Into the Wild: Pushing a Telepresence Robot Outside the Lab*

Amedeo Cesta[1], Gabriella Cortellessa [1], Andrea Orlandini [1] and Lorenza Tiberio[1]

***Abstract*— Most robotic systems are usually used and evaluated in laboratory setting for a limited period of time. The limitation of lab evaluation is that it does not take into account the different challenges imposed by the fielding of robotic solutions into real contexts. Our current work evaluates a robotic telepresence platform to be used with elderly people. This paper describes our progressive effort toward a *comprehensive, ecological* and *longitudinal* evaluation of such robots outside the lab. It first discusses some results from a twofold short term evaluation performed in Italy. Specifically we report results from both a *usability* assessment in laboratory and a subsequent study obtained by interviewing 44 healthcare workers as possible secondary users (people connecting to the robot) and 10 older adults as possible primary users (people receiving visits through the robot). It then describes a complete evaluation plan designed for a long term assessment to be applied "outside the lab" dwelling on the initial application of such methodology to test sites in Italy.**

I. Introduction

The area of social robotics is receiving increasing attention and the task of "robot as companion" has received attention at research level [1]. Several projects have also proposed different types of solutions with robots that both interact with humans and are connected to heterogeneous technology to build innovative living environments (e.g., [2], [3], [4]). This paper aims at underscoring one aspect connected to such a line of innovation that deserves special attention: the study of *attitude and perceptions of people who share the environments in which the robot operates* over *long periods of time*.

It is also worth noting how in robotics there is a deep-rooted tradition in developing technology usually shown in sporadic events and for short periods, i.e., for demos or live show cases. These demonstrations usually aims to present the "enhanced" characteristics of a prototype, making them attractive while "hiding" or at least "containing" the technical problems connected with any long term use within a comprehensive application. Indeed, a key requirement for social companions (e.g., robots assisting older adults at home) is their continuous operation, their robustness and the continuous interaction with humans over time. Such continuity of use has significant implications on the technology development but it also highlights the need to design a methodology for assessing human reactions with respect to prolonged use of the proposed solutions. The challenges for the Intelligent Technology and the Human Robot Interaction researchers are numerous and mainly related to two aspects: (a) in terms of *users perspective*, robots must adhere to user requirements and be acceptable in the long term, (b) in terms of *technology*, the need exists to create usable, robust, efficient and secure solutions. More specifically, the transition from a use in the laboratory to an actual deployment into real contexts, highlights the need for a shift from short term to long term experiments. In particular we underscore how *long-term use* and *evaluation* are key points to be addressed to ensure that robotic technology can make a leap forward and be used in real environments.

In the framework of the EU Ambient Assisted Living (AAL) Joint Program[1] we are part of a project called EXCITE[2], which is performing a wide program of evaluation in the field of an industrial mobile telepresence platform called GIRAFF produced by GIRAFF Technologies AB[3], Sweden. More specifically, we take part in an evaluation spanning three different EU countries – Italy, Spain and Sweden. The evaluation takes social and psychological factors into account to study users attitude and reaction, but also analyzes the emergence of "undesired behaviors" like technological weaknesses in continuous operation, possibly leading to human rejection. In this work, we present the results gathered in Italy after the short term evaluation phase and, then, we present and discuss the general long term evaluation methodology, showing its current application to real test sites. The paper[4] introduces the context of work (Section II), then analyzes and reasons about the work both to realize short term experiments with real users and to develop a methodology for addressing long term evaluation (Section III, Section IV, Section V); finally it describes the status of the first test sites in Italy where the long-term evaluation methodology is being applied (Section VI).

II. Context of work

Telepresence robots have been increasingly proposed to be used in workplace and Mobile Remote Presence (MRP) systems have been studied as a means to enable remote collaboration among co-workers [5], [6]. Furthermore, MRPs are also being used to provide support to elderly people. In this respect, some research exists which aims to understand the acceptance of older adults, their concerns and attitude toward the adoption of MRPs [7], [8], [9]. Our work is motivated by the participation to the EXCITE project, aiming at promoting the use of MRPs to foster interaction and social

*Authors are partially supported by the EU under the Ambient Assisted Living Joint Program – EXCITE Project (AAL-2009-2-125) – and under the ICT Call 7 – *GiraffPlus* Project (GA 288173).

