Intelligent Service Robotics (2024) 17:303–313
https://doi.org/10.1007/s11370-023-00505-2

ORIGINAL RESEARCH PAPER

Robot-assisted language assessment: development and evaluation of feasibility and usability

Sukyung Seok[1,2] · Sujin Choi[3] · Kimun Kim[1] · Jongsuk Choi[1] · Jee Eun Sung[3] · Yoonseob Lim[1]

Received: 19 December 2022 / Accepted: 6 December 2023 / Published online: 22 January 2024

Abstract

Many studies have shown that robots can provide medical help to patients, such as supporting physical movements, managing mood, or simulating cognitive function. However, robotic cognitive/language assessment, which is vital for mental health care, has not been fully explored and is limited to only a few types of assessment. The aim of this study is to present and evaluate a social robot equipped with a web-based language assessment for sentence comprehension test (SCT) with a dialogue system involving yes/no questions. A total of 50 participants took the test with 36 items conducted by a robot (robot-SCT), while a total of 55 participants took the same test but conducted by a human examiner (human-SCT). Comparative analyses were performed to evaluate the validity of the robot-SCT in terms of test scores and time-related measures. Usability was evaluated through the system usability score and interview feedback. With regard to the validity of the robot-SCT, the test scores indicated no significant differences between the robot-SCT and human-SCT. In addition, conditional differences in reaction time for the test items were observed, similar to the previous paper-and-pencil researches. The high system usability scores (i.e., mean = 78.5, SD = 11) demonstrated the high usability of the robot-SCT. This study demonstrates the validity and usability of robotic language assessment among normal adults. However, further evaluation is required for people with dementia or mild cognitive impairment.

Keywords Sentence comprehension test (SCT) · Robotic cognitive/language assessment · Service robotics · Human–robot interaction (HRI)

1 Introduction

As people age, cognitive abilities such as executive functions, memory, and language will decline. Distinguishing or detecting the changes in cognitive abilities during the early stages of a disease becomes important as it not only influences the lives of patients and family members but also has severe societal and economic impacts [1, 2]. Several cognitive assessments have been established to ensure that psychiatrists can evaluate cognitive impairment, diagnose the possible cause of the change in cognitive abilities, and monitor the progress of the disease. These methods have been proven to have reliable psychometric rigor, but certain subgroups of patients may have limited access to the tests. A study showed that simple cognitive screening tools could be partially incomplete due to stroke-related impairments or symptoms [3]. Furthermore, changes in cognitive functions are not easy to notice until daily activities become severely disrupted [4], hindering the utilization of cognitive tests in clinics or primary care settings.

The most common methods for evaluating the cognitive functions of people with mild cognitive impairment (MCI) and dementia (or stroke victims) include paper-and-pencil tests and interviews [5–8]. The mini-mental state examination (MMSE) is widely used to screen for dementia severity, although it is not sufficiently accurate for detecting early-stage cognitive changes, especially in MCI [9]. To address this issue, the Montreal Cognitive Assessment (MoCA) test

Sukyung Seok, Sujin Choi and Kimun Kim have contributed equally.

✉ Jee Eun Sung
jeesung@ewha.ac.kr

✉ Yoonseob Lim
yslim@kist.re.kr

[1] Center for Intelligent and Interactive Robotics, Korea Institute of Science and Technology, Seoul, Republic of Korea

[2] Department of Korean Language and Literature, Korea University, Seoul, Republic of Korea

[3] Department of Communication Disorders, Ewha Womans University, Seoul, Republic of Korea

was developed as a brief cognitive screening tool, which is widely used in clinical and research areas [10]. However, MoCA is not recommended for individuals with limited or poor abilities in reading or writing in Korea. Additionally, there are limitations to consider due to variations in examiners' characteristics, such as voice, speaking speed, questioning styles, and expressiveness, which can influence the validity of assessment results [11]. To overcome the limitations of the short version of cognitive assessment tools, Korean clinical practice includes the Seoul Neuropsychological Screening Battery (SNSB) tool [12], which assesses cognitive abilities across various domains, including the MMSE. The SNSB sensitively evaluates early cognitive impairment (such as MCI) and allows for the assessment of the severity of dementia and the analysis of patterns of cognitive decline to estimate underlying causes [13]. However, due to the inclusion of numerous items to provide comprehensive information across different cognitive domains, the SNSB has a longer administration time of almost 2 h and requires expertise from the examiner, making it challenging to apply as a robot module.

The sentence comprehension test (SCT) [14] is a cognitive assessment tool that specifically evaluates deficits in sentence-level processing, which can be observed in the early stages of cognitive decline. This test serves as both an aging effect indicator and a screening tool for MCI. The SCT manipulates sentence types and word order, with a focus on reflecting the characteristics of Korean as a free word order language. For example, the typical canonical word order in Korean is subject–object–verb, while a non-canonical word order refers to a change in the position of the object and subject. The SCT performance of the elderly population indicates a decline in non-canonical word order items compared to canonical word order items on the SCT [14]. The results of a study by Sung and colleagues using the SCT on a population of people who were either aging normally or had MCI confirmed that people with MCI exhibit different patterns from those who were aging normally, particularly in tasks involving increased syntactic complexity [15]. Moreover, the researchers identified linguistic markers in the SCT that could be used to detect MCI. Therefore, as a cognitive assessment tool that reflects the characteristics of the Korean language, the SCT is suitable for application as a robot module due to its relatively short administration time, limited number of test items, and simple testing procedures.

To facilitate the delivery of cognitive assessments, computerized test batteries and cognitive tests running on mobile devices such as smartphones and tablets have been proposed [16–18]. For example, a study exploring the validity of tablet-based administration of the Brief Assessment of Cognition in Schizophrenia (BACS) converted the paper-and-pencil-based cognitive assessment tool to a tablet-based version of the app and showed that the tablet-based assessment can achieve results that are consistent with the traditional paper-and-pencil-based BACS. However, participants often have a hard time maintaining their attention on the questionnaires as the tasks can be boring and repetitive, which may affect the quality of the collected data [19]. It has also been shown that participants may simply stop using the online program and drop out before the completion of the test [20]. To successfully involve participants in cognitive assessment and training, game design elements have been imported into cognitive tasks. A study introduced a non-immersive virtual reality cognitive assessment for the Aphasia App and reported its preliminary evidence on the validity of assessing non-linguistic task performance by comparing the performance with paper-and-pencil tasks [21]. Other studies had developed spoken dialogue systems to detect dementia through verbal interaction with virtual agents [22, 23]. By analyzing acoustical and linguistic features of speech data obtained based on structured dialogues with predetermined questions, these studies showed the potential of automated detection of dementia through spoken dialog through interaction with virtual avatars. However, the dialogues and exchanges are abrupt and unnatural, with long pauses or keyboard strokes used to indicate turn-taking to a virtual agent.

