Title: Reliability and acceptability of using a social robot to conduct cognitive tests for community-dwelling older adults

Short running title: Robot cognitive test's reliability & acceptability

Authors

Kana Takaeda[1], Tomoko Kamimura[1], Takenobu Inoue[2], Yuko Nishiura[2]

[1]Department of Medical Sciences, Graduate School of Medicine, Shinshu University
Matsumoto, Japan
[2]Department of Assistive Technology, Research Institute, National Rehabilitation Center for Persons with Disabilities
Tokorozawa, Japan

Corresponding author
Tomoko Kamimura
Department of Medical Sciences, Graduate School of Medicine, Shinshu University
3-1-1 Asahi, Matsumoto, Nagano Pref. 390-8621, Japan
E-mail tkamimu@shinshu-u.ac.jp
Tel +81(263)37-2395

Abstract

Aim: To improve access to cognitive testing for older adults, the reliability and acceptability of a speech-based cognitive test administered by a social robot were investigated.

Methods: The Japanese version of the Telephone Interview for Cognitive Status was administered by a social robot to participants recruited from retirement homes and adult daycare facilities. The robot's dialogue and gestures were preprogrammed, while the researcher controlled the timing of proceeding to the next question and scored participants' responses. We examined the internal consistency, alternate form reliability (Experiment 1), and test-retest reliability (Experiment 2) of the cognitive test. The acceptability of the cognitive test was also examined using a questionnaire in Experiment 2.

Results: Sixty-six individuals (mean age: 81.2 ± 5.8 years) participated in Experiment 1; the internal consistency (Cronbach's α) of the test was 0.691, and its alternate form reliability (measured by interclass correlation coefficient) was 0.728. Forty of these individuals (mean age: 82.0 ± 5.4 years) also participated in Experiment 2, and the test-retest reliability was 0.818. According to the questionnaire responses, over half of the participants wanted (or very much wanted) to use the robot version of the test to measure the deterioration of their cognitive function.

Conclusions: A robot-administered cognitive test might have satisfactory reliability and acceptability to community-dwelling older adults if those aspects of the test implemented by the researcher can also be automated successfully.

Keywords: acceptability, cognitive dysfunction, older adults, reliability, robotics

Introduction

Early recognition of cognitive impairment is necessary for diagnosis and appropriate treatment, education, and psychosocial support, as well as to ensure that patients can engage in decision-making regarding their own life planning and health care.[1] In accordance with the expansion of these needs caused by rapid population aging, the use of computerized cognitive tests (CCTs) is rapidly increasing.[2] With appropriate selection, use, and interpretation, the use of CCTs has the potential to contribute to cost-effective delivery of services, improved healthcare resource allocation, early identification of patients in need of more comprehensive diagnostic evaluation, and improved cognitive outcomes.[2] However, the CCTs' advantages and limitations have been debated since the 1980s,[3,4] and ongoing discussion regards how best to use such tests in healthcare.[4-6]

In this study, we explore a new topic in computerized cognitive testing: the possibility of a social robot as a new interface for speech-based CCTs to improve access to cognitive testing. Social robots in this article are defined (see Dautenhahn[7] and Lee[8]) as robots that both fulfill a useful purpose and possess social intelligence and skills that enable them to interact with people in a socially acceptable manner. Recently, a social robot began to function as an interface for cognitive training, affective therapy, social facilitator, companionship, and physiological therapy for older adults.[9]

Although currently mainstream CCTs involve performance-based testing using a touch screen, keyboard, or mouse,[5,6] widely used screening tests to measure global cognitive function mainly take a question-and-answer format, such as the Mini Mental State Examination (MMSE),[10] the Telephone Interview for Cognitive Status (TICS),[11,12] and the Montreal Cognitive Assessment (MoCA).[13] To take these tests, the examinees must meet professionals either in person or remotely. However, if a social robot exists in the community and administers these tests, the community-dwelling people may be able to take them easily, before a more comprehensive diagnostic evaluation. This may help improve access to such cognitive testing. However, it is unknown whether a cognitive test administered by a social robot would be considered acceptable by older adults, because few studies have focused on this issue. Furthermore, the reliability of such cognitive tests administered by a social robot has not been verified.

In this study, as initial exploration of the potential for the speech-based computerized cognitive testing administration by social robots, we investigated the reliability of data obtained in such a way, and the acceptability of this type of interface among older adults. Of the 4 facets of reliability,[14] we investigated the internal consistency, alternate form reliability, and test-retest reliability of a robot-administered cognitive test.