[1]Authors are with CNR, Italian National Research Council, ISTC, Rome, Italy. http://www.istc.cnr.it/

[1]http://www.aal-europe.eu/

[2]http://www.excite-project.eu/

[3]http://www.giraff.org

[4]This is a revised version of this work: A. Cesta, G. Cortellessa, A. Orlandini, L. Tiberio. Evaluating Telepresence Robots in the Field. To appear in a special issue of Springer's LNCS TCCI Journal.

participation of older adults as well as to provide an easy means to possible caregivers to visit and interact with their assisted persons in their living environment. GIRAFF is a remotely controlled mobile, human-height physical avatar integrated with a videoconferencing system (including a camera, display, speaker and microphone). It is powered by motors that can propel and turn the device in any direction. An LCD panel is incorporated into the head unit. The robotic platform is accessed and controlled via a standard computer/laptop using a software application. From a remote location the *client*, or *secondary user* (member of family or healthcare professionals) with limited prior computer training teleoperates the robotic platform while older adults (*end users, or primary user*) living in their own home (where the robot is placed) can receive their visit through the MRP. The remote user can charge the robot batteries by driving it onto a docking station.

Key pursued ideas

The EXCITE project aims at assessing the validity of an MRP in the field of elderly support in different European countries. The project fundamental concepts are the following:

- *User centered product refinement*. This approach is based on the idea of obtaining users feedback during the time they use the robot and cyclically refine the prototype in order to address specific needs;
- *User tests outside labs*, rather than testing the system in laboratory setting, the MRP is placed in a real context of use. This approach is in line with several research that highlights how systems that work well in the lab are often less successful in real world settings [10]. The evaluation of robots made in a laboratory environment does not favor the emergence of robotic aid suitability to support elderly who are able to stay in their own homes. For this reason an essential step is to assess the technology in the specific contexts in which the technology is supposed to be used [11];
- *Use on a time period long enough*, to allow habituation and possible rejection to appear. Indeed, interviews and survey conducted after a short period of time, though useful and valuable can not be the only way to assess techonology since they can be limited and can prevent other effects to emerge. On the contrary, a key aspect of relationship is that it is a persistent construct, spanning multiple interactions [12]. In this light, in order to assess the human-robot interaction it is important to investigate how people interact with robots over long periods of time.
- *Analysis of cultural and societal differences*, an interesting part of our project stems from the idea of comparing the long term deployment of the telepresence platform in different countries so as to allow an analysis of cultural and societal differences over European countries.

Different GIRAFF prototypes are being deployed for long periods of time (at least three months, and possibly 1 year) in real context of use. Feedback obtained from the users (both primary and secondary) is used to improve the robot. In what follows, we describe our progressive work toward a long-term human-robot interaction assessment showing how we combing short term evaluation sessions with long term efforts.

III. THE EVALUATION APPROACH

We have conceived a twofold path for evaluating the human-robot interaction gathering both feedback from short interactions between potential users and the GIRAFF robot and also focusing on a long term assessment plan. More specifically we identified two tracks for our effort:

- *Short Term Evaluation*, which consists of a collection of immediate users feedback (i.e., after a short interaction with the robot) on the telepresence robot, connected to different aspects of the interaction mainly related to the usability, willingness to adopt it, possible domains of applications, advantages and disadvantages.
- *Long Term Evaluation*, which relates to the study of the long-term impact of GIRAFF's social and physical presence on elderly users using the system both to communicate with their relatives and friends and to receive visits from healthcare providers and in general caregivers.

The short term evaluation effort, though not sufficient alone, still provides immediate feedback that can be used to quickly improve the technological development, to possibly add functionalities to the system or to simply confirm the validity of some technological choices. In addition it can give valuable guidance to the long-term assessment. For this reason we adopted a combined approach involving participants representative of both types of users: the *secondary* and *primary* users.

Following this schema, we first present results for the short term evaluation performed in Italy, then we introduce our complete design for a methodology to assess the long-term impact of the GIRAFF in EXCITE also reporting the status of the Italian long-term test sites that are currently running according to this methodology.

IV. SHORT TERM EVALUATION

For the short term evaluation effort we first realized some usability experiments in laboratory, so as to identify possible problems in the user interaction with the system. Subsequently we organized user evaluation sessions with real potential users of the system to investigate other complementary aspects.