Meanwhile, robots for healthcare, often called assistive robots or socially assistive robots, have been utilized in unique ways to provide physical or psychological help to the sick and the elderly [24]. Due to its physical appearance and movable components, robots have been widely used to provide physical support for surgeons to perform surgery, caregivers to transfer their patients, patients to move their limbs, and so on [25–27]. Equipped with diverse human–robot interaction skills, such as verbal communication and gestures, robots have also been used to mentally assist people with cognitive impairments [28–30]. With regard to cognitive assessments, Brief Cognitive Testing (BCT) and the MoCA have been studied to demonstrate the validity and usefulness of their application through robots by comparing the assessment performance of robots with that of human examiners [31, 32]. However, apart from these two assessments, it is difficult to find examples of robot applications in the context of cognitive assessment, and more attempts are required to reveal the validity and utility of robot-based cognitive assessment.

All in all, cognitive assessments in various forms such as mobile device applications, virtual agents, and robots have been studied and their validity has been proved by comparing their performance with traditional paper-and-pencil-based tests. It has been suggested that the physical presence of robots plays a more crucial role in social interaction with agents than the physical embodiment of virtual agents [33]. Therefore, robots can be more beneficial to cognitive assessment than virtual agents or mobile devices.

However, relatively few studies have explored the validity of robot-based cognitive tests such as MoCA and BCT [31, 32]. To the best of our knowledge, the validity of language assessment through verbal interaction with a robot has been less studied.

In this study, we employed a validated cognitive assessment (SCT) on a robot to test the validity and usability of a language assessment through verbal interaction with a robot in healthy adults, with the aim to expand the range of cognitive assessments that robot can perform. Specifically, we adapted the sentence-picture matching paradigm of SCT, where the robot presented a sentence verbally and presented a picture on a display, then the user responded "yes" or "no" verbally to indicate whether the sentence and picture matched. The robot-SCT system was designed to recognize the user's voice and conducted the test automatically. Additionally, the system recorded the test score and response time automatically. We evaluated the validity of the robot-SCT system through the SCT score and response time, and the usability through the System Usability Scale (SUS) score, which is a user's perceptual evaluation tool for system usability. By doing this, we aimed to demonstrate the potential of our robot-SCT system, contributing to the field of robot cognitive assessment.

2 Hypotheses

Based on related works on SCT and robots in cognitive assessment, we constructed three hypotheses for our experiments. Hypotheses 1 (HI) and 2 (H2) are related to the validity, while the hypothesis 3 (H3) is related to the usability of robot-SCT.

H1. Concurrent Validity: The robot-SCT is concurrently valid with the human-SCT.
H2. Linguistic Validity: response time from the robot-SCT can reliably elicit linguistic complexity.
H3. Robot-SCT Usability: SUS scores are analyzed to evaluate the usability of the robot-SCT.

3 Methods

3.1 Online SCT system for robots

The robotic platform used in this study was the humanoid-like robot developed by RoboCare [34], which is a prototype robot designed for social human–robot interaction studies that is equipped with various interfaces: speech recognition and production, object and face recognition, facial expression, human-like gestures with 6-DOF arm and 2-DOF waist, and a touch screen (Fig. 1a).

We implemented the robot to lead the administration of the SCT. The robot gives instructions for the SCT and records the interaction data based on the verbal response of the user. The entire process is automated and controlled via the robotic operating system (ROS). The overall framework is shown in Fig. 1b. The verbal response of the user can be detected by the microphone installed on the left shoulder of the robot or via a wireless headset. The obtained speech signal is subsequently sent to a web server for speech-to-text (STT) speech recognition via the data manager, which addresses the processing of incoming raw sensory data and recognition data. Dialogflow handles the complete verbal interaction process with the user, evaluating the accuracy of the user's response at each step and selecting suitable robot gestures and utterances that match the relevant questions for the SCT. The behavior manager creates a series of body movements and coordinates them with the robot's speech, using task information obtained from the data manager. The robot can make emotional expressions through 40 different body gestures and 20 different eye expressions. For example, the robot can wave its hand to greet the user or make a questioning gesture by lifting both hands when the user gives a wrong response.

We converted the previous paper-and-pencil-based SCT into a web-based assessment test, which can be accessed by the user through commercial web browsers or via web-socket-based communication with the robot platform (dashed rectangle shown in Fig. 1b). Unlike paper-and-pencil-based tests, in which the assessment data are collected manually through an examiner, the web-based SCT module can recognize the voice of the user by using the speech recognition of the Web Speech API [35] and automatically measures the test score and the response time (RT) of the user's verbal response. The assessment data, including individual verbal response and response delay, are automatically stored in a real-time database (Google Firebase). Additionally, a speech synthesis module was also implemented to ensure that the assessment can be independently taken through mobile devices if the robot is not available.

3.2 Participants

This study was conducted with the approval of the Institutional Review Board of the Korea Institute of Science and Technology (KIST IRB No. 2021–024). A total of 105 native Korean speakers in their 20 s and 30 s participated in this study. All participants did not report any subjective memory impairments in the Subjective Memory Complaints Questionnaire [36], and their results of the Korean Mini Mental State Examination (K-MMSE) [37] were within the normal range (> 16%ile).