Methods

Robot-administered cognitive test

In this study, we used the Japanese version of the Telephone Interview for Cognitive Status (TICS-J)[15] as our exemplar of a cognitive test. The original TICS is a brief test of cognitive functioning administered by a professional via telephone, not requiring face-to-face interaction. We adopted the TICS in this study because there are no visual components, and it seems to be more easily applicable and less time-consuming than MMSE and MoCA. It consists of 11 test items (own name, date, address, counting backward, word list learning, subtractions, responsive naming, repetition, prime minister’s name, finger tapping, and word opposites).[11] The total possible score is 41, with higher scores indicating better functioning. TICS scores correlate highly with MMSE scores[10], and the tool has high test-retest reliability and specificity for the detection of cognitive impairment.[12] The TICS-J has the same structure as the TICS, and its high test-retest reliability and excellent specificity for the detection of cognitive impairment among Japanese older adults have been demonstrated.[15]

In this study, the Communication Robot PaPeRo R500® (NEC Corporation, Figure 1) conducted the TICS-J face to face; this administration of a cognitive test by a social robot is hereinafter referred to as a "robot test." The test was conducted according to a fixed scenario that we developed, in which the robot first introduced itself to the participant, introduced and explained the TICS-J; thereafter, the robot alternately asked a question from the test and reacted to the participant's response, and finished the test with a closing greeting. This fixed scenario included specification of the robot's dialogue and gestures. For example, after saying, "Please tell me your full name," the robot's ears glowed orange and its head tilted while waiting for the response, after which the robot repositioned its head and said, "Thank you" and proceeded to the next question by saying "Well then…." The timing of the robot's utterances and gestures, however, was not automatic. Researchers sitting in another room monitored the participants' responses via a microphone and camera and directed the robot to proceed to the next piece of dialogue or gesture. Researchers also scored participants' responses. Participants were initially told that the test was fully administered by the robot, and were only informed about researchers' involvement after the test ended.

Experiment 1

We recruited participants at five retirement homes and two adult daycare facilities. The selection criteria were that participants were aged 65 years or over and capable of verbal communication. The exclusion criteria were severe hearing or visual impairment, and diagnosis or suspicion of psychiatric disorder. Participants were given a cash voucher worth one thousand yen per test for their participation. Data were collected between November 2015 and June 2016.

In Experiment 1, we investigated internal consistency and alternate form reliability. The TICS-J was conducted face to face twice for each participant, once in the form of a robot test (administered by the social robot) and once as a human test (administered by an occupational therapist). The order of these two tests was counterbalanced across participants, and the interval between them was about one week.

Experiment 2

Participant recruitment and the eligibility criteria were same as for Experiment 1. In Experiment 2, we investigated the test-retest reliability and acceptability of the robot test. Data were collected between January and August 2016.

Data for Experiment 2 were obtained through an additional robot test conducted about six weeks after the second test in Experiment 1 with some of the original participants. A questionnaire on the acceptability of the robot test was simultaneously administered.

The questionnaire contained 4 questions: 1) How trustworthy do you feel this robot test is? 2) How favorable is your opinion of this robot test? 3) To what extent would you want to use this robot test to measure your deterioration in cognitive function, if it existed at your neighborhood supermarket or community center? 4) Which do you prefer, a test administered by a robot or one administered by a human, and why? Participants responded to questions 1) to 3) on a 5-point scale from 0 (not at all) to 4 (very much so).

Statistical analysis

Cronbach’s α was calculated to measure internal consistency, with values greater than 0.80 considered excellent, 0.75–0.79 good, 0.70–0.74 moderate, 0.65–0.69 fair, and 0.65 or lower unsatisfactory.[16] The interclass correlation coefficient (ICC) between the human and robot test and between the two instances of the robot test, along with

95% confidence intervals (CIs), were calculated to measure alternate form reliability and test-retest reliability, respectively, with ICC values greater than 0.90 considered very high, 0.80–089 high, 0.70–0.79 adequate, 0.60–0.69 marginal, and 0.59 or lower low.[17] IBM SPSS Statistics 22.0 was used for these analyses.

Descriptive statistics are presented on data from the questionnaire on the acceptability of the robot test.

