International Journal of Social Robotics
https://doi.org/10.1007/s12369-024-01158-6

Exploring the Viability of Socially Assistive Robots for At-Home Cognitive Monitoring: Potential and Limitations

Matteo Luperto[1] · Marta Romeo[2] · Francesca Lunardini[3] · Javier Monroy[4] · Daniel Hernández García[2] · Carlo Abbate[5] · Angelo Cangelosi[6] · Simona Ferrante[3] · Javier Gonzalez-Jimenez[4] · Nicola Basilico[1] · N. Alberto Borghese[1]

Accepted: 10 June 2024

Abstract
The early detection of mild cognitive impairment, a condition of increasing impact in our aging society, is a challenging task with no established answer. One promising solution is the deployment of robotic systems and ambient assisted living technology in the houses of older adults for monitoring and assistance. In this work, we address and discuss a qualitative analysis on the feasibility and acceptability of a socially assistive robot (SAR) deployed in prospective users' houses to monitor their cognitive capabilities through a set of digitalised neuropsychological tests and spot questions conveniently integrated within the robotic assistant's daily tasks. We do this by describing an experimental campaign where a robotic system, integrated with a larger framework, was installed in the house of 10 users for a duration of at least 10 weeks, during which their cognitive capabilities were monitored by the robot. Concretely, the robots supervised the users during the completion of the tests and transparently monitored them by asking questions interleaved in their everyday activities. Results show a general acceptance of such technology, being able to carry out the intended tasks without being too invasive, paving the way for an impactful at-home use of SARs.

Keywords Cognitive monitoring · Mild cognitive impairment · Field study · Human–robot interaction

1 Introduction

The global population is rapidly ageing. For instance, in Europe, the expected proportion of adults above 65 years old is projected to rise from 20 to 30% from 2019 to 2070 [13]. As a consequence, the old-age dependency ratio, measuring the number of people aged 65 and above relative to those aged 20–64, is estimated to increase from 34 to 59% by 2070, causing a shift in social structures and highly impacting healthcare programs [13]. Enhancing independent living and "ageing in place" provides a number of benefits in terms of cost, effectively meeting the needs of older people, and delaying nursing home admission [10]. Towards this end, researchers have directed their efforts into developing novel technologies for remote monitoring and everyday assistance of older adults, especially focusing on their acceptability. In this context, socially assistive robots (SARs) [14] emerged as a valuable asset, able to carry out health-monitoring functionalities directly in the users' homes. This category of robots is often based on autonomous mobile platforms that, thanks to a high level of integration into ambient assisted living

Matteo Luperto and Marta Romeo contributed equally to this work.

✉ Matteo Luperto
matteo.luperto@unimi.it

✉ Marta Romeo
m.romeo@hw.ac.uk

1 Department of Computer Science, University of Milan, Milan, Italy

2 School of Mathematics and Computer Science, Heriot-Watt University, Edinburgh, UK

3 Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

4 Machine Perception and Intelligent Robotics Group, Department of System Engineering and Automation, Biomedical Research Institute of Málaga, University of Malaga, Malaga, Spain

5 Istituto Palazzolo, IRCCS Fondazione Don Carlo Gnocchi, Milan, Italy

6 Department of Computer Science, The University of Manchester, Manchester, UK

Published online: 12 July 2024

environments (AAL), can convey services such as delivering messages and reminders, teleconferencing with family members or caregivers, and guiding users through physical and cognitive exercises [4]. For these applications, SARs have already demonstrated a high level of acceptability by their end users [12, 21, 37].

When dealing with the effects of ageing, cognitive decline is among the most pressing concerns [25] as it has a strong impact on the life quality of older adults and can progress into dementia. There is currently no effective treatment for dementia, but available therapies have proven to be effective in slowing down the decline if detected in its early stages, a phase commonly referred to as mild cognitive impairment (MCI) [5]. At present, neuropsychological assessment is the most useful tool for identifying patients with MCI. It is usually conducted through standardised *cognitive tests*, commonly carried out in a controlled clinical environment under experts' supervision.

Early assessment of MCI can enable effective prevention strategies but it still represents a challenging task for which a solution is largely missing. In fact, cognitive tests are usually scheduled only after evident symptoms have already manifested. Moreover, conducting tests in clinical facilities is not ideal as it might introduce factors that could interfere with the patient's behaviour, potentially jeopardizing the test's validity [6]. Administering tests in a familiar home environment can contribute to overcome these limitations.

For the aforementioned reasons, we developed, within the EU-funded MoveCare project [29], a multi-actor framework comprising a SAR to support the independent living of older adults through engagement, assistance, and monitoring, both cognitive and physical. This framework has been designed to operate in the users' house, in complete autonomy, for a long period of time. In this paper, we focus on describing the deployment of the cognitive-monitoring module, a component of this multi-actor framework, in the houses of potential end users. The objective of the cognitive-monitoring module is not to be a replacement for a clinical cognitive assessment. Instead, it aims at being an instrument that, if used at home, could ease the first steps toward a proper clinical evaluation, thus increasing the chances for early MCI assessment. In this context, cognitive monitoring was carried out, under the supervision provided by a SAR, in two scenarios: through classical *neuropsychological screening tests*, and through a series of *spot questions*.

With the NeuroPsychological Test scenario (NPT, Sect. 3.1), we aim to investigate the at-home feasibility of a complex and time-consuming interaction where the robot explains and guides older adults through the tests.

In the Spot Questions scenario (SQ, Sect. 3.2), the robot asks the users single repeated-in-time short questions on a weekly basis. SQ collect answers over different months to enable a longitudinal data analysis for monitoring purposes. In this work, we assess the feasibility and intrusiveness of such kind of longitudinal assessment, exploiting the fact that the SAR has been deployed in the users' apartments for a long time.

We report and discuss the results of an experimental campaign for the two aforementioned scenarios where the SARs were in operation in the house of 10 older adults for at least 10 weeks, interacting autonomously with the users during their regular activities of daily living. With our work, we contribute to a better understanding of how these systems are used at home, shedding light on their actual usage within the day-to-day life of end users. Our data suggest how long-term home deployments of SARs can efficiently achieve cognitive monitoring of older users, without requiring the presence of a human supervisor.

The first stage of this study has been discussed in Luperto et al. [27], where we focused on the feasibility of the NeuroPsychological Test scenario in a controlled setting. In this work, we extend the evaluation to a home setting, in which the SAR administers the tests autonomously.

The two scenarios defining the cognitive-monitoring module were tested in the experimental campaign of the MoveCare project, which involved a complex and heterogeneous system designed for at-home assistance, engagement, and monitoring of older adults. The works of Luperto et al. [29, 30] describe the global results of such a campaign, beyond the scope of this paper, while the work of Luperto et al. [31] describes the architecture of the entire system. Differently from such works, this paper presents previously undisclosed results that focus specifically on the at-home testing of the NeuroPsychological Test and Spot Questions scenarios.

2 Related Work

SARs were first defined by Feil-Seifer and Matarić [15] as the intersection of Assistive Robots and Socially Interactive Robots, whose aim is to provide assistance through social interactions.

The benefits of using SARs for remote health monitoring functionalities [18, 24] are widely recognised [1, 43], and have been increased by their integration in Ambient Assisted Living (AAL) arrangements. AAL is becoming fundamental to answering the needs of "ageing in place". The technological advances and affordability of smart sensors, as well as the consolidation of software platforms for their integration with robots, are enabling the creation of high-tech living environments capable of actively improving the life quality of older adults. An example of such integration is given by Bellotto et al. [4], where a set of smart sensors and a companion robot contributed to provide monitoring of older adults affected by MCI. The objective of Bellotto et al. [4] was to develop and demonstrate the general feasibility, effec-

tiveness, and acceptability of a system composed of three robotic platforms, integrated into intelligent environments, which actively worked in real conditions and cooperated to favour independent living. Another example is the work of Garzo et al. [20], where a robot was developed to improve the well-being of older adults by providing connectivity, therapy reminders, and fall detection by being interfaced with a set of home devices. The work of Fischinger et al. [18] represents another instance of testing a SAR in a short-term human-robot interaction. In this case, the robot was providing services like fetching and delivering objects, giving reminders, entertaining the user, and detecting falls but it did not carry out cognitive monitoring.

Among the different platforms available as SARs, telepresence robots are specifically used to enhance independent living by providing a direct line between older adults and their carers. A well-known example of a telepresence robot is the Giraff robot, used in works such as Coradeschi et al. [8] to achieve monitoring of older adults in their daily life activities. This robot uses a Skype-like interface to allow relatives or caregivers to virtually visit an older person in their home. The feasibility and acceptability of a long-term deployment of a telepresence robot in the houses of older adults were studied in Fiorini et al. [17], where they found that participants in the study were positive towards the proposed technology. Therefore, the use of SARs is a promising approach for home-based monitoring of cognitive functionalities with the final aim of detecting early alerts. In this context, the Guardian Ecosystem Ciuffreda et al. [7], starting from a user-centred and value-sensitive co-design approach, provides interesting insights about the key user requirements for such robots to be actively used in real-world deployments. Similar remarks arise from the eWare Project, described in Amabili et al. [2]. While most works focus mainly on older adults, as they are the main users of these systems, the work of Amabili et al. [2] investigates how dyads of caregivers and older adults can benefit from a platform comprising a SAR and an IOT system.