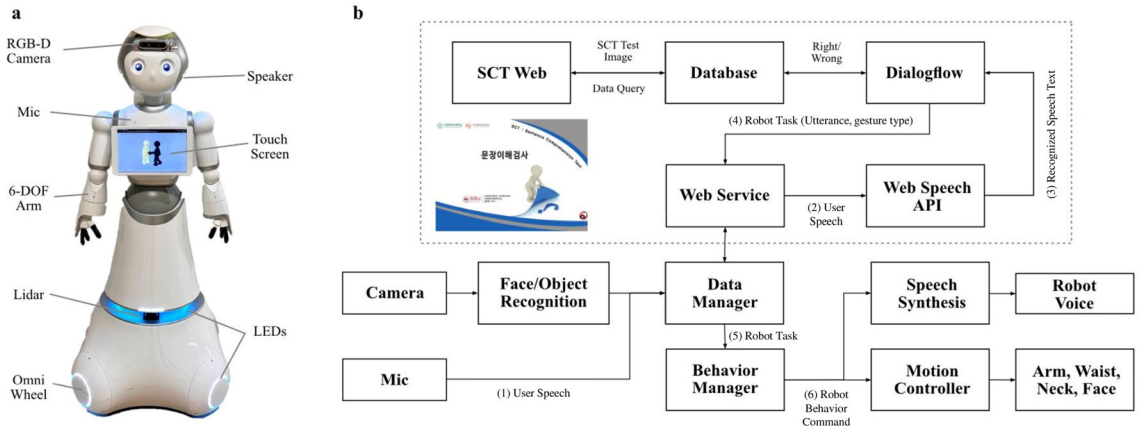

Fig. 1 **a** Robot platform and **b** the architecture of the whole system used in this study

Table 1 Descriptive information by groups

	Human-SCT ($N = 55$)	Robot-SCT ($N = 50$)	P(Human-SCT vs. Robot-SCT)
Age (mean (SD))	26.78 (1.82)	27.16 (3.08)	0.44[a]
Years of education (mean (SD))	17.15 (1.06)	17.24 (1.89)	0.75[a]
K-MMSE score (mean (SD))	29.36 (0.93)	29.66 (0.62)	0.06[a]
Gender (M:F)	14:41	11:39	0.68[b]

[a]One way ANOVA
[b]Pearson chi-square test

The participants were divided into two groups (human-SCT and robot-SCT) according to the presentation type of the SCT. Of these participants, 55 participants (41 women, 14 men; $M_{age} = 26.78$, $SD_{age} = 1.82$) were assigned to the human-SCT group and the other 50 participants (39 women, 11 men; $M_{age} = 26.15$, $SD_{age} = 3.57$) were assigned to the robot-SCT group. There was no statistically significant difference in age, $F_{(1,104)} = 0.597$, $P = 0.66$, or years of education, $F_{(1,104)} = 0.102$, $P = 0.33$, between the two groups. Table 1 provides the descriptive information about both groups.

3.3 Experimental stimuli and measurements

The SCT performed in this experiment was carried out by converting the version from [14] to a sentence validation task based on voice recognition. For the picture stimuli of the SCT, a sentence-picture paradigm with humanized pictograms was employed using three-color words. The SCT was constructed based on the freedom of word order in Korean. Although there is a canonical word order of 'Subject (S) + Object (O) + Verb(V),' non-canonical word order of 'OSV' is also allowed. The SCT was further subdivided into three sentence types: (1) active sentences with 2-place verbs (A2), (2) active sentences with 3-place verbs (A3), and (3) passive sentences (P). In summary, the SCT has a total of six syntactic structures according to the canonicity, canonical word order sentences (C), non-canonical word order sentences (NC), and sentence types (A2, A3, P), and each structure consists of six items for a total of 36 items. The examples of each item are presented in Table 2.

In this study, participants were asked to listen to the sentence and answer 'yes' if the sentence matched the picture; otherwise, they should answer 'no.' The experimental stimuli consisted of three steps as follows: (1) color-blindness test, (2) action verb learning, (3) practice trials, and (4) main trials. There were a total of 36 questions, where one point was given for each correct response for a total of 36 possible points. An example of the SCT is provided in Fig. 2.

For the subjective evaluation of the robot-SCT, we utilized the SUS, which is composed of ten statements rated on a 5-point Likert scale for a user's subjective rating of a system's usability [38]. We modified the phrase "the (this) system" in the original SUS statements to "the robot-SCT" to accurately indicate the evaluation target. For example, the first original SUS statement, "I think that I would like to use this system frequently," was changed to "I think that I would like to use the robot-SCT frequently." We calculated the final SUS score by following the calculation method in [38]. In addition, we conducted interviews on the robot-SCT to obtain more detailed feedback.

Table 2 Examples of target sentence of the SCT

Type		Examples			
A2[a]	Target	The yellow pushes the blue			
	C[d]	Nolangi-ka The Yellow-NOM[f]	Palangi-lul The Blue-ACC[g]	Mil-ta pushes-ACT[i]	
	NC[e]	Palangi-lul The Blue-ACC	Nolangi-ka The Yellow-NOM	Mil-ta pushes-ACT	
A3[b]	Target	The blue gives a box to the black			
	C	Palangi-ka The Blue-NOM	Kemcengi-eykey The Black-OBL[h]	Sangca-lul a box-ACC	Cwu-ta gives-ACT
	NC	Kemcengi-eykey The Black-OBL	Palangi-ka The Blue-NOM	Sangca-lul a box-ACC	Cwu-ta gives-ACT
P[c]	Target	The blue is pushed by the yellow			
	C	Palangi-ka The Blue-NOM	Nolangi-eykey The Yellow-OBL	Mil-li-ta is pushes-PASS[j]	
	NC	Nolangi-eykey The Yellow-OBL	Palangi-ka The Blue-NOM	Mil-li-ta is pushes-PASS	

[a]Active sentences with 2-place verb
[b]Active sentences with 3-place verb
[c]Passive sentences
[d]Canonical word order sentences
[e]Non-canonical word order sentences
[f]Nominative case marker
[g]Accusative case marker
[h]Oblique case marker
[i]Active verb
[j]Passive verb