Ethical considerations

This study was approved by the Medical Ethics Committee of Shinshu University and the Ethics Committee of the National Rehabilitation Center for Persons with Disabilities. All participants provided written informed consent before data collection.

Results

Experiment 1

Of 72 people who applied to participate, 66 took part in Experiment 1 (mean age: 81.2 ± 5.8 years; 52 women); demographic details are presented in Table 1. Of the 6 potential participants excluded, 2 were suspected of having severe hearing impairment and 4 were in poor health on the test day.

Mean TICS-J scores were 29.9 ± 5.6 (robot test), and 32.7 ± 5.0 (human test) (Table 1). The ICC, an alternate form of reliability, was 0.728 (95% CI: 0.218–0.884), in the “adequate” range.

The robot-administered TICS-J had a Cronbach’s α of 0.691, rated as “fair” internal consistency (Table 1).

Experiment 2

Of the 66 participants in Experiment 1, all 40 who participated in January 2016 or later also participated in Experiment 2 (mean age: 82.0 ± 5.4 years; 33 women); further demographic details are in Table 2. The average period between the first robot test in Experiment 1 and the second in Experiment 2 was 49.6 ± 6.5 days.

The average score on the first robot-administered TICS-J was 30.1 ± 5.7, and the average on the second test was 32.0 ± 5.6 (Table 2). The ICC between these tests was 0.818 (95% CI: 0.682–0.899), in the "high" range.

Regarding impressions of the robot-administered TICS-J, 32 participants (80.0%) perceived the robot test as "very trustworthy" or "trustworthy" (Figure 2). Twenty-four (60.0%) reported "very favorable" or "favorable" impressions. Twenty-three participants (57.5%) answered that they "very much wanted to use" or "wanted to use" this test to measure their deterioration in cognitive function. Fifteen participants (37.5%) reported that they preferred a robot, 14 (35.0%) that they preferred a human, and 10 (25.0%) that they had no preference. Reasons for preferring a robot included "The robot is cute. I do not feel nervous, as if it were a toy or a game" and "When I cannot answer correctly, sometimes I am embarrassed with a human, but it is alright with a robot." Reasons for preferring a human included "I feel nervous with a robot," "Questions from a human are easier to hear because he or she talks at the right speed to match my pace," and "It is possible to ask a human a question."

Discussion

The results of this study suggest the possibility that cognitive tests employing social robots as user interfaces, such as Communication Robot PaPeRo R500®, can be reliable for and acceptable to community-dwelling older adults.

The internal consistency and the test-retest reliability of the cognitive test administered in this study (Experiment 1 and Experiment 2, respectively) were satisfactory (Cronbach's α=0.691, ICC=0.818), as has been shown in a previous study of CCTs. Yu et al. developed a CCT for the Beijing version of the MoCA for older adults and reported that its internal consistency (Cronbach's α) is 0.72, and its test-retest reliability (ICC) is 0.82.[18] Our results and those of Yu et al. suggest that cognitive tests such as the TICS and the MoCA might be reliable when administered to older adults by a social robot or a computer.

However, the test-retest reliability of the cognitive test (Experiment 2) was not as high as that in previous studies, in which professionals conducted the test either remotely or in person, and this aspect of the results was also the same as in previous studies. Konagaya et al. report that the ICC for the TICS-J conducted remotely in a standardized way is 0.946,[15] which is higher than the ICC of 0.818 from the present study. In previous studies, ICCs of 0.862[19] and 0.87[20] were observed for the Beijing version of the MoCA conducted in person in a standardized way, higher than the ICC of 0.82 reported by Yu et al.[18]

Additionally, although the alternate form reliability of the cognitive test administered in the present study was adequate (Experiment 1), it was not as high as that measured by professionals who conducted the test remotely via videoconferencing.[21-23] According to Castanho et al., the degree of coincidence in scores on the modified TICS (TICSm) when administered by professionals via videoconferencing and telephone, measured by Pearson's correlation coefficient, is 0.885,[21] representing a higher level of alternate form reliability than the ICC of 0.728 from the present study.

Thus, although the robot test has a satisfactory reliability, it should not replace detection of cognitive impairment by professionals under the present circumstances.