SARs have already been included in rehabilitation and mental healthcare applications. In Rabbitt et al. [35], SARs deployed for people suffering from dementia are presented and analysed. However, their at-home use for such domains has not been exhaustively investigated yet. In particular, SARs could play a fundamental role in administering tests in a familiar home environment and overcoming the barriers to early-stage detection of cognitive weakening. The idea of using robots for cognitive monitoring is strengthened by works like Mann et al. [32] stating that users respond more positively to robots than tablet computers delivering health care instructions.

Nonetheless, the validity of such a robotic psychometric approach is still an open question. While there are works that employ SARs to deliver cognitive games to older adults, and that demonstrate they are accepted by their end users for such tasks [21], little effort has been directed to cognitive monitoring and cognitive assessment, especially in an at-home setting. Varrasi et al. [41] shows how SARs can provide advantages for early MCI detection thanks to their capabilities of administering specific standardised tests and automatically recording the answers for further analysis while engaging the user. The study stems from an experimental setting where several healthy adults completed the "Montreal Cognitive Assessment" (MoCA) [33] under the guide of the humanoid robot Pepper in a laboratory setting. Another work looking at the feasibility of cognitive assessment for older adults is Sorrentino et al. [38]. Differently from Varrasi et al. [41], they pay particular attention to the role of personality and emotions in the test delivery (the test chosen is the Mini-Mental State Examination [19]). Their system was tested by 11 older adults interacting with the robot only once and in a controlled laboratory setting.

Our work extends the above literature by focusing on the at-home employment of SARs for MCI early detection. As part of an experimental campaign run for the MoveCare project, our study is corroborated by data gathered on the field, where robots and users shared the same domestic environment over a period of at least 10 weeks.

3 SAR-Based Cognitive Monitoring

Clinical cognitive assessment is primarily supported by neuropsychological tests under the supervision of a clinician [39]. The tests follow standard and validated protocols that rule the supervision of the test administration, and that define the evaluation metrics. The tasks that are performed during the tests can be various (based on paper and pencil, requiring verbal interactions, etc.) and the metrics used for their evaluation can be quantitative and qualitative, the latter being more subjected to the clinician's judgement.

Keeping this context in mind, we investigated two different modalities to perform at-home neuropsychological assessments with a SAR supervising their administration.

3.1 Digitalised Neuropsychological Tests

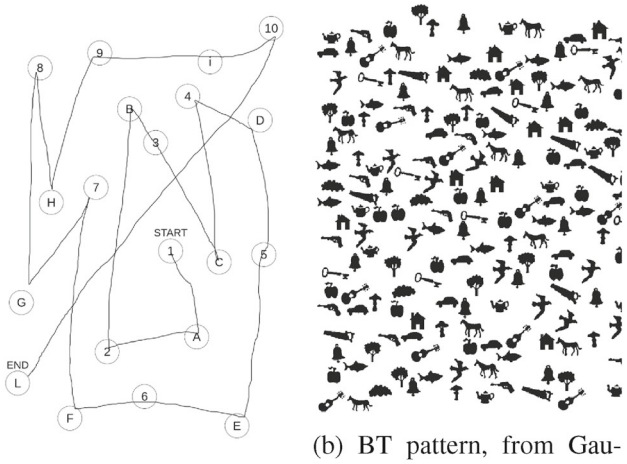
The technical requirements that guided the choice of the tests, identified with the help of clinicians from the hospital Maggiore Policlinico in Milan (Italy), were the following:

- the digital neuropsychological tests should be easy to interact with, and conveyed by standard consumer technology (like tablets),
- the SAR's supervision (as required by the test protocol) should be structured by simple actions and triggered by clearly recognizable events,

- the evaluation metrics should be primarily quantitative and allow automated computation,
- the SAR's intervention (as required by the test protocol) should have a minimal impact on the test's evaluation metrics, allowing an easier interpretation of the results.

Guided by the above criteria, our attention narrowed down to paper-and-pencil tests, since the digital implementation of this input modality can take advantage of many robust development technologies. Our final choice fell on the Trail Making Test (TMT) [36] and the "Bells" Test (BT) [22], exemplified in Fig. 1. The first one, TMT, aims to test visual attention and task switching, while the second one, BT, aims to evaluate visual attention and visual neglect.

The tests are provided through a 10.1-inch tablet where a dedicated application, called Test-App is installed, a capacitive pen, and a stand to support it (Fig. 2).



(a) TMT execution (b) BT pattern, from Gauthier et al (1989)

Fig. 1 The digital cognitive tests of the proposed system. Taken from Luperto et al. [27]

Fig. 2 The tablet used in our project.

Trail Making Test (TMT)

The TMT requires drawing a continuous line traversing a number of symbols according to the sequence suggested by their labels, e.g., 1, A, 2, B, . . . , 10, L. The test is performed in two sessions called TMT-A and TMT-B. In TMT-A, symbols are labelled with numbers 1, 2, . . . , 25; in TMT-B, labels interleave letters and numbers 1, A, 2, B, . . . , 10, L. The two sessions are performed in sequence, starting with the TMT-A. The layout is similar to the original paper-based TMT but, given the reduced dimensions of the tablet screen (the original test is performed on an A4 paper), the number of targets was decreased to 20 as proposed by Dahmen et al. [9] and Fellows et al. [16].

Directions on how to execute the test are provided with a simple tutorial, where a limited number of symbols are displayed, and where the first symbols are connected by the user during training [23]. The evaluation metrics should indicate whether the symbols have been traversed in the correct order and the amount of time required to complete the test. Errors made during the execution of the test are signalled, but not corrected. Figure 1a depicts a typical result of the test.

Bells Test (BT)

The "Bells" test requires localizing and circling or marking, within a time budget, an icon shaped like a bell mixed up with different icons called "distractors". A portion of the layout of our digital implementation is shown in Fig. 1b and contains 35 target icons and 280 distractors as proposed in Gauthier et al. [22]. The graphical elements have been scaled to fit the tablet's screen size. As in the case of TMT, a simple tutorial with a limited number of icons is provided, where the test protocol is explained. Errors made during the execution of the test are not corrected nor signalled. The test is considered completed when the user declares that all the bells have been identified. The evaluation metrics include the number of bells correctly marked, the number of errors, the distribution of the bells that have been found/missed, and the time required to complete the test.

TMT and BT are tests typically included in batteries for MCI detection and have been proven useful for assessing early stages of it [42]. This is particularly true for the TMT, of which several digital versions have been proposed [16, 26]. These implementations, however, are mere transpositions of the original paper-and-pencil tests over a digital device and are designed to be performed in a clinical site under the supervision of a clinician. Instead, the digital versions of TMT and BT employed in the work we are presenting in this paper feature supervision carried out by a SAR (see Sect. 4.1) and are meant to be administered at home.

Table 1 Example of spot questions asked by the SAR deployed in our system.

Question type	Question
Episodic Memory	To be able to offer you some new activity, can you tell me, please, whether you have played cards in the last 3 days?
Apathy	On a scale from 0 to 5, are you more tired today than usual?
Temporal Orientation	I need your help to set the system configurations: what day of the week is today?
Confabulation	What was wearing the person seated next to you on the bus last time?

Moreover, the tests' digitization allows for automated computation of their evaluation metrics. Additional details on the implementations of our digital tests can be found in Lunardini et al. [26] where their validity (providing predicting capabilities comparable to those of their traditional counterparts) is demonstrated.

3.2 Spot Questions

SQ have been designed in collaboration with neuropsychologists from the Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano (Italy) and Gerencia del Area de Salud de Badajoz, Junta de Extremadura (Spain). They test different cognitive domains that are of crucial importance when it comes to age-related cognitive alteration: episodic memory, apathy, temporal orientation, and confabulation [3, 40]. Examples of such questions are reported in Table 1.

SQ in our cognitive monitoring module was referring to either common knowledge (i.e., temporal orientation through questions on the day of the week or the exact date) or activities performed by the user in the recent past. For this second type of question, we resorted to asking the user details about activities performed within the MoveCare platform, which were logged by the system and known to the robot (see Table 1).

The questions we selected can belong to two ways of testing cognitive alterations: "recovery" (i.e., stimulating a voluntary and direct process to extract information from memory) and "recognition" (i.e., stimulating a voluntary but indirect way to recover stored information). To assess the recovery ability of the subjects, they are asked to provide the information through "Free Recall", i.e., they are allowed to answer freely to the robot's question, without a set of predefined answers to choose from. On the contrary, to assess recognition they are usually asked to say what is the answer among several options (which can include false information). An example of a "Free Recall" SQs are those for Episodic Memory (Table 1).

The key feature of SQ is that they enable longitudinal assessment by asking several questions to the user on a weekly basis and over a timespan of several months.

4 Proposed System

Our setup exploits a SAR called Giraff-X to provide supervision during the execution of TMT and BT, and to collect the answers to SQ. We developed Giraff-X, a modified ROS-based [34] version of Giraff mobile robot [8], as a part of the MoveCare project where new functionalities (like autonomous navigation) enable its use as a fully-autonomous assistive robot [30] to support older adults' independent living.

Within MoveCare, the robot is the main actor of an AAL domestic system composed of an Internet of Things (IoT) network of sensors, used to monitor the user activity and to provide information to the other modules, a tablet application called Community-Based Activity Center (CBAC) [28], used to provide stimulation, socialization, and cognitive assessment to older adults through online entertaining activities, and a Virtual Caregiver (VC), a cloud component acting as platform coordinator that, among other things, is in charge of scheduling and issuing the interventions performed by the robot, including NPT and SQ).

A full list of the robot interventions can be found in Luperto et al. [29], while a functional description of the system is presented in Luperto et al. [31]. For the scope of this paper, we focus on NPT and SQ interventions.