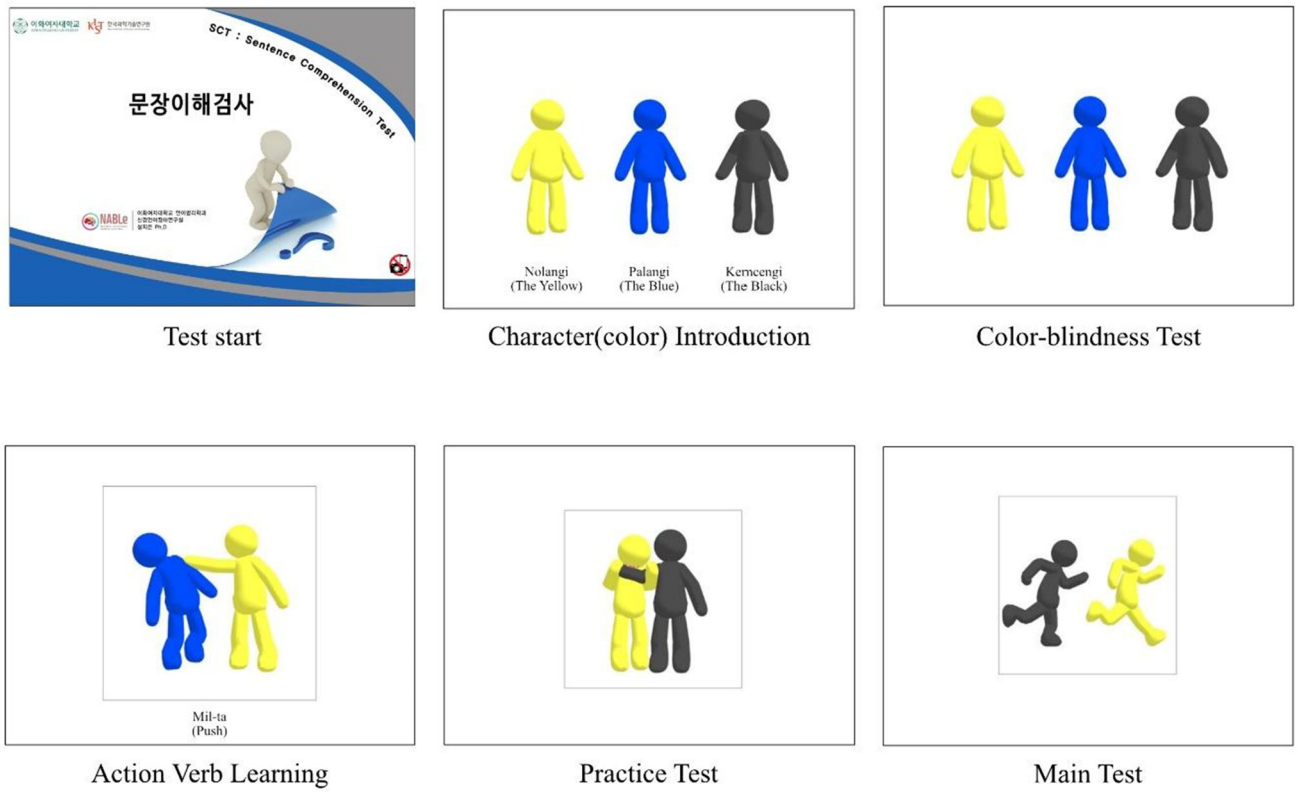

Fig. 2 Example screenshots of web-based SCT

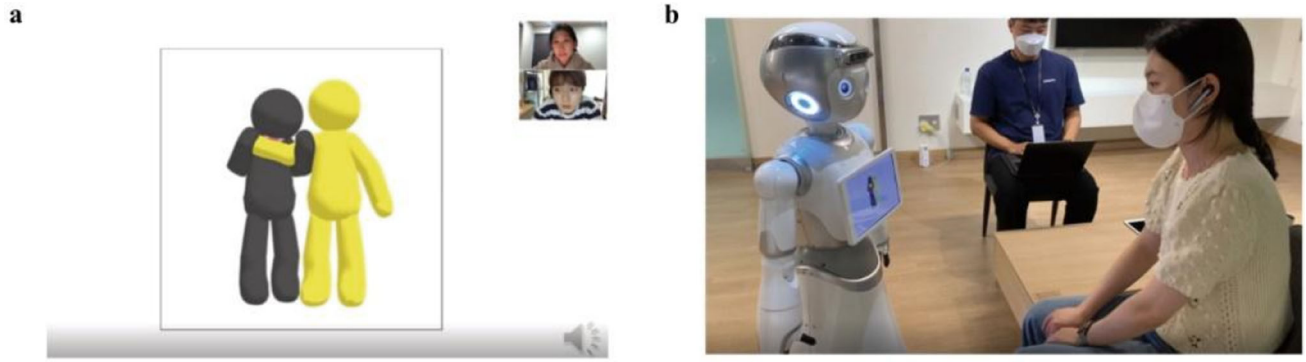

Fig. 3 Experimental setup. **a** Example of human-SCT group, **b** example of robot-SCT group (Supplementary Video)

3.4 Procedures

The participants were divided into two groups (human-SCT and robot-SCT) according to the type of examiner (see Fig. 3). Based on previous research that indicated no significant differences in language assessment results between face-to-face and web-based remote methods [39–41], we conducted a human-SCT using the web-based method. The human-SCT group participated in the SCT using the Zoom application, a web-based video conferencing tool that allows for one-on-one meetings with human examiners. A human examiner presented the test stimuli using the screen-sharing function of the Zoom application, and all responses of the participants were video-recorded.

For the robot-SCT group, the participants physically met with the robot and the human experimenter. They were brought into the room, where they sat facing the robot with the human experimenter by their sides. The participants responded verbally while watching the visual stimuli presented through a monitor at the robot's chest and listening to the audio stimuli from the headset (the same visual and audio stimuli were used for the human-SCT group). Although the user's verbal responses can be acquired through the microphone installed on the left shoulder of the robot, participants had to use the headset due to the high noise level in the environment (refer to the supplementary video for the robot-SCT experiment conducted in a quieter area). The human experimenter remained with the participants and the robot and observed the entire procedure to assist participants if they had any questions or encountered technical difficulties.

The overall procedures for the human-SCT and robot-SCT groups were designed to be consistent in all aspects except for the testing module. During the pre-test or practice session, the participants were instructed to indicate their response as either "yes" or "no" based on the presented stimuli. In the human-SCT condition, the human examiner would respond with a smile for correct answers and with rolled eyes and raised palms for incorrect answers. Similarly, in the robot-SCT condition, the robot would also display these non-verbal cues. Nonetheless, during the main exam, neither human nor robot examiners provided any feedback.