A. Usability evaluation

The usability assessment has been made by using both an observational technique and a usability questionnaire. Specifically, we relied on the *Thinking Aloud* evaluation technique [13], which consists of asking the users to verbalize their thoughts while performing certain tasks and interacting with the system. The experimenter observes silently the interaction session, and records user's actions and thoughts, focusing on the difficulties and problems encountered. In

addition, the System Usability Scale (SUS) [14] was administered as an additional measure[5].

1) Participants and procedure: five participants took part in our usability experiment (see Figure 1). Four of them were male students (with a mean age of 18,4) and one was their teacher (male, age 54)[6]. All the participants had experience in software and computer and received training prior to the test consisting of a tutorial presentation of 20 minutes and a practical session. After the tutorial each participant received written instructions on specific tasks and how to carry them out. Four main tasks have been considered that can be grouped as the following: (a) *make a video call*; (b) *navigate in the environment*; (c) *read a text through the robot*; (d) *perform the docking*.

During the sessions participants were encouraged to "think aloud" to verbalize their opinions while completing the assigned tasks. The sessions were recorded and the experimenter took notes during the session.

At the end of the test, the SUS questionnaire was administered and a final interview was conducted to understand opinions with respect to the telepresence system experience and to discover further problems and take note of additional advices. Also this interview was recorded. The recordings have been analyzed and experiment results have been written in the form of Usability Aspect Reports (UARs)[7].

2) Results: Overall the interface was judged usable, even though some specific problems still emerged. The detailed UARs have been examined and have been organized according to four main categories:

a) Video and audio: the control and audio quality were judged overall very good. The video instead has been considered not completely satisfactory. The quality seems, in fact, sufficient to allow for general navigation in the environment but not entirely satisfactory in case you need to perform specific visual inspections such as reading a text or recognize the state of some specific objects within the environment. One solution would be to improve the quality of the camera and also to provide it with a zoom feature.

b) Navigation: the navigation in the environment was generally satisfactory. Some difficulties were encountered when the robot had to move in extremely narrow spaces or with obstacles. A suggestion from participants regards the possibility to insert a map and a position indicator of the robot within the environment. This feature could possibly be superfluous in case the secondary user is a son or a person who knows the environment in which the elderly live. On the contrary, it would be particularly useful if the secondary users is a person less familiar with the explored environment (e.g. a formal caregiver or a health professional). In addition some autonomy for helping the remote operator of the robot, when the driving is more critical could ease the navigation.

(a)

(b)

Fig. 1. Pictures from the "Thinking Aloud" evaluation session: (a) Reading task; (b) Driving task

c) Client Interface: the client interface was satisfactory. The commands for the control of the robot have been judged as clear and easily identifiable. A possible improvement concerns the indicator of the level of charge that could be implemented with a more visible color or through a flashing signal that would attract the attention when the battery is reaching a critical level.

d) Docking: this was the most critical functionality from the point of view of usability. At least half of the participants had difficulties in the docking. This is both because of poor video quality, and the manual docking conducted without visual aids. Possible solutions to this problem are: implementation of an automatic docking functionality or alternatively, providing the base with more visible indicators (e.g. colored) and simultaneously put directional indicators in the interface which can "guide" during manual docking.

As for the SUS usability questionnaire, results show that the GIRAFF application scored 77 of 100 points. Our result

[5]The SUS instrument is a reliable tool for measuring the usability of a wide variety of products and services. It is composed of 10 statements that are scored on a 5-point scale of strength of agreement. Final scores for the SUS can range from 0 to 100 where scores above 70 indicate products which are at least passable. Scores in the high 70s to upper 80s guarantee products with a good acceptability. Greatly superior products score better than 90.

[6]The specific choice of this sample was motivated by the fact that the participants were somehow representative of the secondary users we had contacted for the long term test sites. Specifically, the main secondary users were: a man with experience in using PC and technology in general and young boys with skill in both computer usage and video games. Our plan is however to enlarge the sample size also considering other age brackets.

[7]The detailed UARs are not reported for the sake of space. They have been analyzed and grouped into four main categories.

can be interpreted as an index of a good acceptability and ease of use. Therefore, the general usability assessment was quite good, though some aspects could still be improved.