4.1 Integrating Giraff-X and the Neuropsychological Tests

NPTs are obtained through the integration of two key elements of the clinical practice: the administration of cognitive tests and its real-time supervision by a clinician. We provide the first by developing a tablet-based version of the selected standard clinical tests, while the real-time guidance is carried out by the SAR.

The tests' execution develops in two steps. During the first step, the `setup` (Sect. 4.1.1), the robot needs to find the user to inform them that it is time to complete the tests. It then instructs them to get the tablet on which the Test-App is installed and to find a comfortable place where they can carry out the tests. The second part of the intervention, the `execution`, requires the robot to locate the user once again and start with the test steps described in Sect. 4.1.2. As

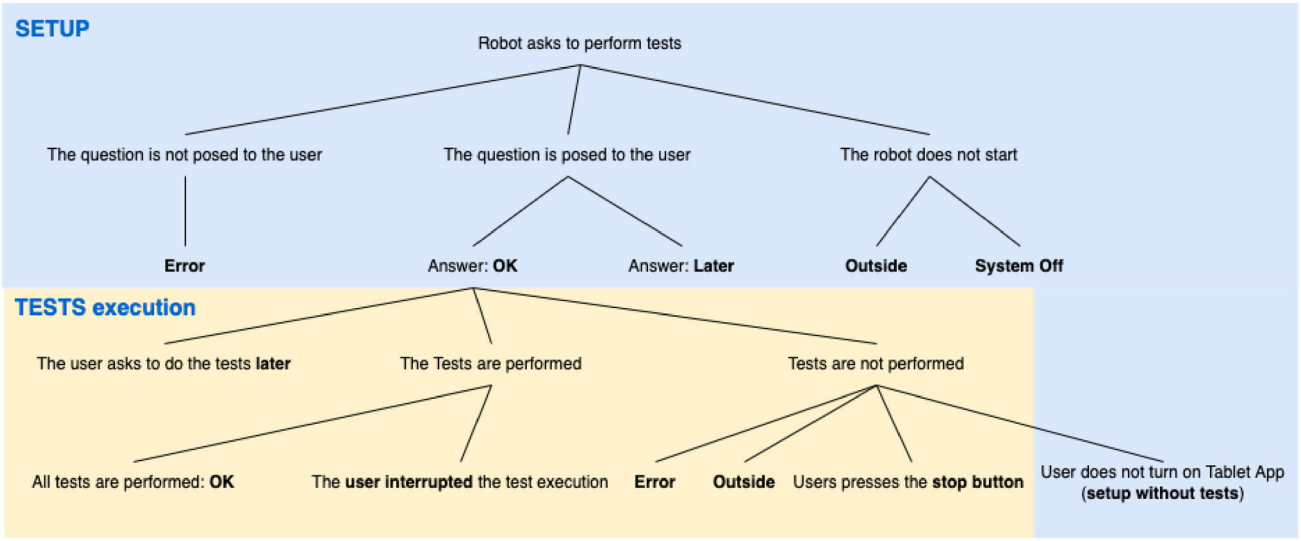


Fig. 3 The possible outcomes of the execution of a NPT scenario

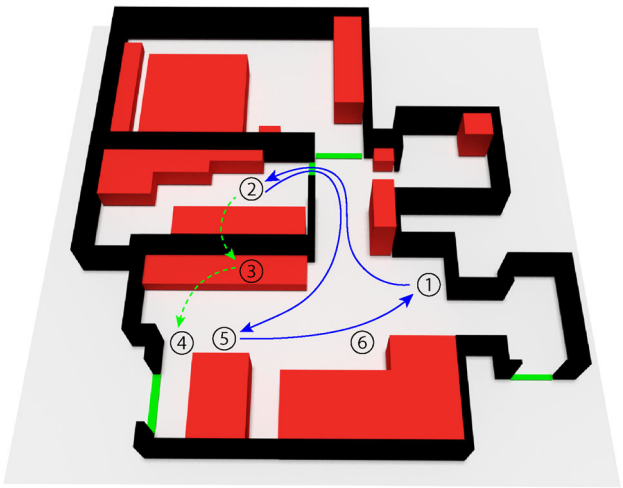


Fig. 4 An example of the NPT scenario. We show in blue the trajectory followed by the robot and in green the actions performed by the user. (1) the robot, while charging at the docking station, receives from the VC an intervention request to perform NPT; (2) the robot finds the user in the kitchen, asks them to perform the tests, and describes the procedure of tests execution; (3) the user agrees to perform the scenario, gets the tablet, and moves to a table. (4) the user sits at the table and turns on the Test-App on the tablet; (5) the robot finds the user again; the test execution starts, and the robot oversees the execution of the tests; (6) the robot returns to the docking station

different events can happen during test execution, we provide an overview of the possible development of the tests in Fig. 3, which are also highlighted in the text with a `different font`.

4.1.1 Preparation Phase

This step aims to prepare the execution of the tests, which require a specific setup, i.e., to have the user sitting at a table, with the tablet (charged), and with their glasses (if needed). The user is then informed that they need to pay attention to the robot for approximately 20 min.

Figure 4 presents an example of the full scenario, where steps 1 to 4 refer to the Preparation Phase. When the robot receives the request to perform the intervention from the Virtual Caregiver (VC) (1), it starts to search for the user in the house. In this illustrative example, the user is found in the kitchen (2). The robot asks the user if they are willing to perform the scenario, explaining that it requires approximately 20 minutes and that it should be performed sitting at a table, with the tablet charged, and with the glasses on (if needed). This first part of the scenario is called `setup` and is concluded successfully upon receiving the answer from the user; if the user answers positively (label `ok`), the robot waits for the next step to start; if the user answers that they do not want to perform the NPT at that moment (answer `later`), the answer is recorded, and the scenario is concluded. The VC will reschedule the scenario for another day.

In Fig. 4, we show the case when the user answers positively to the request of the robot. The user finds the tablet (3), turns it on, and sits with their glasses at the table in the living room (4). Then, the user turns on the Test-App, the event

that triggers the second step of the scenario: the robot starts looking for the user again and eventually finds them (5). The robot then approaches the user in the new location, and asks for confirmation that they are ready to carry out the tests; if the answer is positive, it starts guiding the users through the NPT intervention; if the answer is negative (answer `later`), the scenario is postponed to another time.

4.1.2 Tests Execution

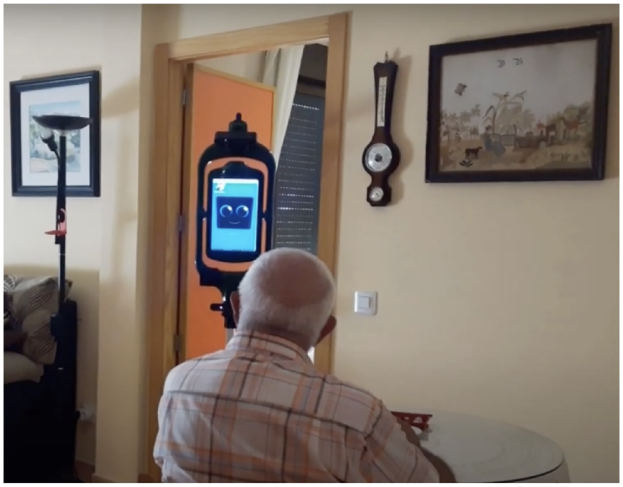
The second step starts after the user launches the Test-App, and the robot has successfully re-located them. The main role of the robot is to explain the tests' protocol by showing a small demonstration and by assessing its understanding by the user. Once the actual tests are completed, the robot thanks the user and returns to its docking station. The data acquired during the tests are saved and sent to the VC by the Test-App.

There are seven possible outcomes to the scenario:

- `ok`: the user successfully completes all scheduled tests.
- `user interrupted`: the user interrupts the tests' session in between the tests, after having completed only a subset of them (e.g., only the TMT).
- `button stop`: the user interrupts the session within a non-completed test (e.g., due to an unexpected event), by pressing the robot emergency button.
- `error`: the session is interrupted by an external event (e.g., loss of internet connection).
- `outside`: tests not performed because the user left the house.
- `setup without tests`: the setup phase is completed, but the user never started the execution phase (the Test-App on the tablet is not started by the user).
- `later`: the user postpones the execution of the test by answering negatively in the setup or execution phase; the session is considered not completed and marked with `later`, to be rescheduled for another day.

During the tests, and in between two consecutive ones, the robot requests confirmation from the user to determine if they want to continue or stop the session.

It must be noticed that the robot remains silent for the duration of the tests unless an error or an idle situation is detected. For the proper administration of these tests, the robot does not give too much feedback to the users while they are completing the task to not introduce biases or distractions, as indicated by the tests' protocol described in Sect. 3.1 and provided by our clinical partners. Fig 5 shows a user executing the TMT, where the robot is supervising the test execution from a side of the table.



(a) The robot asks a user to perform the tests.

(b) The user performs the test on the tablet under the supervision of the robot

Fig. 5 A user executes a neuropsychological test under the supervision of a Giraff-X robot

4.2 Integrating Giraff-X and Spot Questions

SQ encompass different cognitive domains of great importance for age-related cognitive alteration. Usually, this type of assessment is performed by having other subjects that live in close contact with the user, posing the questions. With our work, we investigated whether the same dynamics can be replicated by a SAR. Our at-home setting, where a SAR interacts with older adults regularly, offers many opportunities for posing these questions, as they can be integrated into, and refer to, the daily activities of the users.