The objective of the pre-test was to familiarize the participants with the SCT procedure and provide them with an opportunity to practice their responses. The non-verbal feedback (i.e., a smile for correct answers and palms raised for incorrect answers) helped the participants to understand the expected responses and reinforced the learning process. This feedback also provided immediate reinforcement and assisted the participants in comprehending the correct and incorrect aspects of their answers. In contrast, the objective of the main test was to evaluate the performance of the participants without any feedback or external cues. By removing feedback from the main test, the focus was solely on evaluating the participants' ability to respond accurately without any external reinforcement.

3.5 Statistical analysis

The statistical analysis was performed using the PASW statistics package version 26 (SPSS Inc.), with the statistical significance set at an alpha level of 0.05.

H1. Concurrent Validity: The robot-SCT is concurrently valid with the human-SCT.

The SCT performance data were analyzed using a three-way mixed analysis of variance (ANOVA) with the canonicity (C, NC) and sentence types (A2, A3, P) as the within-subject factors and the group (human vs. robot) as the between-subject factor. The scores of the SCT were set as the dependent variable.

H2. Linguistic Validity: Response time from the robot-SCT can reliably elicit linguistic complexity.

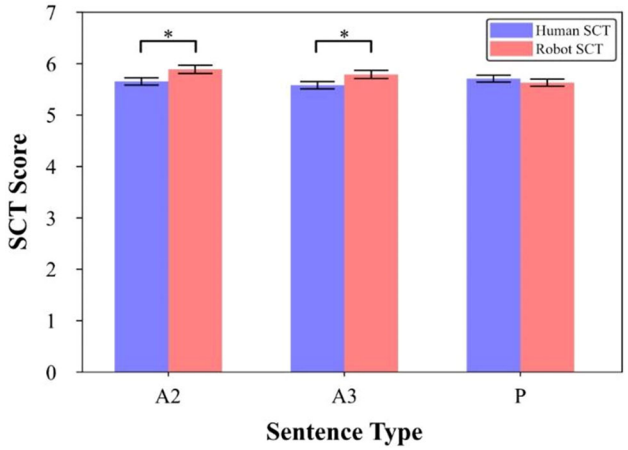

Fig. 4 Result of post hoc analysis on the two-way interaction of the SCT performance

To verify the significance of the response time (RT) according to the canonicity and sentence types in the robot group, a two-way repeated ANOVA was performed. The dependent variable was set as the RT, while the independent variables were the canonicity (C vs. NC) and the sentence types (A2, A3, P).

H3. Robot-SCT Usability: The SUS scores were analyzed to evaluate the usability of the robot-SCT.

The process of determining the SUS score involved summing the score contributions of each item following [38]. Each item's contribution was determined based on its position on the scale, which ranges from 0 to 4. For the positively worded items (1, 3, 5, 7, and 9), the contribution was equal to the scale position minus 1. In contrast, for the reversed items (2, 4, 6, 8, and 10), the contribution was calculated as 5 minus the scale position. To obtain the final SUS value, the sum of all the scores was multiplied by 2.5.

4 Results

4.1 Concurrent validity: the robot-SCT is concurrently valid with the human-SCT

The results of the three-way mixed ANOVA (group, canonicity, sentence types) revealed that the main effect of the canonicity was significant, with the score of the canonical sentences (mean = 5.88, SD = 0.03) being higher than the non-canonical sentences (mean = 5.53, SD = 0.06),$F_{(1, 103)} = 37.058$, $P < 0.001$. The sentence types also had a significant main effect with $F_{(2,206)} = 3.058$, $P = 0.049$, where the score of A2 (mean = 5.77, SD = 0.05) was higher than that of A3 (mean = 5.68, SD = 0.05) and P (mean = 5.67, SD = 0.04), but there was no significant difference between A2 and P. On the other hand, the group had no significant main effect, $F_{(1, 103)} = 2.034$, $P = 0.15$.

The two-way interaction between group and sentence type was statistically significant, $F_{(2,206)} = 7.612$, $P < 0.001$. The results of the post hoc analysis of the one-way ANOVA showed that for A2 and A3 types, the score of the robot-SCT group was significantly higher than that of the human-SCT group, as shown in Fig. 4. The other two-way and three-way interactions were not statistically significant. Table 3 provides the descriptive statistics of the SCT scores for each group.

4.2 Linguistic validity: response time from the robot-SCT can reliably elicit linguistic complexity.

To ensure linguistic validity in robot-SCT, we performed an analysis of RT based on the syntactic structure within the robot-SCT group. The results of the two-way repeated ANOVA revealed that canonicity had a significant main effect ($F_{(1, 75.538)} = 37.656$, $P < 0.001$) and the RT (ms) was shorter in the canonical sentences (mean = 1000.93, SD = 37.85) than in the non-canonical sentences (mean = 1221.21, SD = 42.82). However, the sentence type ($F_{(2, 92)} = 2.320$, $P =$

Table 3 Descriptive statistics of performance on SCT for each group

		A2[a] Mean (SD)	A3[b] Mean (SD)	P[c] Mean (SD)
Human-SCT	C[d]	5.85 (0.55)	5.80 (0.62)	5.87 (0.51)
	NC[e]	5.45 (0.97)	5.36 (0.86)	5.55 (0.89)
Robot-SCT	C[d]	5.98 (0.14)	5.96 (0.19)	5.86 (0.35)
	NC[e]	5.80 (0.49)	5.62 (0.63)	5.40 (0.78)

[a]Active sentences with 2-place verb
[b]Active sentences with 3-place verb
[c]Passive sentences
[d]Canonical word order sentences
[e]Non-canonical word order sentences

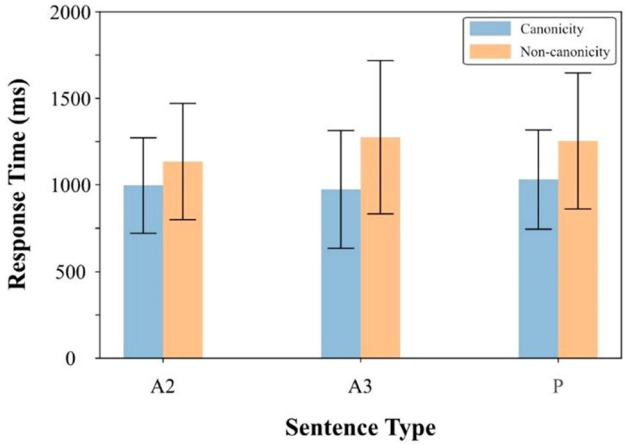

Fig. 5 Result of response time on the two-way repeated ANOVA for robot-SCT Group

0.10) and two-way interaction between canonicity and sentence types ($F_{(2, 92)} = 2.345$, $P = 0.10$) had no significant main effect, as shown in Fig. 5.

