	Orientation	0.157	−0.037	−0.041	−0.020	−0.052	0.073	**0.515****	0.195
	Overall	**0.578****	0.308	**0.563****	**0.486****	**0.608****	**0.505****	**0.454****	**0.819****

Table 8 Descriptive statistics of participants potentially excluded by Watson ($N = 16$, Watson score ≥20)

	Paper MoCA	Supervised score	Watson score
Mean	25.69	25.25	24.13
Std. Dev.	1.58	3.02	2.00
Median	26	27	25
Minimum	23	18	21
Maximum	29	28	27

Table 9 Paired samples *t*-test on the participants potentially excluded from further assessment by Watson or the supervisor ($N = 18$, Watson score ≥ 20 and Supervised score ≥ 23)

Pair	Mean difference	Std. Dev.	Significance (*p*)
Supervised score – Watson score	1.13	2.06	.045
PaperMoCA – WatsonScore	1.56	2.37	.018
PaperMoCA – SupervisedScore	0.44	3.16	.588

participants would avoided going to a clinic and being assessed by a practitioner in person. In fact, if we consider the participants with the higher scores, there is no significant difference ($p > 0.01$) when analysing the differences via the paired samples *t*-test. Table 9 reports the results of the 18 participants potentially excluded with Watson score ≥ 20 and Supervised score ≥ 23.

4 Discussion

The results show that the CATHI system can be a promising tool for assisting the doctors in the screening of cognitive decline. The analysis demonstrates that the system may achieve its objective to significantly reduce the human assessment via the automated scoring and to allow remote evaluation thanks to the audio and video recordings. The results show that the CATHI system is consistent with the benchmark (MoCA) for the participants that performed well and achieved high scores in the Paper and Pencil test (Table 8), while CATHI amplifies the errors and undervalues the score for the low performers. Nonetheless, this limited impact on the application as the primary concern does not produce false negatives, i.e. does not recognise potential cases of MCI.

However, it should be noted that we tested CATHI on participants who are almost all young and very likely to not suffer from any cognitive impairment. Indeed, before the system can be adopted for clinical diagnosis, it needs to be validated in future clinical trials with a larger sample, which should better represent the target population. It should be also proven in comparison with other forms of computerised self-assessment of cognitive functionalities.

An additional topic for discussion is the lower performance of the non-native speakers, whose performance was significantly lower than the natives despite they were all well-educated and proficient in English. This was expected because MoCA localisations can be different because they are not just translated but also adapted to the local culture. Furthermore, being assessed in a secondary language can generate a cognitive overload and have a negative effect on the performance[54]. In fact, the non-natives were included in the experiment to simulate the MCI and allow a wider range of scores. However, the difficulty was amplified in the case of the robotic assessment, causing an unexpected number of extremely low scores. A further negative effect can be seen in the case of Watson speech recognition, which failed to recognise words in the case of participants with strong accents. We are confident that using appropriate cultural and language localisations, the Watson performance can certainly improve.

However, the CATHI system was in almost all cases providing a lower score than the reference MoCA score, which can be also due to the rigid administration procedure, for instance, similar to the MoCA, the robot never repeated the task instructions or cut the answers if these exceeded the time allocated. Human assessors can be more flexible, e.g. repeat or give more time in case of distraction, even if repeating is explicitly forbidden by the test manual. There is also the case of the "draw a clock" task in which the Paper-and-Pencil version indirectly suggests what to draw while the robot doesn't.

Finally, we would remark that similarly to the MoCA, whose applicability as a screening tool has been extended to a variety of psychiatric conditions that can affect also the younger population (see section **2.1**), the CATHI system could be used for screening the MCI related to schizophrenia and depression. Future studies will also address this application.

5 Conclusion

This article presents a novel intelligent robotic service, named CATHI, for cognitive assessment via human-robot interaction with the autonomous scoring system based on the IBM Watson AI cloud services. The analyses of the experimental results show that the CATHI system could autonomously administer the cognitive test and calculate a score with an acceptable reliability. Those cases that do not score the minimum threshold can be referred to a medical practitioner, who could review the recordings and decide whether to prescribe an examination in person with a practitioner. We underline that the system is proposed as a screening and data collection tool, therefore, the automatic Watson score should be always revised and validated by a professional supervisor for clinical use. We envision the robotic administration could favour large-scale screening tests of MCI by providing preliminary remote diagnostic information about patients, therefore, reducing the workload for human doctors and increasing the population that can be screened. The screening will favour early detection of neurological disorders associated with older (e.g. dementia) and younger population (schizophrenia and depression). To this end, a new psychometric instrument should be created and tailored to the artificial intelligence and robotics in order to fully take advantage of the opportunities given by these novel technologies, then its validity should be tested via clinical trials in comparison with alternative computerised solutions.