Their integration with the SAR is based on randomized interventions where, during the day, users could be asked some questions by the robot. Each question was chosen randomly among a list of 24 pre-defined ones (see Table 1 for some examples), covering different cognitive domains. First, a question domain was randomly selected and then one question from that domain was extracted. This question was then marked as "used" and excluded from the next randomization.

Table 2 Differences between the exploratory study of Luperto et al. [27] and the main experimental campaign for the NPT scenario described in this paper

Characteristic	Exploratory study	Main experimental campaign
Cognitive monitoring	Cognitive tests	Cognitive tests, spot questions
Number of participants	16	10
Expert presence	In the room in case of need	Absent
Robot familiarity	Interaction for the duration of the study (approx 20. min)	Sharing the apartment for at least 10 weeks
Human–robot interaction	Via voice only	Via voice and green/red buttons
Setup	Tablet already on the table, robot static already placed close to the table	Robot needs to find the user twice in the house to prepare the scenario and carry it out
Questionnaire answering	Immediately after the study, interview with the experimenter	After the whole system is uninstalled from the house, alone

During the execution of the SQ intervention, the robot first searches for the user, starting from the estimated user location. Once the user is found, the robot approaches them and asks a question. All questions follow the same structure: "Hello <*user-name*>, I need to check some data that will help me give you a better service. Do you have time to answer a question for me now?". If the user accepts, the robot asks the question and records the answer (label `answer`); if the user does not answer or declines (label `later`), the robot notifies the VC that will reschedule the intervention. At the end of the interaction, the robot always thanks the user for their input with sentences like: "Thank you very much. I have updated my data. It is a pleasure talking to you.". No additional feedback was given, as there was no correct or wrong answer to most of the questions. Moreover, we did not want to cause distress to participants in case they made a mistake. The answers from the users are converted into text and stored on a server data repository in the cloud, becoming available to the VC for its analysis.

4.3 Preliminary Campaigns

Before the deployment of the system in the users' apartment, where it was going to work autonomously, we evaluated the NPT scenario on two different occasions: an exploratory study where the functionalities for the execution of the tests were under scrutiny [27], and an on-the-field preliminary campaign in the presence of technicians. With the exploratory study, we collected insights on the usability and acceptability of the execution phase only of the scenario in a controlled environment. We used these user-given insights to complete the development of the NPT. During these tests, the robot was performing a simpler form of intervention with respect to the final one used in the main experimental campaign (Sect. 5), where only the execution phase was evaluated. The main differences between these tests and those in the main experimental campaign are listed in Table 2. During the on-the-field preliminary campaign, the system was set up within an apartment of an assistive living facility and we asked 5 older adults to perform the full set of scenarios, including NPT and SQ, and to qualitatively evaluate their feasibility. We tested the full NPT scenario, setup, and execution, with the presence of caregivers and technicians assisting the users. The positive evaluation allowed us to integrate both NPT and SQ in the bigger framework before the main experimental campaign. A video of the NPT as performed during these tests is available online.[1]

5 Experimental Evaluation

The main experimental evaluation was performed within the experimental campaign of the project MoveCare, where the entire MoveCare system, of which NPT and SQ are two main functionalities, was installed and tested inside the private apartments of end users. During this period, the robot performed interventions with full autonomy, by moving inside the apartment. After the initial installation, nobody was present to oversee the robot's behaviour or to help the users. However, we gave the possibility for the users to ask for explanations on the behaviours of the robot, or to signal an inconsistency. The experimental campaign consisted of two rounds: the first took place from September to December 2019, and the second from January to April 2020. After the first round, a few users were asked to continue and participated also in the second round.

A total of 14 older adults were selected for the experimental campaign. They were over 65, did not suffer from any cognitive impairment, and lived independently on their own. These users were from two different regions: 7 were residing in Milan, Italy (we shall refer to this group as `ITA`),

[1] https://www.youtube.com/watch?v=Xla4Ue3hCrQ&ab_channel=MovecareProject.

while the other 7 were from Badajoz, Spain (group ESP). We report the data extracted from 10 of the 14 participants of the field study (average age of 75.3) that performed the NPT and SQ scenarios. We include in our current analysis only those participants who actively used the system for at least 10 weeks: 5 ITA and 5 ESP. For further details about the experimental campaign, please refer to Luperto et al. [29, 30].

5.1 Test Scheduling Within the Experimental Campaign

The system was designed to schedule a scenario to be performed by the robot until the desired target was reached. If a scenario was scheduled but not executed, it was rescheduled for another day, intertwined with the other tasks that the robot needed to perform.

NPTs were supposed to be performed twice: once at the very beginning of a campaign round, used to establish a baseline for the participants, and once at the end of that round. Those users who participated in both rounds were requested to perform the NPT scenario three times (at the beginning of round 1, between the two rounds, and at the end of round 2). The three tests were scheduled by the VC in sequence (TMT-A, TMT-B, and then Bells) and the VC stopped requesting the execution of the NPT when all three tests were completed.

SQ were scheduled differently every week. An example of SQ scheduling is the following:

1. Week 1: 1 question per day for 4 consecutive days.
2. Week 2: day 1 and 3, 2 questions; day 5, 1 question
3. Week 3: 4 questions on the same day
4. Week 5: 1 question per day for 5 consecutive days.

While the users did not know specifically that their answers to the SQ would be used for monitoring purposes, they were aware that the robot would ask them to perform some tasks during the experimental campaign to collect data that could be used in the future as an additional monitoring tool.

The VC was in charge of deciding when to start an intervention. As each type of intervention has its own schedule to follow, and a priority, the VC decides which intervention should be performed during a given day following the overall scheduling of all interventions. To start an intervention, the VC is also subject to a set of constraints designed to not burden the users: interventions should be performed during daytime when the user is at home, interventions should be spread during the whole day and not performed altogether. One of our main objectives was to design a system that could adapt to the needs and lifestyle of the users, and not pretend that the users adapt to it.

Not to overload the users, we set a constraint that only from 2 to 5 robot interventions per day could be executed when the user was available. The VC, using IoT data, was able to start an intervention only when the user was at home. However, during the experimental campaign, it could happen that the user left the house after the intervention had already started. On these occasions, the robot immediately cancelled the intervention and returned to the docking station for charging. We labelled such events as outside.

Due to these constraints, and as NPT and SQ were intertwined with the other interventions, the robot ultimately performed fewer interventions than initially planned. We assigned to NPT a high priority, and to SQ a low priority. As we explain in Sect. 7, this reduced the amount of SQ that was asked to the user.

5.2 User Feedback and System Inconsistencies

As described in Luperto et al. [30], the users' feedback was particularly important during the long-term deployment of the robots.

A user, ESP-1 experienced an error in the system. Although they had performed the NPT correctly, the robot kept asking to perform NPT in the following days, due to a scheduling malfunction. However, as the robot instructed them that the tests should be performed only every few months, the user replied later to the intervention and signalled the issue as an inconsistency in the system. This fact allowed us to identify and fix the issue, without compromising the execution of the protocol.

During the first days of the experimental campaign, participants reached out to us to complain about the weird questions that the robot was asking them [30]. They were referring to the *confabulation* type of SQ which asked questions that were not linked to anything that participants actually experienced (see Table 1 for an example). Not being able to answer those questions was considered frustrating for the majority of our participants. For this reason, we decided to eliminate the confabulation questions from the SQ scenario. We discuss this fact in Sect. 7.2.

6 Results

6.1 NPT Execution

The outcome of the executions of the NPT scenario is summarised in Table 3.

The robot started a session of NPT 62 times (performed and error summed up in the table). In 8 cases (resulting in error) the robot could not start the interaction with the user for several reasons including the presence of an obstruction, network connectivity loss, and unsuccessful localization

Table 3 Number and outcome of each time the NPT scenario is performed, per user

Users		ITA-1	ITA-2	ITA-3	ITA-4	ITA-5	ESP-1	ESP-2	ESP-3	ESP-4	ESP-5	ITA	ESP	ALL
setup	Performed	6	4	6	4	6	15	5	4	2	2	26	28	54
	ok	5	3	6	0	1	8	2	1	2	2	15	15	30
	Later	1	1	0	4	5	7	3	3	0	0	11	13	24
	Error	1	0	0	0	2	3	0	2	0	0	3	5	8
	System Off	0	0	0	0	1	1	1	0	1	0	1	3	4
	Outside	1	0	0	8	1	5	1	1	0	0	10	7	17
setup without tests		1	0	2	0	0	3	0	0	1	1	3	5	8
tests	Performed	3	3	2	0	1	3	0	1	1	1	9	6	15
	ok	3	2	2	0	1	3	0	0	1	0	8	4	12
	User Interrupted	0	1	0	0	0	1	0	1	0	1	1	2	3
	Later	0	0	1	0	0	3	0	0	0	0	1	3	4
	Error	0	0	0	0	0	0	2	0	0	0	0	2	2
	Button Stop	1	0	1	0	0	1	0	0	0	0	2	1	3
	Outside	0	0	0	0	0	0	0	0	0	0	0	0	0

of the user. However, not all the remainder 54 interactions led to completing the tests. In 24 cases, the user did not accept to start the test answering `later` to the robot. Two users (ITA-4 and ESP-2) always answered negatively to the robot's request to perform NPT. Of the 30 cases (49%) in which the user accepted to carry out the NPT, only 12 completed all three tests in a row (22%). In 3 cases, the users did not complete all the tests together and closed the Test-App after completing only the TMT-A and/or TMT-B (leading to a `user interrupted`). In 8 cases, the user did not pick up the tablet and did not start the Test-App (the tablet was out of battery, or the user got distracted by other events resulting in `setup without tests`). In 3 cases, the user pressed the emergency button that stopped the robot and halted the intervention after having launched the Test-App (resulting in `button stop`). In 2 cases, a system error occurred and the robot returned to the docking station (resulting in `error`). In 4 cases, after they started the app, the user said to the robot that they were not ready to perform the tests (resulting in `later`).