4.3 Robot-SCT usability: SUS scores are analyzed to evaluate the usability of the robot-SCT

4.3.1 SUS score on the robot-SCT

The average SUS score was 78.5 (SD = 11), surpassing the threshold of 68 suggested in [42]. According to Bangor and colleagues, systems with scores above 68 are considered acceptable, while those with scores below 51.6 are deemed unacceptable.

4.3.2 Interview feedback on the robot-SCT

A total of 36 types of feedback were received through the interviews. Among the robot-related feedback, the most frequent (25%) was about the robot's voice. In particular, the participants highlighted prosodic factors, such as flat intonation, slow speech rate, and unnatural rhythm. The remainder of the feedback was about the robot's movement and appearance, voice recognition, and screen position, which were only mentioned by one or two participants. Voice recognition was not perceived as a major problem because there was a low voice recognition error of 0.9% for all items across all participants.

5 Discussion

5.1 Principal findings

This study presents a language assessment conducted by a social robot through verbal human–robot interaction with healthy adults. A web-based sentence comprehension test was implemented on a robot platform so that the robot itself can manage the process of the assessment and automatically collect the verbal response from the user. In this study, concurrent validity was assessed by comparing the SCT scores between the robot- and human–SCT groups. Additionally, an analysis of RT based on syntactic complexity was conducted to verify the linguistic validity of the robot-SCT. Furthermore, system usability was evaluated through subjective measures and interviews.

The SCT score results show that the robot-SCT can be applicable for young adults without cognitive impairment when compared to the human-SCT (Fig. 4). Similar results have been reported for another cognitive assessment, namely BCT [3]. Desideri and colleagues compared the results of two BCT tasks, a recall and a subtraction, conducted by a robot and an expert clinician, where it was found that the task accuracy was equivalent regardless of the conductor [31]. Meanwhile, Di Nuovo and colleagues, who presented the robotic system for MoCA, reported relatively lower scores in the robot-based assessment than the paper-and-pencil-based assessment for some subtests, where speech recognition errors due to different accents of participants were mentioned as one of the main causes [32]. Some of MoCA's subtests require participants to say the date and place, names of animals, or words that start with the same letter, or repeat a sentence. As a result, the robot must process various forms and lengths of speech. The robot-SCT, on the other hand, requires only short responses, i.e., 'yes' or 'no', resulting in fewer speech recognition errors and less impact on the evaluation results with little difference between the robot-SCT and the human-SCT group in terms of SCT score.

Furthermore, Desideri and colleagues found that compared to the assessment conducted by a clinician, participants in the assessment involving a robot spent more time looking at the robot and made fewer gaze aversions, suggesting that the participants might be more attracted to the robot as a consequence of the novelty effect and had a lower cognitive load from the conductor's face [31]. Similarly, in our experiment, a relatively small standard deviation was observed in the robot-SCT scores compared to those of the human-SCT, which can be interpreted as a relatively high concentration and a relatively lower cognitive burden. However, this needs further study since our experiment was conducted as a between-subject design with a difference in the population of the experimental groups; thus, the demographic features of the groups might affect the results.

We also compared the RT between items according to linguistic complexity within the robot-SCT group to verify the linguistic validity. The results showed that the RT of the canonical sentences was shorter than that of the non-canonical sentences. Our result was consistent with previous paper-and-pencil-based studies, which showed that the linguistic processing of sentences with canonical word order was easier than that of sentences with non-canonical word order [14, 43]. In addition, in our results, the performance of RT is supported by the result of the SCT score in that the performance for canonical sentences was higher than that of non-canonical sentences (Table 3). This study identified significant implications because the linguistic validity was assessed through RT analyses using the automatic measurement of robot-SCT. With the process to automatically extract RTs based on the detection of the offset of the robot's audio stimuli and the onset of the user's voice responses, the utilization of robot-SCT facilitated easier analysis of time-related measurements compare to human-SCT. In contrast, human-SCT or traditional paper-and-pencil-based assessments require manual segmentation of voice recorded files and calculations by examiners, which can compromise the reliability and validation of the time-related measurements. By utilizing robot-SCT, we expected to overcome the limitations of human-SCT and conduct a more valid assessment.

Finally, the average SUS score suggests that the robot-SCT system was perceived as an assessment with good ease-of-use, usability, and learnability for young adults. Even though there was negative feedback about the robot's voice, the non-rhythmic and slow voice was intentionally designed to deliver messages clearly to more vulnerable users. Thus, it needs to be reevaluated by the elderly or people with mild cognitive impairment, who are the expected users of the proposed robot-SCT system.

5.2 Limitations

Our study has a limitation regarding the medium, which may lower the naturalness of the human–robot verbal interactions involved. During the entire robot-SCT experiment, participants had to wear a headset to avoid speech recognition failures in a noisy environment. A previous study found that inaccurate speech recognition could affect the performance of the cognitive assessment by a robot [3]. Although the proposed system achieved a low error rate for speech recognition through the use of a headset, it is necessary to improve the speech recognition performance to allow for a more reliable verbal interaction in an uncontrolled and open space. In addition, although the fixed short yes/no answer could minimize the unnecessary effect of the answer on the evaluation results, it may have decreased the naturalness of the human–robot interaction during the experiment. Ultimately, it would be desirable to enable cognitive evaluation in various linguistic aspects through the use of more words. Further study is, therefore, needed for more natural and richer dialogue-based language assessments [44]. Finally, we only included young adults with normal cognitive functions. Therefore, we cannot generalize the results of the current study to people with dementia or mild cognitive impairment, who can be one of the target users of the proposed system. However, we would like to mention that MCI symptoms can be observed in the younger population due to psychiatric disorders like schizophrenia and depression, and therefore, the current system may be applicable for the evaluation of cognitive assessment in such groups [45, 46].