Acknowledgment

The authors gratefully thank all university staff and students who participated in this study. This work won the IBM Shared University Research (SUR) Award. The work of Daniela Conti and Alessandro Di Nuovo was supported by the European Union's H2020 research and innovation program under the MSCA-Individual Fellowship grant agreement no. 703489 (CARER-AID). Alessandro Di Nuovo was also partially supported by the EPSRC through project grant EP/P030033/1 (NUMBERS).

References

[1] Feil-Seifer D, Matarić M J. Defining socially assistive robotics. *Proceedigns of the* 9*th International Conference on Rehabilitation Robotics*, Chicago, IL, USA, 2005, 465–468.

[2] Tapus A, Mataric M J, Scasselati B. Socially assistive robotics [Grand Challenges of Robotics]. *IEEE Robotics & Automation Magazine*, 2007, **14**, 35–42.

[3] Matarić M J, Scassellati B. Socially assistive robotics. In Siciliano B, Khatib O eds., *Springer Handbook of Robotics*, Cham, 2016, 1973–1994.

[4] Furht B, Escalante A. *Handbook of Cloud Computing*, Springer, New York, NY, USA, 2010.

[5] Hu G, Tay W P, Wen Y. Cloud robotics: Architecture, challenges and applications. *IEEE Network*, 2012, **26**, 21–28.

[6] Kehoe B, Patil S, Abbeel P, Goldberg K. A survey of research on cloud robotics and automation. *IEEE Transactions on Automation Science and Engineering*, 2015, **12**, 398–409.

[7] Novoa J, Wuth J, Escudero J P, Fredes J, Mahu R, Yoma N B. DNN-HMM based automatic speech recognition for HRI scenarios. *Proceedings of the* 2018 *ACM/IEEE International Conference on Human-Robot Interaction*, Chicago, IL, USA, 2018, 150–159.

[8] Di Nuovo A, Broz F, Wang N, Belpaeme T, Cangelosi A, Jones R, Esposito R, Cavallo F, Dario P. The multi-modal interface of Robot-Era multi-robot services tailored for the elderly. *Intelligent Service Robotics*, 2018, **11**, 109–126.

[9] Wang N, Di Nuovo A, Cangelosi A, Jones R. Temporal patterns in multi-modal social interaction between elderly users and service robot. *Interaction Studies*, 2019, **20**, 1–9.

[10] Di Nuovo A, Broz F, Cavallo F, Dario P. New frontiers of service robotics for active and healthy ageing. *International Journal of Social Robotics*, 2016, **8**, 353–354.

[11] Cavallo F, Esposito R, Limosani R, Manzi A, Bevilacqua R, Felici E, Di Nuovo A, Cangelosi A, Lattanzio F, Dario P. Robotic services acceptance in smart environments with older adults: User satisfaction and acceptability study. *Journal of Medical Internet Research*, 2018, **20**, 264.

[12] Iroju O, Ojerinde O, Ikono R. State of the art: A study of human-robot interaction in healthcare. *International Journal of Information Engineering and Electronic Business*, 2017, **3**,

43–55.

[13] Conti D, Di Nuovo S, Trubia G, Buono S, Di Nuovo A. Use of robotics to stimulate imitation in children with autism spectrum disorder: A pilot study in a clinical setting. *Proceedings of the* 24*th IEEE International Symposium on Robot and Human Interactive Communication*, *ROMAN*, Kobe, Japan, 2015, 1–6.

[14] Conti D, Di Nuovo S, Buono S, Di Nuovo A. Robots in education and care of children with developmental disabilities: A study on acceptance by experienced and future professionals. *International Journal of Social Robotics*, 2017, **9**, 51–62.

[15] Conti D, Cirasa C, Di Nuovo S, Di Nuovo A. "Robot, tell me a tale!": A social robot as tool for teachers in kindergarten. *Interaction Studies*, 2019, **20**, 1–16.