In total, the tests of NPT were performed successfully 15 times (12 completed and 3 interrupted halfway). The `setup` phase of the test, required 176 s on average (σ = 109 s); the shortest intervention required 68 s while the longest one took 421 s. The time needed for the intervention is mostly due to the robot trying to locate the user. On average, the robot searched in 1.38 rooms.

The execution phase of the tests required, on average 15 min and 6 s (σ = 459 s); the shortest intervention required 10 min and 12 s and the longest one 38 min and 32 s. To locate the user after the `setup` phase, the robot searched in 0.38 rooms, on average; this is because the user was often already nearby or in front of the robot when they launched the app on the tablet. Therefore, the time reported is almost entirely due to the interaction time spent during the execution of the tests themselves.

Figures 6 and 7 show the trajectories of the robot while executing the NPT intervention for ITA-1 and ITA-2. It can be seen that during the initial `setup` phase of the tests, the robot needed to travel across multiple rooms before finding the user. After the `setup` phase is completed, the robot heads toward the table at which the user is seated.

Figures 8 and 9 show the results of the first TMT test performed by users ITA-1 and ITA-2. It can be seen how users managed to perform the test successfully, despite having some difficulties (due to the complexity of the task) during TMT-B, which is indeed difficult, as shown by the various marks made on the numbers.

6.1.1 SQ Execution

Table 4 shows the results of SQ interventions. Of the 251 SQ asked the users, 119 times the robot could not start the interaction with the user (43% of the cases, the sum of the intervention non completed and non-performed in Table 4). The robot could not complete the interaction due to unexpected events 72 times: 51 for a system malfunction (16 times for a connectivity loss, 29 times because the pathway to the user was blocked, and 6 times the robot was not able to locate the user); 21 for a misinterpretation from the HRI module; 20 times because the user was *outside* so the intervention did not even start; 27 times the system or the robot had been turned off by the user. Therefore the robot started the interaction with the user 132 times (57% of the cases) and it got an answer to the spot question 89 times (35% of the cases). In 43 times, 21% of the cases, the user answered the initial

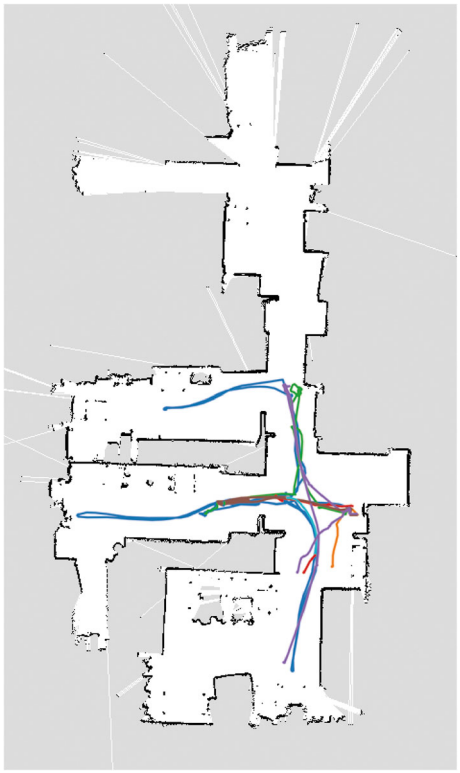
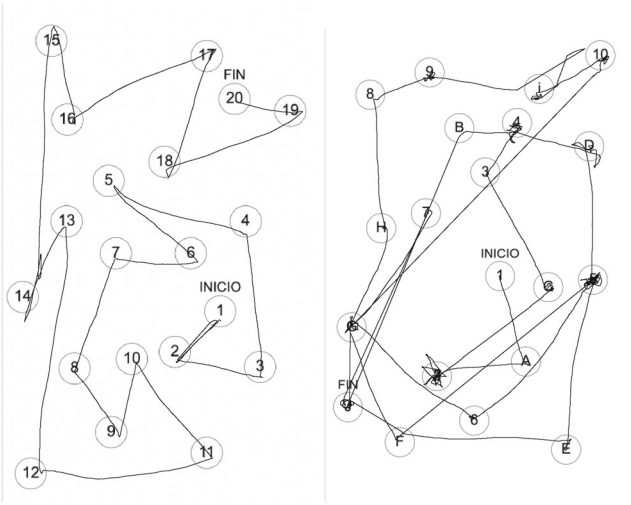


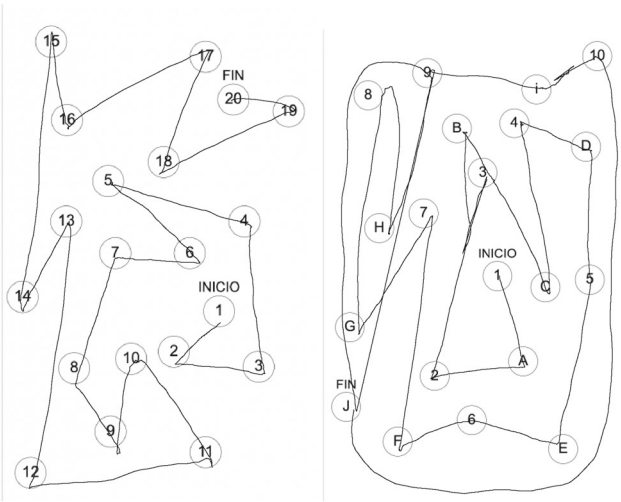
Fig. 6 The trajectories performed by the robot of ITA-1 to perform NPT

Fig. 7 The trajectories performed by the robot of ITA-2 to perform NPT



(a) Execution of TMT-A. (b) Execution of TMT-B.

Fig. 8 TMT tests performed by user ITA-1 under the supervision of the robot



(a) Execution of TMT-A. (b) Execution of TMT-B.

Fig. 9 TMT tests performed by user ITA-2 under the supervision of the robot

question ("Do you have time to answer now?") posed by the robot, completing the interaction but answering that they did not want to answer the question (outcome later).

Some users answered the robot most of the time (e.g., ITA-1 answered over 70% of the questions), and some others answered the robot a few times (ITA-4 answered 6% of the questions).

6.2 User Questionnaires Evaluation

At the end of the experimental campaign rounds, participants were asked to answer a set of questionnaires related to their experience. The questionnaires were administered by

Table 4 Number and outcome of each time the spot questions (SQ) are asked, per user

Users		ITA-1	ITA-2	ITA-3	ITA-4	ITA-5	ESP-1	ESP-2	ESP-3	ESP-4	ESP-5	ITA	ESP	ALL
Scheduled		22	26	17	30	41	42	15	30	11	17	136	115	251
Completed	Performed	19	17	11	10	24	18	7	14	8	4	81	51	132
	Answer	16	8	5	2	18	16	5	11	5	3	49	40	89
	Later	3	9	6	8	6	2	2	3	3	1	32	11	43
Not completed	HRI error	2	2	2	0	4	6	1	1	1	2	10	11	21
	Robot error	0	6	2	0	4	12	7	14	2	3	12	39	51
Not performed—system off		0	1	0	10	3	6	0	0	0	7	14	13	27
Not performed—user outside		1	0	2	10	6	0	0	1	0	0	19	1	20

Table 5 NPT user questionnaire. Given the limited number of answers we report the scores for all the participants who actively completed the tests, the median scores (M) per country, and and the total median score per question

	Question	ITA-1	ITA-2	ITA-3	ITA-5	ESP-1	ESP-3	ESP-4	ESP-5	M ITA	M ESP	M ALL
Q1	The proposed tests are interesting and stimulating	5	4	4	3	4	4	4	4	4	4	4
Q2	I easily understand the test instructions	5	4	5	3	4	1	4	1	5	2.5	4
Q3	I easily understand how to execute the tests	5	5	5	3	4	1	4	1	5	2.5	4
Q4	The system reacts rapidly to my moves	5	4	4	4	3	3	4	2	4	3	4
Q5	It is easy to execute the tests using the tablet and the pen	5	5	5	4	4	2	4	1	5	3	4
Q6	I don't need any kind of help to perform the tests	5	4	5	2	3	1	4	1	5	2	4
Q7	I enjoyed the graphical interface of the number and letter tests	5	5	2	3	4	4	5	4	3	4	4
Q8	I enjoyed the graphical interface of the bell test	5	5	4	3	4	4	5	4	4	4	4

a professional caregiver during a meeting where the users also performed cognitive assessments to measure changes that happened during the campaign rounds. We administered to each user 21 questionnaires, for a total of 150 combined questions. Three of the questionnaires we used were designed to evaluate the overall experience of the participants in the campaign, while six other questionnaires evaluated a single component of the platform. An example of the former type of questionnaire is the System Usability Scale (SUS), while an example of the latter type is the Robot Acceptance Questionnaire, which evaluates the robot; full results of these questionnaires are described in Luperto et al. [29]. In addition, 12 targeted functionalities were evaluated with an ad-hoc questionnaire. As requested by the professional caregivers involved in the project and the design of the questionnaires, and in order not to burden the users with too many questions, only a subset of the components and functionalities of the project were directly evaluated through an ad-hoc questionnaire. All questionnaires were based on a 1-to-5-Likert scale, with 1 for "Strongly Disagree" and 5 for "Strongly Agree".