Some common aspects emerged also from the analysis of the content of semi-structured interview. Specifically, referring to the experience of use participants were asked to judge the interaction through the robot relying on a semantic differential with six adjective pairs on 6 point scale. The participants agreed in judging the telepresence experience as *active*, *participatory* and *exciting*. The GIRAFF's height was judged adequate but its base was considered cumbersome.

B. Assessment of primary and secondary users attitude

After the usability assessment results, we started involving possible users of the telepresence system in order to study their opinions on the use of telepresence systems. As stated in [7], before intelligent technologies would be accepted, it is important to understand their perception of the benefits, concern and adoption criteria. In our study, we aim at reproducing as much as possible an "ecological" setting for the experiment. To this purpose we distinguished the role of the users and recruited different participants according to their expected role. Specifically for the secondary users group we recruited users representative of the potential visitors of the elderly users among caregivers, nurses, health workers, etc. For the end user side we interviewed older adults living alone, or possibly receiving some kind of health care assistance. This evaluation was aimed at assessing users reaction toward the possible adoption of the GIRAFF system as a means to visit or provide some kind of service to the elderly users. Aspects investigated were *willingness to adopt the robotic solution*, *possible domains of application*, *advantages* and *disadvantages* and *suggestions for improvements*.

Health workers as secondary users

1) Participants and Procedure: forty-four health workers from different specialist areas were recruited for this study. The sample interviewed so far is composed by 26 women and 18 men with a mean age of 42 years, $SD = 12.2$.

The meeting entailed a tutorial presentation of 20 minutes to describe features and functionalities of the telepresence robot. After this tutorial, a practical session allowed the health workers to operate the system and experience the different functionalities. Following the tutorial a focus group was conducted and a final questionnaire was administrated to assess possible applications of the telepresence robot, the perceived advantages and disadvantages of the system, the patient profile best suited to benefit from the use of telepresence.

2) Results: the results have been grouped according to the following categories:

a) General assessment: a first analysis of the results showed a positive reaction of the participants to the system. In particular 66% of participants would be willing to use GIRAFF as an aid support in his/her profession and no one opposes to the use of robots (see Figure 2a). In addition most of them judge the telepresence robot as a better tool

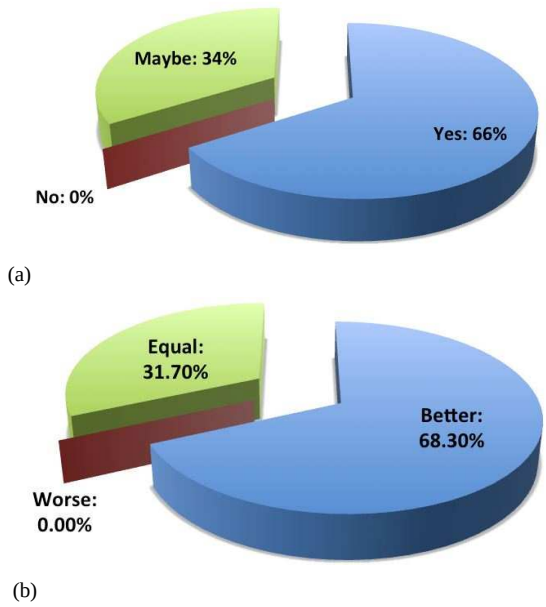

Fig. 2. General assessment of the GIRAFF system: (a) willingness to adopt it; (b) qualitative comparison with traditional teleconference systems like skype

with respect to traditional teleconference system like Skype (see Figure 2b).

b) Profile of potential users: results also identify the categories of people who could benefit from the use of telepresence robots: specifically, the category "self-sufficient or semi-autonomous elderly living alone" has been mentioned by 35% of respondents; 25% of the subjects also indicates "adults and elderly patients in home care and with special needs", such as patients in isolation for infection, dialysis patients or with chronic diseases such as Chronic Obstructive Pulmonary Disease (COPD) or diabetes. A 20% of the responses were grouped into the category "older adults with early or mild dementia". Two other categories were "adults or older adults with physical disabilities" (17%) and "young people and adults with intellectual disabilities" (7%).