[16] Rabbitt S M, Kazdin A E, Scassellati B. Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use. *Clinical Psychology Review*, 2015, **35**, 35–46.

[17] Boumans R, van Meulen F, Hindriks K, Neerincx M, Olde Rikkert M. Proof of concept of a cocial robot for patient reported outcome measurements in elderly persons. *Companion of the* 2018 *ACM*/*IEEE International Conference on Human-Robot Interaction*, Chicago, IL, USA, 2018, 73–74.

[18] Conti D, Trubia G, Buono S, Di Nuovo S, Di Nuovo A. Evaluation of a robot-assisted therapy for children with autism and intellectual disability. *Proceedings of* 19*th Annual Conference on Towards Autonomous Robotic Systems*, Bristol, UK, 2018, 405–415.

[19] Di Nuovo A, Conti D, Trubia G, Buono S, Di Nuovo S. Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability. *Robotics*, 2018, **7**, https://doi.org/10.3390/robotics7020025.

[20] Petric F, Miklic D, Kovacic Z. Robot-assisted autism spectrum disorder diagnostics using POMDPs. *Proceedings* of *Companion of the* 2017 *ACM*/*IEEE International Conference on Human-Robot Interaction*, Vienna, Austria, 2017, 369–370.

[21] Wijayasinghe I B, Ranatunga I, Balakrishnan N, Bugnariu, N, Popa D O. Human-robot gesture analysis for objective assessment of autism spectrum disorder. *International Journal of Social Robotics*, 2016, **8**, 695–707.

[22] Kojima H, Takaeda K, Nihel M, Sadohara K, Ohnaka S, Inoue T. Acquisition and evaluation of a human-robot elderly spoken dialog corpus for developing computerized cognitive assessment systems. *Journal of the Acoustical Society of America*, 2016, **140**, 2963–2963.

[23] Varrasi S, Di Nuovo S, Conti D, Di Nuovo A. A social robot for cognitive assessment. *Proceedings of Companion of the* 2018 *ACM*/*IEEE International Conference on Human-Robot Interaction*, Chicago, IL, USA, 2018, 269–270.

[24] Feingold Polak R, Elishay A, Shahar Y, Stein M, Edan Y, Levy-Tzedek S. Differences between young and old users when interacting with a humanoid robot: A qualitative usability study. *Proceedings of Companion of the* 2018 *ACM*/*IEEE International Conference on Human-Robot Interaction*, Chicago, IL, USA, 2018, 107–108.

[25] Scassellati B, Admoni H, Matarić M. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 2012, **14**, 275–294.

[26] Desideri L, Ottaviani C, Malavasi M, di Marzio R, Bonifacci, P. Emotional processes in human-robot interaction during brief cognitive testing. *Computers in Human Behavior*, 2019, **90**, 331–342.

[27] Rossi S, Santangelo G, Staffa M, Varrasi S, Conti D, Di Nuovo A. Psychometric evaluation supported by a social robot: Personality factors and technology acceptance. *Proceedings of the* 27*th IEEE International Symposium on Robot and Human Interactive Communication* (*RO-MAN*), Nanjing, China, 2018, 802–807.

[28] Petersen R C. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 2004, **256**, 183–194.

[29] Luis C, Loewenstein D, Acevedo A, Barker W W, Duara R. Mild cognitive impairment: Directions for future research. *Neurology*, 2003, **61**, 438–444.

[30] Landau S M, Harvey D, Madison C M, Reiman E M, Foster N L, Aisen P S, Petersen R C, Shaw L M, Trojanowski J Q, Jack C R, Weiner M W, Jagust W J. Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, 2010, **75**, 230–238.

[31] Di Nuovo S, De La Cruz V M, Conti D, Buono S, Di Nuovo A. Mental imagery: Rehabilitation through simulation. *Life Span and Disability*, 2014, **17**, 89–118.

[32] Caraci F, Castellano S, Salomone S, Drago F, Bosco P, Di Nuovo S. Searching for disease-modifying drugs in AD: Can we combine neuropsychological tools with biological markers? *CNS & Neurological Disorders – Drug Targets*, 2014, **13**, 173–186.

[33] Fisekovic S, Memic A, Pasalic A. Correlation between MoCA and MMSE for the assessment of cognition in schizophrenia. *Acta Informatica Medica*, 2012, **20**,

186–189.