The NPT scenario is one of the 12 functionalities evaluated through a questionnaire. Its questions, and the collected answers, are reported in Table 5 and evaluate the usability and acceptability of at–home NPT.

In total, only 8 participants out of 10 actively completed at least a session of the NPT scenario and answered the related questionnaire, 4 from Italy and 4 from Spain.

Overall, considering the general median scores, we can say that our system was positively evaluated and its usefulness was understood by prospective users. By investigating the answers by country we can see that ITA users were generally more satisfied with neuropsychological tests. The main difference can be noticed in Q2, Q3, and Q6 of Table 5 where it is clear that ESP participants had more difficulties in understanding and executing the tests under the robot supervision.

By looking at Q5 and Q7 in Table 5, it seems there is still some margin of improvement for the graphical interface. One of the reasons behind this could be the fact that using the pen on the tablet was still not considered completely user-friendly, in particular for the TMT tests. A similar remark was signalled during the exploratory study described in Sect. 4.3 suggesting the need to find a better tool to allow them to complete such tests. However, another reason behind this could be the fact that those tests are challenging to perform by design, and the look and feel of the interface is strictly limited by the test protocol (i.e., it should be a white page with only a set of numbers and letters to connect, as in Fig. 1a).

Conversely, the SQ were not evaluated with a questionnaire. This is due to the transparent nature of the questions: being integrated with the daily activities of the participants who were unaware that they were part of a wider scenario.

This kind of longitudinal assessment, intertwined with the daily activities of participants and with the other interventions that the robot was programmed to do, could have contributed to making the full MoveCare platform invasive and frustrating, as it involved asking many questions during a week (see Sect. 5.1). However, from the subjective assessment of the entire campaign and from the evaluation of the robot through ad-hoc questionnaires [29], we can see that our SAR-based system was not considered intrusive by participants despite the high number of interventions performed for several weeks, that included SQ.

This is a positive result for the feasibility of a longitudinal assessment based on cognitive monitoring data collected in a home setting: it is possible to carry it out without frustrating the users.

7 Discussion and Lessons Learnt

The NPT and SQ scenarios were chosen to investigate the capabilities and limitations of SARs carrying out cognitive monitoring duties while operating long-term with their users.

Results indicate how NPT could indeed be administered at–home through the supervision of a SAR, while also showing some open points for future investigation.

Single sessions of the NPT scenario were performed without supervision by 8 of the 10 users with success and were positively evaluated by the users, who signalled that the proposed tests were interesting and stimulating (Q1 Table 5). These results provide empirical evidence for the feasibility and acceptability of neuropsychological tests performed in the homes of the users without human supervision but mediated by a SAR.

The positive evaluation provided by the end users is particularly important if we consider the fact that the execution of the NPT scenario, even with the presence of a clinician as a supervisor, could be a challenging task for older adults, as those tests are designed to be difficult to perform. Further support for this indication comes from the answers to Q6 of Table 5, which shows how most users reported that they did not need any kind of help to perform the tests. The fact that the scenario required a long time to be performed did not have a negative impact on the answers to the questionnaires; however, it did impact the number of times the users actually performed the scenario, as we can see that they often opted to reschedule it by answering `later` to the requests of the robot.

We observed a strong discrepancy both in terms of number of times the scenarios were performed and in terms of answers to the questionnaires when comparing the two experiments performed in ITA and ESP. Results indicate that users from ITA perceived NPT as easier and also provided a higher evaluation, while ESP users experienced much more difficulty (e.g., as shown in Q5 Table 5). Many aspects could have contributed to this result: the differences in the organisation and logistics of the experimental campaign in the two countries and some speech recognition problems in the speech interface (especially in the Spanish version) due to the fact that speech interfaces for older adults are yet to be perfected.

This difficulty in using the robot speech interface emerged also in Luperto et al. [30], where the full questionnaire on the usability and acceptability of Giraff-X is presented. There too, users expressed some trouble understanding the robot requests, possibly due to the same sub-optimal speaking ability.

7.1 Users' Availability and Test Scheduling

One important factor that emerged from our experimental campaign is the importance of the user *availability* during long-term robot deployment. With user availability, we describe three different yet highly correlated conditions:

(a) the user is at home;
(b) condition (a) is verified, and the user is not busy;
(c) condition (b) is verified, and the user is willing to interact with the SAR.

During the whole campaign, condition (a) was not always met, as users were often outside their apartments for their regular activities. On average, condition (a) was met for the

58.4% ($\sigma = 0, 97\%$) of the days of the experimental campaign. When the users were available, the robots performed, on average, 2.85 interventions per day ($\sigma = 0.7\%$).

At the same time, when the users were at home (condition (a) fulfilled), they were often busy performing other tasks such as cleaning, cooking, and eating, thus not meeting condition (b). As our system was able to detect the user's presence and location, but not to estimate their activity (see Luperto et al. [31] for further details), the robot often tried to perform interventions that were postponed by the user answering *later*.

While the issues in fulfilling conditions (a) and (b) had an impact on all interventions planned by the system, the consequences are more evident on interventions that should be performed often as SQ or on interventions that require the user to be available for several minutes as NPT.

Interestingly, when condition (b) is met, the users generally are interested in engaging with the platform, fulfilling condition (c). This shows the benefits of a system like ours, which does not encourage users to change their habits to use the system, but is flexible and adapts its behaviour to the user activity.

The users were instructed to perform their regular activities *as if* the system is not present, and not to adapt their behaviour to the system. We chose this approach intentionally, as we believe that robotics platforms should not hinder their users in their routine and everyday lives but should be able to adapt and cohesively provide assistance.

Availability and NPT

Eight out of 10 users performed NPT at least once. Two users never performed the NPT. This event could have been caused by two main reasons. First, the long duration of NPT, at least 20 minutes, as notified by the robot to the users, made it so that the robot asked the user to interrupt their daily activities to do the tests and that the users were not available for that amount of time The second reason regards the availability of the users, who spent most of the daytime outside their apartment and, sometimes, decided to turn the system off (condition (a)). This reduced the possibility of the system to perform interventions. Moreover, the robot was interacting with the users for several other activities that were scheduled on a regular time interval. As a consequence, sporadic interventions such as NPT, which should be performed only once over a few months, are more likely to be postponed if the number of interventions to be performed by the robot is higher than the slots available. This is particularly clear from the data from `ESP` where fewer users performed NPT, but also where the robot asked fewer times to comply with the protocol.

Notice that the robot is not constantly tracking the user's behaviours and activities, in an attempt not to be intrusive. The robot's only knowledge about the user is the estimated location collected through the IoT network. This might increase the chance that the robot chooses the wrong time to ask the user to perform an intervention. An example of this can be seen for user `ITA-4`, who was often outside the apartment but returned several times a day for a few minutes: this caused the system to start interventions that were later dropped when the user left the apartment and labelled as *outside*.

A further side-effect of limited users' availability is that the system tried to perform many more interventions than envisaged, as it was particularly difficult to complete the interventions. As an example, if all NPT interventions were performed at the first attempt, only 20 interventions would have been required to cover the entire experimental campaign. However, the system needed to attempt 83 interventions, of which only 15 resulted in a NPT being performed.

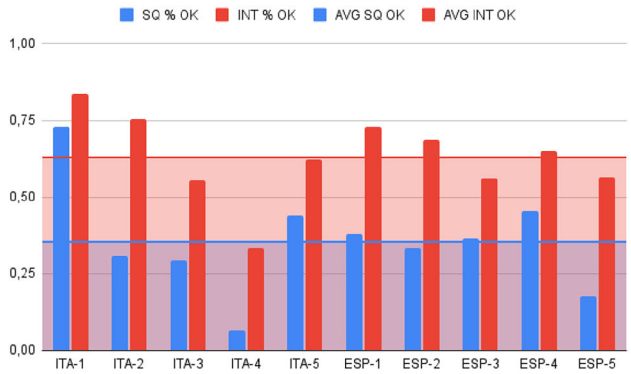
In this work, we performed all tests (TMT A+B, Bells) in a single session, with the aim of performing only a setup phase of the tests. In future works we will investigate if splitting the test into different moments, thus distributing the time needed to perform the full set of tests, can facilitate the users in executing them.

Availability and SQ

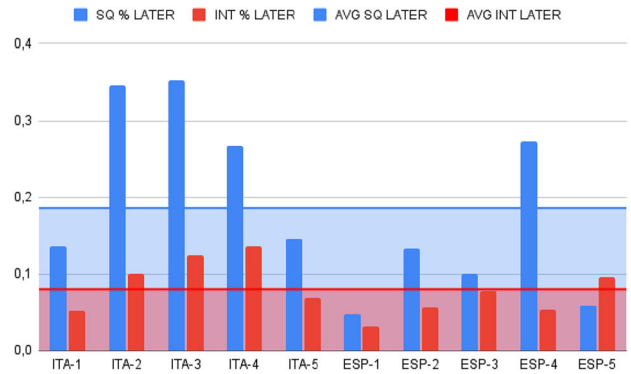
The fact that condition (a) was not met for a significant part of the experimental campaign greatly reduced the possibility of performing interventions that required a tight schedule, as SQ was. Despite we managed to perform a fair amount of SQ, a less dense schedule for longitudinal monitoring is preferred to enforce adherence to the protocol.