c) Application domains: the participants are in favor of the use of robots to train the family caregiver to small nursing tasks and to maintain constant contact with assisted older adult. The possibility of *continuous monitoring* (see Figure 3) of the patient at home is considered the most useful application (59% of participants were in favor of this kind of application). The *support* application follows at 23%, while the *companionship* and *communication* applicative domains seem less suitable. More specifically, 45.5% of the health workers advocate the use of the robot to train a family caregiver to perform small nursing tasks (e.g., treat a bedsore, administer an enema, measuring of vital signs) and to maintain a constant contact with the patient and his family (75% of participants). Finally 60% of participants also says that the robot could alleviate the workload of the family caregiver, but not that of the health workers themselves (50% of people admit to be uncertain about the real possibility of the robot to diminish their daily workload).

d) Advantages and Disadvantages: among the advantages in using the robot, participants listed the following: a) ability to monitor remotely *via* visual communication the physical state of health; b) possibility to follow the management of medication and certain health practices (e.g., control of vital parameters such as level of blood glucose for diabetic patients, supervision of practices related to their care and medication like deep breathing exercises for patients with COPD); c) the possibility for the operator to improve his/her night surveillance activity in hospital and home care cases. Among the disadvantages they reported the poor quality of the video, the bulky size of the base unit, the fact that the robot might not be suitable for all patients, issues related to cost and privacy.

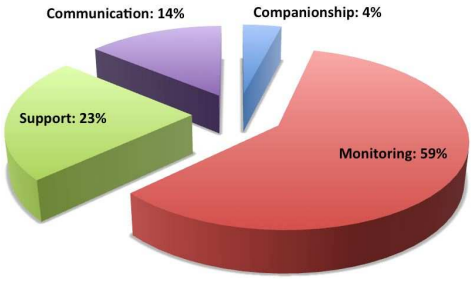

Fig. 3. Favorite GIRAFF's domains of application

e) Suggested improvements: The focus group conducted at the end of this analysis, highlighted some aspects considered as particularly relevant for using the platform in the healthcare domain for long-term period. These aspects specifically refer to improvements and integration of additional functionalities. Specifically according to participants, the need exists to improve the video quality, especially in relation to night vision; it would be useful to add the zoom functionality to the webcam; the battery duration and recharging modality should be improved (e.g., it would be better if the robot could reach autonomously the docking station); the safe navigation of the robot should be guaranteed. In addition it would be beneficial to enable the call transfer if the client is not connected to the robot via the PC. Finally the transmission of vital parameters to the doctor should be supported. All these suggestions for technical improvements are currently inspiring the future modifications of the GIRAFF system in line with the user centered approach pursued in the EXCITE project.

Older adults as primary users

1) Participants and Procedure: To investigate aspects connected to the end-user interaction with the telepresence system we contacted 10 older adults. Four of them were potential end users who have been asked to participate in the long-term evaluation described in Section V. The remaining participants are involved in a parallel study, also connected to the project that aims to validate the GIRAFF system as a tool for providing remote rehabilitation [9].

The procedure followed in this qualitative research entailed an explanation of the main idea underlying the telepresence system, showing some descriptive materials, a video of the system and, where possible, a practical demonstration of the system itself. The selection of the material and the modality to present the system were decided according to the time availability, and the specific situation presented in each evaluation session. We recorded the interview and we then opted for a qualitative analysis, summarizing the main recurrent cited positive and negative aspects, given the relatively small number of the sample. A more structure study is in our future research plan.

2) Results: A qualitative analysis of the interview have been conducted and the most relevant feedbacks are here reported in terms of positive and negative aspects of the MRP.

a) Positive Aspects: Among the positive aspects most of the subjects reported the following: participants judged the visit through GIRAFF as engaging and "real"; the robot was pleasant to see; the ability of the robot to move in the environment was positively assessed; users felt physically involved during the interaction; participants think that the robot would help someone living alone at home to feel safer; participants judged positively both the audio and the video functionalities; participants think that interaction through the robot was spontaneous.

b) Negative Aspects: Among the most negative aspects we mention: the GIRAFF system is too big and consequently may not be well integrated in a domestic environment due to its size; the battery power may be too short; there may be some problems due to the privacy issue; there were some concerns related to the safe movement of the robot and to its ability of obstacle avoidance; some "intelligent features", like the autonomous recharging of the battery, are missing; the connection to the docking station is "not very intuitive".