The acceptability of a speech-based CCT (administered by a social robot) in this study was as good as that of performance-based CCTs from previous studies. Fredrickson[24)] and Darby[25)] investigated the usability of a performance-based CCT of

CogState conducted serially, which is an internet-based self-administrated cognitive test. They reported that a majority (85%, 95%) of older participants successfully completed the baseline test. Hansen[26)] investigated users' preferences of another internet-based self-administered performance-based CCT (Memoro); almost twice as many older participants preferred the self-administered computerized test to the analog pen-and-paper test owing to it being less difficult and allowing participants more control and less scrutiny by examiners. In this study (Experiment 1 and Experiment 2), all participants completed the test successfully. Moreover, according to the findings of Experiment 2, several individuals (80%) found the robot test trustworthy, while 57.5% wanted to use it to measure their own cognitive deterioration.

The number of participants in Experiment 2 who preferred the robot test to the human test was almost equal to the number who gave the opposite answer. When asked why the test administered by the social robot was preferred to the human test, participants' comments referred to the cute and game-like nature of the robot test, which helped them avoid nervousness or embarrassment. This indicates that cognitive tests administered by a social robot with an affectionate appearance that interacts based on a pre-programmed scenario would reduce the psychological barriers to tests experienced by some older adults. Therefore, the use of such cognitive tests may help increase the number of older adults making first contact with a cognitive test.

This study suggests that social robots as user interfaces of CCTs might improve access to cognitive testing for community-dwelling older adults; however, the test requires a user's ability to take it (i.e., correctly understand the questions and answer them without being distracted by the robot's behaviors or ambient noise) of their own will, along with a system in which professionals appropriately administer the advanced cognitive test.

There are several limitations to this study. The sample size was small, and consisted predominantly of healthy older adults. Therefore, it is unknown whether the findings are generalizable to other older adults, including individuals diagnosed with or suspected of having psychiatric disorders. Additionally, all participants volunteered to participate in the robot test, which may have produced a sampling bias: in particular, the sample contained a disproportionate number of individuals who found a robot-administered cognitive test to be acceptable. Therefore, the finding that a large proportion of participants gave positive answers regarding trustworthiness, favorability, and so on cannot be assumed to hold true for the general population of older adults. Finally, not all aspects of the cognitive test were conducted by the social robot in this study; researchers monitored the timing of the participant's responses, decided when the robot should proceed to the next question, and provided feedback on the results after scoring. Therefore, to determine whether social robots can be reliable and acceptable as a new type of interface for CCTs, it is necessary to develop the above-mentioned technical functions and to investigate whether social robots with such capabilities are also found to be acceptable and reliable by older adults.

This study is, to our knowledge, the first to suggest that social robots are preferable to some older adults as a new user interface for cognitive tests taking a question-and-answer format like the TICS. It also shows that such tests may have fair internal consistency and acceptable reliability if additional functions can be developed to a good standard to enable the social robot to proceed to the next question appropriately, score the user's responses, and provide feedback on the results.

Acknowledgments

This research was supported by a Strategic Promotion of Innovative Research and Development grant from the Japan Science and Technology Agency.

We wish to thank Hiroaki Kojima and Ken Sadohara from the National Institute of Advanced Industrial Science and Technology for installing the scenario

dialogue for the cognitive test and operating the robot during the experiments.

Disclosure statement

Since the paid rental business of the Communication Robot PaPeRo R500® ended in March 2016, it was provided for free between April and August 2016 by the manufacturer (NEC Corporation).

References

1. Morley JE, Morris JC, Berg-Weger M, et al. Brain Health: The importance of recognizing cognitive impairment: An IAGG Consensus Conference. *J Am Med Dir Assoc* 2015; **16**: 731-39.
2. Roebuck-Spencer TM, Glen T, Puente AE, et al. Cognitive screening tests versus comprehensive neuropsychological test batteries: A National Academy of Neuropsychology Education Paper. *Arch Clin Neuropsychol* 2017; **32**: 491-98.
3. Schlegel RE, Gilliland K. Development and quality assurance of computer-based assessment batteries. *Arch Clin Neuropsychol* 2007; **22**: S49-S61.
4. Wild K, Howieson D, Webbe F, Seelye A, Kaye J. Status of computerized cognitive testing in aging: A systematic review. *Alzheimers Dement* 2008; **4**: 428-37.
5. Zygouris S, Tsolaki M. Computerized cognitive testing for older adults: a review. *Am J Alzheimers Dis Other Demen* 2015; **30**: 13-28.
6. Aslam RW, Bates V, Dundar Y, et al. Automated tests for diagnosing and monitoring cognitive impairment: a diagnostic accuracy review. *Health Technol Assess* 2016; **20**: 1-73.
7. Dautenhahn K. Socially intelligent robots: dimensions of human-robot interaction.