Another indication about issues related to user availability comes from Figure 10, which shows how the SQ (in blue) has a significantly lower percentage of answers than the entirety of the interventions developed for MoveCare and performed by the robot throughout the experimental campaign (in red). This could indicate that having to answer a question to the robot was perceived as time-consuming while the same does not apply when the robot is simply giving instructions to the users to complete a task that did not involve the robot. In the former case, the user had to perform the task (answering to the robot) at that moment; in the latter case, the user could perform the task the robot was asking whenever they wanted.

The system was designed to perform SQ in a carefully designed daily and weekly schedule: when the user was not at home the robot entered a busy waiting loop until the user came back home. However, asking SQ just after the user came back home was revealed to be a non-ideal choice as the user was often engaged in several tasks and could not answer the robot right away (condition (b)). This could be



(a) % of interventions with an answer.



(b) % of interventions when the user asked to postpone the interaction, answering `later` to the robot's request.

Fig. 10 Comparison against how many times the users answered a Spot Questions (SQ, in blue) against the replies to all robot interventions during the main experimental campaign (INT, in red). Labels 3 and 4 are the area, indicating the average value

the reason for the many times in which the user postponed the SQ intervention.

As a result, the number of SQ interventions that were postponed is high (43% of the cases, against an average of all interventions of 27%).

Stemming from this consideration, future works should investigate how to plan the daily tasks of a SAR to increase their chances of engaging the users when they are available.

7.2 Validity of the At-Home Assessment

While our work empirically shows the feasibility of at-home cognitive monitoring, several methodological issues are still open. On one side, such a modality of administration, distributed over time, and the fact that the monitoring is carried out at home, in an uncontrolled environment, makes it difficult to identify the standard controlled conditions invoked by clinicians. On the other side, conducting neuropsychological evaluations in a clinical facility may interfere with the behaviour of the patient, potentially jeopardizing the assessment [6]. Future work should compare the results of the at-home monitoring modality against those obtained by analogous evaluations carried out in the clinics and possibly pave the way to robust at-home cognitive monitoring. Another open point is how to provide the results of tests performed at home to professional caregivers. Digitalized tests can be particularly useful as they can be automatically scored and they can be evaluated by the clinician through a replay functionality [26]. Digitalized tests can also be used to compute new features that are commonly not considered in paper-and-pencil-based evaluations [11].

Another positive result of our experimental campaign emerges when considering other aspects of cognitive assessment. The administration of neuropsychological tests in the clinical setting involves a time-consuming effort on the part of the neuropsychologists. They not only have to administer the tests, but also explain the task assignments, motivate the patients to perform to the best of their ability, and calm them to alleviate their performance anxiety that clinical assessments inevitably increase. Our campaign showed positive preliminary results considering all these aspects. In particular, the robot turned out to be a viable at-home substitute for the clinician. Furthermore, the possibility of voluntarily participating in the test and being able to postpone its execution, together with the comfortable setting of one's own home, proved to be valuable elements in sustaining the participant's motivation and ease in taking the test.

In addition, the fact that users judged confabulations SQ as "strange", and identified them as a technical error of the system is intriguing and instructive as it highlights the differences to be kept in mind when implementing clinical instruments in a different context. The perception of the participants that confabulation-inducing questions are strange is exactly the kind of response that clinicians expect from a patient who does not confabulate, and who is therefore healthy. During clinical assessment, the healthy patient's reactions to confabulation-inducing questions are often of amused surprise, and it is not uncommon for some to ask the psychologist if they are making fun of them. So, in the case of the robot, the participants' embarrassed responses to confabulation-inducing questions could be clinically valuable responses that indicate preserved cognitive functioning. This fact indicates that it is possible to monitor clinical signs and symptoms not only deriving from the concrete answers given to the tests and questions asked by the robot but also the emotional or behavioural reactions exhibited by the participants while performing the tasks. This opens up new scenarios and opportunities in the neuropsychological monitoring and assessment of patients at home. However, our experience, and the consequence that we needed to remove confabulation questions from SQs, indicates how

these insights should be investigated more thoroughly in controlled studies before being deployed in a complex setting as those of our study.

It must be noted that, starting from the design stages of the experimental campaign to its full deployment, our clinical partners were involved in (1) identifying the most appropriate tests to be carried out at home; (2) ensuring that they were delivered most similarly and unobtrusively with respect to their paper-pencil counterparts. Having clinical partners in the project has proved fundamental for implementing good practices into our cognitive monitoring system. The development of our cognitive monitoring intervention has been performed in an iterative way, and our clinical partners were involved in both the design and evaluation phases. Digitalized cognitive tests were first evaluated in a clinical setting with a human caregiver, using a tablet but with no SAR's involvement (see Lunardini et al. [26]), and later tested completed with the SAR's supervision (see [27]). Finally, a preliminary campaign (see Sect. 4.3) was conducted in controlled apartments before the final deployment of the wider experimental campaign.

7.3 Environmental Complexity

One thing that emerges from Table 3 and 4 is how several interventions have failed due to `errors` that prevented the robot from completing its task. For example, the robot was unable to move, find the user, or reach a viable position for the interaction. These errors do not always depend on the robot and, in several cases, the exact cause of a failure is difficult to ascertain. This is due to the fact that older adult apartments are challenging environments for SARs. This fact is particularly important, as it can prevent SARs from successfully collecting monitoring data, while also affecting the perception of the robot from the end users' perspective.

The behaviour of SARs can vary significantly from one environment to another. As an example, the robot of the user `ITA-2` was able to find them quickly, as seen in Fig. 9, without following complex and more prone-to-navigation-failure paths. The one time when the robot of the user `ITA-2` lost its localization (in red) was when it tried to locate the user by searching throughout the whole environment. Instead, the robot of the user `ITA-1` (Fig. 8) needed to perform more complex paths across the whole environment multiple times.

Another aspect of this is that not all events of the environment are observable by the robot: while we can have, through IoT, a knowledge of the expected user location, we cannot infer where the user will go next (e.g., to perform the `setup` phase of NPT). Consequently, when a `setup without tests` event happens, we do not have any means to identify the exact cause of such an event.

7.4 Cognitive Monitoring as a Part of a Larger Framework

Integrating SQ and NPT in the MoveCare system and testing them within the main experimental campaign has proved challenging, due to the availability problems and environmental complexities described above. It could be argued that the reported results on the cognitive monitoring module are hindered by the global experience of having a complex system deployed in the houses of users. We argue that the strength of our results lies in the fact that they have been collected within a complex experimental campaign. The contribution that we report here is part of a more general experience by design, and it is not and should not be evaluated independently from it. The core idea of our project is to develop an at-home heterogeneous framework around the needs of the users, to test both the system as a whole and specific advanced functionalities outside controlled environments and into the houses of the target population. Even when integrated with the larger framework that required the users to comply with a large set of requests, NPT and SQ were not deemed tedious or invasive. On the contrary, by looking at the results for NPT, the longest intervention of the whole experimental campaign, we can see that the users were engaged and reported feeling challenged but intrigued and happy to carry it out. This is to underline that, even if the self-reported results on the single specific interventions were biased from the global experience, the objective compliance to the interventions (when available) demonstrates that the scenarios were a success within the bigger experience. At the same time, following the same reasoning, we can say that the global experimental campaign was successful as the users did not report distress or discomfort in any of the self-reported scales. As the users were accustomed to having the robot in their houses, and they knew how to interact with it, they were comfortable performing long and difficult tasks as the NPT (as shown in Table 5). As an example, during the lockdown due to the COVID-19 pandemic, a user asked to continue to have the robot after the end of the experimental campaign as they felt that the robot's presence was friendly and companionable and helped them during the difficult period. In the future, when robots are at the stage of entering their users' houses, they will not be performing one single intervention; they will have multiple functionalities and it will be the sum of the parts that will result in a successful or abandoned robotic companion.

7.5 Limitations and Future Work

The experimental campaign we describe here aimed at investigating the feasibility of deploying a SAR in the house of older adults to carry out cognitive monitoring, interleaved with other daily interventions. Due to its preliminary nature,

the campaign involved only 10 participants, 2 of whom did not contribute to the results. This is a limitation that should be acknowledged. Future efforts should be directed to organise a larger campaign, which will be most useful for gathering more quantitative data on the acceptability and usability of usability of the system, to complement the results discussed here and in [27].

Furthermore, in terms of acceptability, we plan to explore with prospective users how they perceive the impact of the proposed solution on their finances. This is a societal consideration that must be taken into account when introducing SARs as potential clinical helpers. Having this discussion with end users and clinicians will help us inform technology providers and policy makers. The proposed setup was developed as a prototype by a broad consortium of partners, involving the company that produced the robot. During the project, the robot underwent several upgrades with respect to its standard setup. It is thus not a commercial product. To estimate the costs of such a product, a business plan assessing potential use cases and their markets is required. Such a study would require efforts that are beyond the scope of this paper and the associated research project.

One of the main lessons learnt from our experimental campaign is that user availability is key. One of the main limits of our system is that to be able to gather the data it needed, it had to approach the users when they were available to comply with the requests of the robot. However, this requires the system to estimate the users' availability, which is a complex task. A possible solution to increase the chances of a successful interactions will be giving the possibility to the users themselves to reschedule the intervention by providing a precise time in which they are available to complete it.

8 Conclusion

The work presented in this paper aims at developing a non-intrusive system, to be deployed in the house of older adults, that could help the detection of early signs of MCI. It consists of a robotic platform able to guide its users through the completion of neuropsychological tests and to transparently ask a series of spot questions for longitudinal monitoring.

While our work shows how the use of SQ for collecting answers from users in long-term monitoring is feasible, several points remain open. Future work will involve assessing the critical problems regarding the validity of spot questions such as what type of questions to ask and with what frequency.