Also this effort showed an overall positive reaction to the system, even though some improvements are desired in view of a real usage of the system. It is worth underscoring how the key point here is the fact that qualitative data has been gathered by interviewing "real potential users" like for example a group of caregivers and older adults who can receive visits through the robot.

V. LONG TERM EVALUATION

One of the original features of the EXCITE project consists of realizing long-term experiments involving older adults hosting the robot in their living environment both to communicate with others and to receive assistance services.

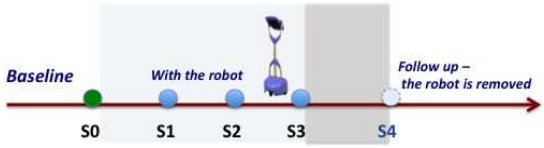

Fig. 4. The Long Term Evaluation timeline

A. Method

Figure 4 gives a general idea of the designed method to evaluate features over time. The evaluation entails a period of N months (with $3 \leq N \leq 12$) during which the end user will have the robot at home and the clients can visit him/her through it. Assessment happens at milestones S_i. Specifically, after an initial assessment (S0 in figure) at the

beginning of the experimentation (*baseline*), the variables of interest are measured at regular intervals (S1-3) to observe changes over time. At the last month the GIRAFF will be removed from the end user apartment and the same variables will be assessed again after 2 months from this removal (S4). The general idea is to use a repeated measures method to see changes over time during the long term usage of the robot.

1) Participants and Procedure: Three different cases have been identified to cover different situations in which the robot can be deployed. Specifically, for the secondary user typology we considered (a) a *formal caregiver* belonging to an Health care organization; (b) a *family member (informal caregiver)*; (c) *other relatives or friends* who may visit the elderly person through the robot. The type of material used in the long term evaluation for both the client and the end user depends upon the type of interaction for which the telepresence is used. For this reason, for each of the three mentioned situations we had developed (or selected) a set of questionnaires (almost all validated in the three languages of the involved countries) aimed at monitoring specific variables and to be administrated at specific time both to end users and to clients.

2) Material: For each of the described case we prepared the material to assess the variables under study at the specified intervals. Table I lists in detail the different variables and the related instruments to be used to measure the variables over time.

a) Client side: Specifically on the client side, during the initial step (S0), we use: (a) an informed *consent form* describing the aim and procedure of the study; (b) the *socio demographic data* form to gather some relevant information on the user; (c) we developed on purpose a questionnaire aimed at assessing the client expectation on the GIRAFF's ability to ease the support (*Support Expectation*). It is worth highlighting that we developed two slightly different types of questionnaires for the *formal* and *informal* caregivers, while for the *other relatives and friends* category we designed a questionnaire (*Influence on Relationship Expectation*) on the expectation on GIRAFF as a means to ease and support the remote communication and consequently the social relationship.

During the following step (S1), for all three types of secondary users introduced above we will use: (a) questionnaires based on the SUS inventory [14] to assess the *usability* of the client software; (b) we will ask participant to keep a *diary* to register the "salient" events of the visit through telepresence in terms of encountered problems, good features and so on.

During the subsequent step (S2), in addition to the diary that clients have to keep along the whole experience with the robot, we make a first assessment of ability of GIRAFF to ease the support (or the communication) between the client and the end user through the *Support Assessment* and *Impact on Relationship Assessment* questionnaires. In addition, during this phase we will also use the Temple Presence Inventory [15] that is a tool to measure dimensions of (tele)presence and the Networked Minds Social Presence Inventory ([16]).

At step S3 we use the Positive Affect Negative Affect Scale, PANAS, [17], the Psychosocial Impact of Assistive Devices Scale, PIADS, [18] and a final structured interview to assess the overall experience in terms of the most relevant variables considered in the study.

After two months from the robot removal, S4 will allow assessing the impact of its absence through the *Support Assessment* questionnaire.

b) End user side: For the end user receiving the robot we followed a similar approach, but we focused on some additional variables that is worth dwelling on (see next table). Specifically, we measure: (a) the *perceived loneliness* through the UCLA Loneliness Scale [19], which was developed to assess subjective feelings of loneliness or social isolation; (b) the perceived health status through the Short Form Health Survey (SF12) [20]; (c) the Multidimensional Scale of Perceived Social Support [21]; (d) Geriatric Depression Scale [22]: a modified version of the Health Service Satisfaction Inventory. Finally the Almere [23] model will allow assessing dimensions of technology acceptance.