Philos Trans R Soc Lond B Biol Sci 2007; **362**: 679-704.

8. Lee KM, Park N, Song H. Can a robot be perceived as a developing creature? *Human Communication Research* 2005; **31**: 538-563.

9. Abdi J, Al-Hindawi A, Ng T, Vizcaychipi MP. Scoping review on the use of socially assistive robot technology in elderly care. *BMJ Open* 2018; **8**: e018815. doi: 10.1136/bmjopen-2017-018815.

10. Folstein MF, Folstein SE, McHugh PR. "Mini-Mental State", a practical method for grading the cognitive state of patients for the clinician. J *Psychiatr Res* 1975; **12**: 189-198.

11. Welsh KA, Breitner JCS, Magruderhabib KM. Detection of dementia in the elderly using telephone screening of cognitive status. *Neuropsychiatry Neuropsychol Behav Neurol* 1993; **6**: 103-10.

12. Brandt J, Spencer M, Folstein M. The Telephone Interview for Cognitive Status. *Cog Behav Neurol* 1988; **1**: 111-117.

13. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 2005; **53**: 695-99.

14. Bauer RM, Iverson GL, Cernich AN, Binder LM, Ruff RM, Naugle RI. Computerized neuropsychological assessment devices: Joint Position Paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Clin Neuropsychol* 2012; **26**: 177-96.

15. Konagaya Y, Washimi Y, Hattori H, Takeda A, Watanabe T, Ohta T. Validation of the Telephone Interview for Cognitive Status (TICS) in Japanese. *Int J Geriatr Psychiatry* 2007; **22**: 695-700.

16. Ponterotto JG, Ruckdeschel DE. An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Percept Mot Skills* 2007; **105**: 997-1014.

17. Lezak MD, Howieson DB, Bigler ED, Tranel D. Neuropsychological Assessment 5th ed. New York: Oxford University Press, 2012.

18. Yu K, Zhang SG, Wang QS, et al. Development of a computerized tool for the Chinese version of the Montreal Cognitive Assessment for screening mild cognitive impairment. *Int Psychogeriatr* 2015; **27**: 213-19.

19. Hu JB, Zhou WH, Hu SH, et al. Cross-cultural difference and validation of the Chinese version of Montreal Cognitive Assessment in older adults residing in Eastern China: Preliminary findings. *Arch Gerontol Geriatr* 2013; **56**: 38-43.

20. Chen X, Zhang R, Xiao Y, Dong JQ, Niu X, Kong WJ. Reliability and validity of the Beijing version of the Montreal Cognitive Assessment in the evaluation of cognitive function of adult patients with OSAHS. *Plos One* 2015; **10**.

21. Castanho TC, Amorim L, Moreira PS, et al. Assessing cognitive function in older adults using a videoconference approach. *EBioMedicine* 2016; **11**: 278-84.

22. Park HY, Jeon SS, Lee JY, Cho AR, Park JH. Korean version of the Mini-Mental State Examination using smartphone: a validation study. *Telemed J E Health* 2017; **23**: 815-21.

23. Cullum CM, Hynan LS, Grosch M, Parikh M, Weiner MF. Teleneuropsychology: evidence for video teleconference-based neuropsychological assessment. *J Int Neuropsychol Soc* 2014; **20**: 1028-33.

24. Fredrickson J, Maruff P, Woodward M, et al. Evaluation of the usability of a brief computerized cognitive screening test in older people for epidemiological studies. *Neuroepidemiology* 2010; **34**: 65-75.

25. Darby DG, Fredrickson J, Pietrzak RH, Maruff P, Woodward M, Brodtmann A. Reliability and usability of an internet-based computerized cognitive testing battery in community-dwelling older people. *Comput Human Behav* 2014; **30**: 199-205.

26. Hansen TI, Haferstrom ECD, Brunner JF, Lehn H, Haberg AK. Initial validation

of a web-based self-administered neuropsychological test battery for older adults and seniors. *J Clin Exp Neuropsychol* 2015; **37**: 581-94.