Among the future working directions for the NPT, we envisage the development of other cognitive tests, perhaps based on different tasks and interactions with the robot and a comparison of the effectiveness of their assessment with the classic paper-and-pencil deployment.

For both scenarios, it will be fundamental to keep on working with clinicians to arrive at the definition of a clinically validated at-home monitoring tool.

Funding The research leading to these results received funding from European Commission H2020 - project MoveCare - under Grant Agreement No ICT-26-2016b - GA 732158 and project Essence grant SC1-PHE-CORONAVIRUS-2020-2B - GA 101016112, and the Italian PON project SI-Robotics.

Data Availability Statement Full data acquired from the platform and that document the platform usage by users are not available due to privacy issues of the participants, as they contain details of the daily activities performed by users. Further details on the methods used for the collection of these data and their content are available upon request made to the corresponding author. All data are available to the corresponding author for further inquiries.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Standards The experimental campaign described in this paper has been approved by the ethical committees of Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy (8/04/19) and of Gerencia del Area de Salud de Badajoz, Junta de Extremadura, Badajoz, Spain (13/02/19). The protocol of the campaign was approved by the ethical committee. All participants signed an informed consent where they agreed to participate in the study. Informed consent forms were approved by the ethical committees.

References

1. Abdi J, Al-Hindawi A, Ng T et al (2018) Scoping review on the use of socially assistive robot technology in elderly care. BMJ Open 8(2):e018815
2. Amabili G, Cucchieri G, Margaritini A et al (2022) Social robotics and dementia: results from the eware project in supporting older people and their informal caregivers. Int J Environ Res Public Health 19(20):13,334
3. Belli E, Nicoletti V, Radicchi C, et al (2020) Confabulations in cases of dementia: atypical early sign of Alzheimer's disease or misleading feature in dementia diagnosis? Front Psychology 2597
4. Bellotto N, Fernandez-Carmona M, Cosar S (2017) Enrichme integration of ambient intelligence and robotics for aal. In: Proceedings of the AAAI spring symposium series
5. Budd D, Burns LC, Guo Z et al (2011) Impact of early intervention and disease modification in patients with predementia Alzheimer's

disease: a Markov model simulation. In: ClinicoEconomics and outcomes research: CEOR, vol 3, p 189
6. Chaytor N, Schmitter-Edgecombe M (2003) The ecological validity of neuropsychological tests: a review of the literature on everyday cognitive skills. Neuropsychol Rev 13(4):181–197
7. Ciuffreda I, Amabili G, Casaccia S et al (2023) Design and development of a technological platform based on a sensorized social robot for supporting older adults and caregivers: Guardian ecosystem. Int J Soc Robot. https://doi.org/10.1007/s12369-023-01038-5
8. Coradeschi S, Cesta A, Cortellessa G et al (2014) Giraffplus: A system for monitoring activities and physiological parameters and promoting social interaction for elderly. In: Human–computer systems interaction: backgrounds and applications, vol 3, pp 261–271
9. Dahmen J, Cook D, Fellows R et al (2017) An analysis of a digital variant of the trail making test using machine learning techniques. Technol Health Care 25(2):251–264
10. Davey J (2006) "Ageing in place"—the views of older homeowners about housing maintenance, renovation and adaption. Technical Report
11. Di Febbo D, Ferrante S, Baratta M et al (2023) A decision support system for Rey–Osterrieth complex figure evaluation. Expert Syst Appl 213(119):226
12. D'Onofrio G, Sancarlo D, Raciti M et al (2019) Mario project: experimentation in the hospital setting, ambient assisted living. Springer, Cham, pp 289–303
13. European Commission (2021) The 2021 ageing report. Economic & budgetary projections for the EU member states (2019–2070). European Economy Institutioanl Papers
14. Feil-Seifer D, Mataric MJ (2005) Defining socially assistive robotics. In: Proceedings of the 9th international conference on rehabilitation robotics, 2005. (ICORR 2005), pp 465–468
15. Feil-Seifer D, Matarić MJ (2005) Defining socially assistive robotics. In: Proceedings of the 9th international conference on rehabilitation robotics (ICORR), pp 465–468
16. Fellows RP, Dahmen J, Cook D et al (2017) Multicomponent analysis of a digital trail making test. Clin Neuropsychol 31(1):154–167
17. Fiorini L, Sorrentino A, Becchimanzi C et al (2022) Living with a telepresence robot: results from a field-trial. IEEE Roboti Autom Lett 7:5405–5412
18. Fischinger D, Einramhof P, Papoutsakis K, et al (2016) Hobbit, a care robot supporting independent living at home: first prototype and lessons learned. Robot Autonom Syst 75:60–78. Assistance and Service Robotics in a Human Environment
19. Folstein MF, Folstein SE, McHugh PR (1975) "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12(3):189–198
20. Garzo A, Martinez L, Isken M, et al (2012) User studies of a mobile assistance robot for supporting elderly: methodology and results. In: Proceedings of the workshop on assistance and service robotics in a human environment, international conference on intelligent robots and systems (IROS)
21. Gasteiger N, Ahn HS, Gasteiger C et al (2021) Robot-delivered cognitive stimulation games for older adults: Usability and acceptability evaluation. J Hum Robot Interact 10(4):18
22. Gauthier L, Dehaut F, Joanette Y (1989) The bells test: a quantitative and qualitative test for visual neglect. Int J Clin Neuropsychol 11:49–54
23. Giovagnoli AR, Del Pesce M, Mascheroni S et al (1996) Trail making test: normative values from 287 normal adult controls. Ital J Neurol Sci 17(4):305–309
24. Gross HM, Mueller S, Schroeter C et al (2015) Robot companion for domestic health assistance: implementation, test and case study under everyday conditions in private apartments. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS 2015), pp 5992–5999
25. Johansson MM, Marcusson J, Wressle E (2015) Cognitive impairment and its consequences in everyday life: experiences of people with mild cognitive impairment or mild dementia and their relatives. Int Psychogeriatr 27(6):949–958
26. Lunardini F, Luperto M, Romeo M et al (2020) Supervised digital neuropsychological tests for cognitive decline in older adults: usability and clinical validity study. JMIR Mhealth Uhealth 8(9):e17,963
27. Luperto M, Romeo M, Lunardini F et al (2019) Evaluating the acceptability of assistive robots for early detection of mild cognitive impairment. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 1257–1264
28. Luperto M, Basilico N et al (2021) Luperto, M. et al. (2021). A Community-Based Activity Center to Promote Social Engagement and Counteract Decline of Elders Living Independently. In: Baldoni, M., Bandini, S. (eds) AIxIA 2020 – Advances in Artificial Intelligence. AIxIA 2020. Lecture Notes in Computer Science(), vol 12414. Springer, Cham. https://doi.org/10.1007/978-3-030-77091-4_24
29. Luperto M, Monroy J, Renoux J et al (2022) Integrating social assistive robots, IoT, virtual communities and smart objects to assist at-home independently living elders: the movecare project. Int J Soc Robot 15:517–545
30. Luperto M, Romeo M, Monroy J et al (2022) User feedback and remote supervision for assisted living with mobile robots: a field study in long-term autonomy. Robot Auton Syst 155(104):170
31. Luperto M, Monroy J, Moreno FA et al (2023) Seeking at-home long-term autonomy of assistive mobile robots through the integration with an IoT-based monitoring system. Robot Auton Syst 161(104):346
32. Mann JA, MacDonald BA, Kuo IH et al (2015) People respond better to robots than computer tablets delivering healthcare instructions. Comput Hum Behav 43:112–117
33. Nasreddine ZS, Phillips NA, Bédirian V et al (2005) The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 53(4):695–699
34. Quigley M, Conley K, Gerkey B, et al (2009) Ros: an open-source robot operating system. In: ICRA workshop on open source software
35. Rabbitt SM, Kazdin AE, Scassellati B (2015) Integrating socially assistive robotics into mental healthcare interventions: applications and recommendations for expanded use. Clin Psychol Rev 35:35–46
36. Reitan RM (1958) Validity of the trail making test as an indicator of organic brain damage. Percept Mot Skills 8(3):271–276
37. Rossi S, Conti D, Garramone F et al (2020) The role of personality factors and empathy in the acceptance and performance of a social robot for psychometric evaluations. Robotics 9(2):39
38. Sorrentino A, Mancioppi G, Coviello L et al (2021) Feasibility study on the role of personality, emotion, and engagement in socially assistive robotics: a cognitive assessment scenario. Informatics 8(2):23
39. Strauss E, Sherman EM, Spreen O (2006) A compendium of neuropsychological tests: administration, norms, and commentary. American Chemical Society, New York
40. Tupikov A (1976) Confabulation in atrophic and vascular disease of old age (clinico-psychopathologic findings). Zhurnal Nevropatologii i Psikhiatrii Imeni SS Korsakova (Moscow, Russia: 1952) 76(8):1181–1186
41. Varrasi S, Di Nuovo S, Conti D et al (2019) Social robots as psychometric tools for cognitive assessment: a pilot test. In: Human friendly robotics. Springer, pp 99–112
42. Votruba KL, Persad C, Giordani B (2016) Cognitive deficits in healthy elderly population with "normal" scores on the mini-mental state examination. J Geriatr Psychiatry Neurol 29(3):126–132

43. Wada K, Shibata T, Saito T et al (2003) Effects of robot assisted activity to elderly people who stay at a health service facility for the aged. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS 2003), vol 3, pp 2847–2852