6 Conclusions

In this study, we demonstrate the validity and usability of a language assessment through verbal interaction with a robot by comparing the performance of the assessment with the test result obtained through an assessment conducted by a human examiner. The robot-SCT system proposed in this study can provide a reliable and easy-to-use interface for language assessment. Our research emphasizes the potential of the robot-SCT system that is equipped with the automatic recording of the test scores and the reaction times. This expands on previous efforts in HRI aimed at incorporating social robots into cognitive assessment procedures. Nevertheless, as this study only included young adults, future studies should evaluate the robotic-SCT systems to aid in the assessment of cognitive decline in older adults.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11370-023-00505-2.

Acknowledgements This work was supported by the Technology Innovation Program (10077553) Development of Social Robot Intelligence for Social Human-Robot Interaction of Service Robots and the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No. CAP21052-200) and the Government-wide R&D Fund for Infections Disease Research (GFID), funded by the Ministry of the Interior and Safety, Republic of Korea (Grant Number: 20014463) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2022R1A2C2005062) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1I1A4063209).

Declarations

Conflict of interests The authors have no relevant financial or non-financial interests to disclose.

Ethical approval This work was approved by the Korea Institute of Science and Technology IRB Committee under IRB Number 2021-024.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent to publication The authors affirm that human research participants provided informed consent for publication of the image and video in Fig. 3b and Supplementary 1.

References

1. Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, Ballard C, Banerjee S, Burns A, Cohen-Mansfield J, Cooper C (2017) Dementia prevention, intervention, and care. Lancet 390(10113):2673–2734. https://doi.org/10.1016/s0140-6736(17)31363-6
2. Papastavrou E, Kalokerinou A, Papacostas SS, Tsangari H, Sourtzi P (2007) Caring for a relative with dementia: family caregiver burden. J Adv Nurs 58(5):446–457. https://doi.org/10.1111/j.1365-2648.2007.04250.x
3. Lees RA, Hendry BAK, Broomfield N, Stott D, Larner AJ, Quinn TJ (2017) Cognitive assessment in stroke: feasibility and test properties using differing approaches to scoring of incomplete items. Int J Geriatr Psychiatry 32(10):1072–1078. https://doi.org/10.1002/gps.4568
4. Lyons BE, Austin D, Seelye A, Petersen J, Yeargers J, Riley T, Sharma N, Mattek N, Wild K, Dodge H, Kaye JA (2015) Pervasive computing technologies to continuously assess Alzheimer's disease progression and intervention efficacy. Front Aging Neurosci 10(7):102. https://doi.org/10.3389/fnagi.2015.00102
5. Wall KJ, Isaacs ML, Copland DA, Cumming TB (2015) Assessing cognition after stroke. Who misses out? A systematic review. Int J Stroke 10(5):665–671. https://doi.org/10.1111/ijs.12506
6. Lees R, Fearon P, Harrison JK, Broomfield NM, Quinn TJ (2012) Cognitive and mood assessment in stroke research: focused review of contemporary studies. Stroke 43(6):1678–1680. https://doi.org/10.1161/strokeaha.112.653303
7. Kang Y, NA DL, Hahn S (1997) A validity study on the Korean mini-mental state examination (K-MMSE) in dementia patients. J Kor Neurol Assoc, pp 300–308.
8. Hoops S, Nazem S, Siderowf AD, Duda JE, Xie SX, Stern MB, Weintraub D (2009) Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. Neurology 73(21):1738–1745. https://doi.org/10.1212/wnl.0b013e3181c34b47
9. Devenney E, Hodges JR (2017) The mini-mental state examination: pitfalls and limitations. Pract Neurol 17(1):79–80. https://doi.org/10.1136/practneurol-2016-001520
10. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H (2005) The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 53(4):695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x
11. Choi CH, Park S, Park HJ, Cho Y, Sohn BK, Lee JY (2016) Study on cognitive reserve in Korea using Korean version of Cognitive Reserve Index Questionnaire. J Kor Neuropsychiatr Assoc 55(3):256–263. https://doi.org/10.4306/jknpa.2016.55.3.256
12. Kang Y, Jang SM, Na DL (2012) Seoul neuropsychological screening battery, 2nd edn (SNSB-II). Incheon: Human Brain Research & Consulting Co.
13. Lee AY, Lee J, Oh E, Yoon SJ, Yoon B, Yu SD (2019). Clinical utility of Seoul Neuropsychological Screening Battery-Core for dementia management project in the community. J Kor Neurol Assoc 37(3):277–283.https://doi.org/10.17340/jkna.2019.3.5
14. Sung JE (2015) Effects of syntactic structure on sentence comprehension ability as a function of the canonicity of word-order and its relation to working memory capacity in Korean-speaking elderly adults. Commun Sci Disord 20(1):24–33. https://doi.org/10.12963/csd.15229
15. Sung JE, Choi S, Eom B, Yoo JK, Jeong JH (2020) Syntactic complexity as a linguistic marker to differentiate mild cognitive impairment from normal aging. J Speech Lang Hear Res 63(5):1416–1429. https://doi.org/10.1044/2020_JSLHR-19-00335
16. Zygouris S, Tsolaki M (2015) Computerized cognitive testing for older adults: a review. Am J Alzheimer's Dis Other Dement® 30(1):13–28. https://doi.org/10.1177/1533317514522852
17. Atkins AS, Tseng T, Vaughan A, Twamley EW, Harvey P, Patterson T, Narasimhan M, Keefe RS (2017) Validation of the tablet-administered Brief Assessment of Cognition (BAC App). Schizophr Res 181:100–106. https://doi.org/10.1016/j.schres.2016.10.010
18. Koo BM, Vizer LM (2019) Mobile technology for cognitive assessment of older adults: a scoping review. Innovat Aging 3(1):igy038. https://doi.org/10.1093/geroni/igy038
19. DeRight J, Jorgensen RS (2015) I just want my research credit: frequency of suboptimal effort in a non-clinical healthy undergraduate sample. Clin Neuropsychol 29(1):101–117. https://doi.org/10.1080/13854046.2014.989267
20. Eysenbach G (2005) The law of attrition. J Med Internet Res 7:e11. https://doi.org/10.2196/jmir.7.1.e11
21. Wall KJ, Cumming TB, Koenig ST, Pelecanos AM, Copland DA (2018) Using technology to overcome the language barrier: the cognitive assessment for Aphasia App. Disabil Rehabil 40(11):1333–1344. https://doi.org/10.1080/09638288.2017.1294210
22. Tanaka H, Adachi H, Ukita N, Ikeda M, Kazui H, Kudo T, Nakamura S (2017) Detecting dementia through interactive computer avatars. IEEE J Transl Eng Health Med 5:1–11. https://doi.org/10.1109/jtehm.2017.2752152
23. Mirheidari B, Blackburn D, Walker T, Reuber M, Christensen H (2019) Dementia detection using automatic analysis of conversations. Comput Speech Lang 53:65–79. https://doi.org/10.1016/j.csl.2018.07.006
24. Feil-Seifer D, Mataric MJ (2005) Socially assistive robotics. In: 9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005, pp 465–468. https://doi.org/10.1109/icorr.2005.1501143
25. Qureshi MO, Syed RS (2014) The impact of robotics on employment and motivation of employees in the service sector, with special reference to health care. Saf Health Work 5(4):198–202. https://doi.org/10.1016/j.shaw.2014.07.003
26. Mukai T, Hirano S, Nakashima H, Sakaida Y, Guo S (2011) Realization and safety measures of patient transfer by nursing-care assistant robot riba with tactile sensors. J Robot Mechatron 23(3):360–369. https://doi.org/10.20965/jrm.2011.p0360
27. Burgar CG, Lum PS, Shor PC, Van der Loos HM (2000) Development of robots for rehabilitation therapy: the Palo Alto VA/Stanford experience. J Rehabil Res Dev 37(6):663–674 (**PMID:11321002**)