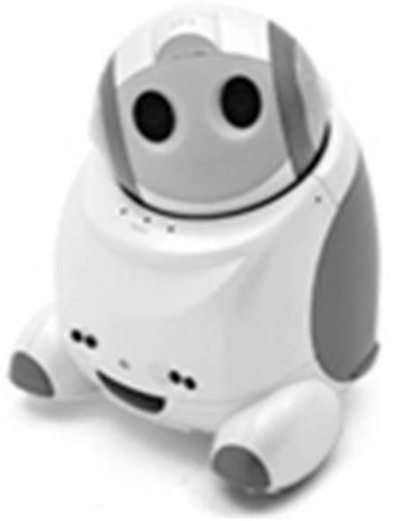

Figure 1
We used this social robot, which was a Communication Robot PaPeRo R500® produced by NEC Corporation..

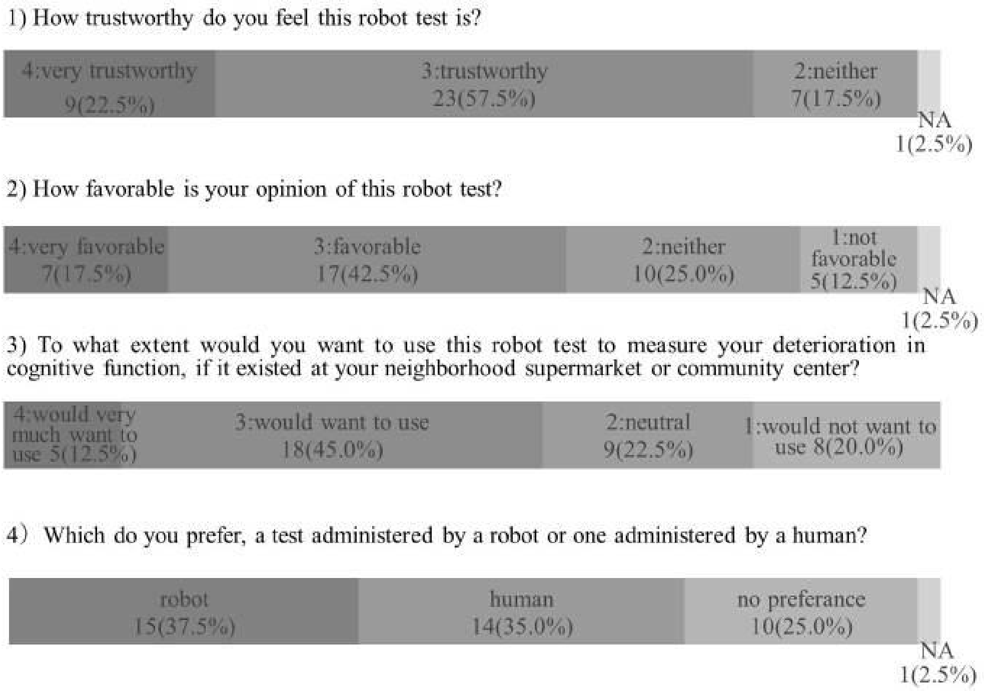

Figure 2
The results of the questionnaire on the acceptability of the robot test in Experiment 2.

Table 1. Participant characteristics and results of Experiment 1

Participant characteristics	
Number of participants	66
Age (years) §	81.2 ± 5.8
Sex Female#	52 (79%)
Male#	14 (21%)
Education (years) §	12.9 ± 3.0
TICS-J scores	
Robot test score§	29.9 ± 5.6
Human test score§	32.7 ± 5.0
Reliability of the robot test	
Alternate form reliability (ICC and 95% CI)	0.728 [0.218–0.844]
Internal consistency (Cronbach's α)	0.691

TICS-J: Telephone Interview for Cognitive Status in Japanese; ICC: interclass correlation coefficient; CI: confidence interval

§: Data are given in the form: mean ± SD. #: Data are given in the form: *n* (%).

Table 2. Participant characteristics and results of Experiment 2

Participant characteristics	
Number of participants	40
Age (years)[§]	82.0 ± 5.4
Sex Female[#]	33 (83%)
Male[#]	7 (17%)
Education (years)[§]	12.3 ± 2.7
TICS-J scores	
1st robot test score[§]	30.1 ± 5.7
2nd robot test score[§]	32.0 ± 5.6
Reliability of the robot test	
Test-retest reliability (ICC and 95% CI)	0.818 [0.682–0.899]

TICS-J: Telephone Interview for Cognitive Status in Japanese; ICC: interclass correlation coefficient; CI: confidence interval

§: Data are given in the form: mean ± SD. #: Data are given in the form: *n* (%).