[34] Moirand R, Galvao F, Lecompte M, Poulet E, Haesebaert F, Brunelin J. Usefulness of the Montreal Cognitive Assessment (MoCA) to monitor cognitive impairments in depressed patients receiving electroconvulsive therapy. *Psychiatry Research*, 2018, **259**, 476–481.

[35] Pandey A K, Gelin R A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, 2018, **25**, 40–48.

[36] Varrasi S, Di Nuovo S, Conti D, Di Nuovo A. Social robots as psychometric tools for cognitive assessment: A pilot test. *Proceedings* of 10*th International Workshop on Human Friendly Robotics*, Naples, Italy, 2019, 99–112.

[37] Varrasi S, Lucas A, Soranzo A, McNamara J, Di Nuovo A. IBM cloud services enhance automatic cognitive assessment via human-robot interaction. In Carbone G, Ceccarelli M, Pisla D eds., *New Trends in Medical and Service Robotics*, Cassino, Italy, 2019, 169–176.

[38] Nasreddine Z S, Phillips N A, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings J L, Chertkow H. The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 2005, **53**, 695–699.

[39] Nasreddine Z. MoCA Montreal Cognitive Assessment, [2018-10-18], www.mocatest.org

[40] Olson R A, Chhanabhai T, McKenzie M. Feasibility study of the Montreal Cognitive Assessment (MoCA) in patients with brain metastases. *Supportive Care in Cancer*, 2008, **16**, 1273–1278.

[41] Wong A, Xiong Y Y, Kwan P W L, Chan A Y Y, Lam W W M, Wang K, Chu W C W, Nyenhuis D L, Nasreddine Z, Wong L K S, Mok V C T. The validity, reliability and clinical utility of the Hong Kong Montreal Cognitive Assessment (HK-MoCA) in patients with cerebral small vessel disease. *Dementia and Geriatric Cognitive Disorders*, 2009, **28**, 81–87.

[42] Dalrymple-Alford J C, MacAskill M R, Nakas C T, Livingston L, Graham C, Crucian G P, Melzer T R, Kirwan, J, Keenan R, Wells S, Porter R J, Watts R, Anderson T J. The MoCA Well-suited screen for cognitive impairment in Parkinson disease. *Neurology*, 2010, **75**, 1717–1725.

[43] Videnovic A, Bernard B, Fan W, Jaglin J, Leurgans S, Shannon K M. The Montreal Cognitive Assessment as a screening tool for cognitive dysfunction in Huntington's disease. *Movement Disorders*, 2010, **25**, 401–404.

[44] Bertrand J-A, Génier Marchand D, Postuma R B, Gagnon J-F. Cognitive dysfunction in rapid eye movement sleep behavior disorder. *Sleep and Biological Rhythms*, 2013, **11**, 21–26.

[45] Carson N, Leach L, Murphy K J. A re-examination of Montreal Cognitive Assessment (MoCA) cutoff scores. *International Journal of Geriatric Psychiatry*, 2017, **33**, 379–388.

[46] Chertkow H, Nasreddine Z S, Johns E, Phillips N A, McHenry C. The Montreal Cognitive Assessment (MoCA): Validation of alternate forms and new recommendations for education corrections. *Alzheimer's and Dementia*, 2011, **7**, S157.

[47] Xu T (Linger), Zhang H, Yu C. See you see me: The role of eye contact in multimodal human-robot Interaction. *ACM Transactions on Interactive Intelligent Systems*, 2016, **6**, 1–22.

[48] Sciutti A, Rea F, Sandini G. When you are young, (robot's) looks matter. Developmental changes in the desired properties of a robot friend. *Proceedigns of the* 23*rd IEEE International Symposium on Robot and Human Interactive Communication* (*IEEE RO-MAN*), Edinburgh, Scotland, UK, 2014, 567–573.

[49] Streiner D L. Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 2003, **80**, 99–103.

[50] Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA, 1988.

[51] Landau S, Everitt B S. *A Handbook of Statistical Analyses Using SPSS*, CRC Press, Boca Raton, FL, USA, 2004.

[52] Field A. *Discovering Statistics Using IBM SPSS Statistics*, SAGE Publications, Los Angeles, CA, USA, 2013.

[53] Lakens D, Scheel A M, Isager P M. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2018, **1**, 259–269.

[54] Purpura J E. An analysis of the relationships between Ttest takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning*, 1997, **47**, 289–325.