28. Shibata T, Wada K (2011) Robot therapy: a new approach for mental healthcare of the elderly—a mini-review. Gerontology 57(4):378–386. https://doi.org/10.1159/000319015
29. Tanaka M, Ishii A, Yamano E, Ogikubo H, Okazaki M, Kamimura K, Konishi Y, Emoto S, Watanabe Y (2012) Effect of a human-type communication robot on cognitive function in elderly women living alone. Med Sci Monit Int Med J Exp Clin Res 18(9):CR550. https://doi.org/10.12659/msm.883350.
30. Kumazaki H, Muramatsu T, Yoshikawa Y, Matsumoto Y, Takata K, Ishiguro H, Mimura M (2022) Android robot promotes disclosure of negative narratives by individuals with autism spectrum disorders. Front Psych 13:899664
31. Desideri L, Ottaviani C, Malavasi M, di Marzio R, Bonifacci P (2019) Emotional processes in human-robot interaction during brief cognitive testing. Comput Hum Behav 90:331–342. https://doi.org/10.1016/j.chb.2018.08.013
32. Di Nuovo A, Varrasi S, Lucas A, Conti D, McNamara J, Soranzo A (2019) Assessment of cognitive skills via human-robot interaction and cloud computing. J Bionic Eng 16(3):526–539. https://doi.org/10.1007/s42235-019-0043-2
33. Li J (2015) The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int J Hum Comput Stud 77:23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001
34. Robocare. http://www.robocare.co.kr/. Accessed 19 May 2022
35. Web speech API, 18-Aug-2020. https://wicg.github.io/speech-api/. Accessed 19 May 2022
36. Youn JC, Kim KW, Lee DY, Jhoo JH, Lee SB, Park JH, Choi EA, Choe JY, Jeong JW, Choo IH, Woo JI (2009) Development of the subjective memory complaints questionnaire. Dement Geriatr Cogn Disord 27(4):310–317. https://doi.org/10.1159/000205512
37. Kang Y (2006) A normative study of the Korean Mini-Mental State Examination (K-MMSE) in the elderly. Korean J Psychology 25:1–2
38. Brook J (1996) Sus: A 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B (eds) Usability evaluation in industry. CRC Press, Boca Raton, pp 207–212 https://doi.org/10.1201/9781498710411-35
39. Choi S, Jo E, Sung JE (2021) Preliminary study on the action naming test: online vs. offline comparisons by presentation type. J Speech-Lang Hear Disord 30(2):87–97. https://doi.org/10.15724/jslhd.2021.30.2.087
40. Dekhtyar M, Braun EJ, Billot A, Foo L, Kiran S (2020) Videoconference administration of the western aphasia battery–revised: feasibility and validity. Am J Speech Lang Pathol 29(2):673–687. https://doi.org/10.1044/2019_AJSLP-19-00023
41. Theodoros D, Hill A, Russell T, Ward E, Wootton R (2008) Assessing acquired language disorders in adults via the Internet. Telemed e-Health 14(6):552–559. https://doi.org/10.1089/tmj.2007.0091
42. Bangor A, Kortum PT, Miller JT (2008) An empirical evaluation of the system usability scale. Int J Human-Comp Interact 24(6):574–594
43. Sung JE, Yoo JK, Lee SE, Eom B (2019) Effects of age, working memory, and word order on passive-sentence comprehension: evidence from a verb-final language. Int Psychogeriatr 29(6):939–948. https://doi.org/10.1017/S104161021700004
44. Mirheidari B, Blackburn DJ, Harkness K, Walker T, Venneri A, Reuber M, Christensen H (2017) An avatar-based system for identifying individuals likely to develop dementia. In: Interspeech 2017, pp 3147–3151. https://doi.org/10.21437/interspeech.2017-690
45. Fisekovic S, Memic A, Pasalic A (2012) Correlation between MoCA and MMSE for the assessment of cognition in schizophrenia. Acta Inform Med 20(3):186–189. https://doi.org/10.5455/aim.2012.20.186-189
46. Moirand R, Galvao F, Lecompte M, Poulet E, Haesebaert F, Brunelin J (2018) Usefulness of the Montreal Cognitive Assessment (MoCA) to monitor cognitive impairments in depressed patients receiving electroconvulsive therapy. Psychiatry Res 259:476–481. https://doi.org/10.1016/j.psychres.2017.11.022