In the table I, measures highlighted in bold will ensure the repeated measures thus allowing to observe the GIRAFF's influence by changes in response over time. It is worth underscoring how this evaluation plan will allow monitoring the human-robot interaction over time, thus contributing to understand the long term impact of a fully deployed robotic solution.

The actual implementation of this plan in three different European countries will also support a cross-cultural analysis, continuing some work started on this specific topics [24]. The following section briefly reports on the current status of the Italian test sites.

VI. FIRST TEST SITES RUNNING

Two test sites have started in Italy that are representative of the *family-member-elderly* user category.

A. Test site 1

A couple of older adults living in the countryside near Rome are the end users of this test site (see Figure 5). The man has reduced mobility, while the woman has problems with her sight. They are quite independent although their health condition is slowly deteriorating. The secondary users are: their son living in Rome and their grandchild.

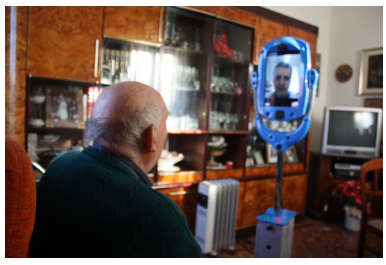

Fig. 5. A picture from the first Italian test site

We initially experienced some problems with the technical set-up of this test site. Specifically, the typical layout of the Italian houses has created some problems due to reduced space (particular difficulty emerged in going through doors and due to some narrow passage in the house) to the connection to recharging station and to smoothly move in the house. This highlighted the need to improve

TABLE I

LONG TERM EVALUATION: VARIABLES MEASURED ALONG THE PHASES (S0–S4) AND RELATED MATERIAL

PHASES	**S0**	**S1**	**S2**	**S3**	**S4**
CLIENT					
Health Professional	Consent Form, Socio-Demographics Data Form, **Support Expectation**, **Diary**	Usability, **Diary**	**Support assessment**, Temple Presence Inventory, Networked Minds Social Presence Inventory, **Diary**	PANAS, PIADS, Final Interview, **Diary**	**Support Assessment**
Family member	Consent Form, Socio-Demographics Data Form, **Support Expectation** (informal carer), **Diary**	Usability, **Diary**	**Support assessment** (informal carer), Temple Presence Inventory, Networked Minds Social Presence Inventory, **Diary**	PANAS, PIADS, Final Interview, **Diary**,	**Support Assessment** (informal carer)
Relatives friends	Consent Form, Socio-Demographics Data Form, **Influence on Relationship Expectation**, Diary	Usability, **Diary**	**Influence on Relationship assessment** (informal carer), Temple Presence Inventory, Networked Minds Social Presence Inventory, **Diary**	PANAS, PIADS, Final Interview, **Diary**	**Influence on Relationship Assessment**
END USER					
Elderly	Consent Form, Socio-Demographics Data Form, **Loneliness (UCLA)**, **Short Form Health Survey (SF12)**, **Multidimensional Scale of Perceived Social Support**, **Geriatric Depression Scale**, **Almere model**, **Health Service Satisfaction Inventory** (if applies)	**Loneliness (UCLA)**, **Multidimensional Scale of Perceived Social Support**, **Geriatric Depression Scale**, Attitude_Acceptance, **Health Service Satisfaction Inventory** (if applies)	Temple Presence Inventory, **Almere model**	**Loneliness (UCLA)**, **Short Form Health Survey (SF12)**, **Multidimensional Scale of Perceived Social Support**, **Geriatric Depression Scale**, **Almere model**, PANAS, PIADS, Final Interview	**Loneliness (UCLA)**, **Short Form Health Survey (SF12)**, **Multidimensional Scale of Perceived Social Support**, **Geriatric Depression Scale**, **Health Service Satisfaction Inventory** (if applies)

the robot's mobility and to provide an automated recharging functionality. Currently the test site is at step S0 of the evaluation plan. Some robot usability problems are emerging due to the particular fragility of the two older adults who participate in the study. The couple is very interested to the GIRAFF robot, even though its use is currently still limited. Our goal is also to monitor the robot's usage over time to assess the effect of familiarity or habituation.

B. Test site 2

A very active woman living alone in Rome is the end user of our second Italian test sites. Her grandchild and daughter are the main current secondary users. Additionally we are also planning to involve a day care center that will connect to the woman. Also this test site is currently at step S0 of the evaluation plan. However, some preliminary comments can be reported. Both the lady and her grandchild are enthusiastic

of the robot. They would also like that the robot do additional things. The lady, as most of the elderly people interviewed, is concerned about possible costs associated to the robots (e.g., the electricity consumption). Overall she really appreciates the possibility to stay in contact with her relatives, also relying on the video capability of the robot. She would also appreciate a sort of service provided by the day care center that would allow her to have a more frequent contact with a doctor or a specialist.

VII. Conclusions

This paper describes the ongoing work that is trying to assess an MRP within the elderly domain. Two important aspects are presented that can be considered as mandatory steps for both a general roadmap in robotics and our specific work.

As a first contribution, we have highlighted the importance of performing *ecological experiments*, i.e., which reproduce as much as possible the actual conditions of use of robotic technology, in terms for instance of real people who use it and real context of use. Although still simple in the results, analysis of the short-term evaluation provides a number of indications "from the field" that are representative of the actual users' expectations, both in relation to the human-robot interaction and to the most urgent technological improvements essential for an effective deployment. In addition to specific suggestions for improving the usability of the systems, we obtained other valuable recommendations that could be used for fielding the system into real world. For example, health workers expressed a number of requests that would be important to fruitfully use the GIRAFF system as a means to support their work. At the same time, the longitudinal tests done in real homes, are highlighting technological barriers that must be necessarily overcome.

The article's second contribution concerns our effort toward a long-term assessment. Other works in the area have highlighted this need but in this article we have proposed a rather elaborated and detailed methodology for the long-term evaluation that is currently being applied to real test sites of elderly people for long periods of time.

In the future we would like to enlarge the sample used in the short term evaluation possibly studying the differences among different groups of people. In addition we hope to continue gathering continuous data from the long term evaluation of the running test sites.

ACKNOWLEDGMENT

Authors are indebted to project partners for the stimulating work environment. Interactions with the colleagues from Örebro University have been fruitful to synthesize the evaluation plan. Authors would like to thank Vittoria Giuliani for comments to preliminary versions of this work.

References

[1] A. Tapus, M. J. Matarić, and B. B. Scassellati, "The Grand Challenges in Socially Assistive Robotics." *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, pp. 35 – 42, 2007.

[2] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards Robotic Assistants in Nursing Homes: Challenges and Results," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 271–281, 2003.

[3] A. Cesta, G. Cortellessa, R. Rasconi, F. Pecora, M. Scopelliti, and L. Tiberio, "Monitoring elderly people with the ROBOCARE Domestic Environment: Interaction synthesis and user evaluation," *Computational Intelligence*, vol. 27, no. 1, pp. 60–82, 2011.

[4] A. Saffiotti, "The Concept of Peis-Ecology: Integrating Robots in Smart Environments," *Acta Futura*, vol. 3, pp. 35–42, 2009.

[5] M. K. Lee and L. Takayama, "Now, I Have a Body: Uses and Social Norms for Mobile Remote Presence in the Workplace," in *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, ser. CHI'11. New York, NY, USA: ACM, 2011, pp. 33–42.

[6] K. M. Tsui, M. Desai, H. A. Yanco, and C. Uhlik, "Exploring Use Cases for Telepresence Robots," in *Proceedings of the 6^{th} Int. Conf. on Human-Robot Interaction*, ser. HRI '11. New York, NY, USA: ACM, 2011, pp. 11–18. [Online]. Available: http://doi.acm.org/10.1145/1957656.1957664

[7] J. B. Beer and L. Takayama, "Mobile Remote Presence for Older Adults: Acceptance, Benefits, and Concerns," in *Proceedings of Human Robot Interaction: HRI 2011*, Lausanne, CH, 2011, pp. 19–26.

[8] A. Kristoffersson, S. Coradeschi, A. Loutfi, and K. Severinson Eklundh, "Towards Evaluation of Social Robotic Telepresence based on Measures of Social and Spatial Presence," in *Proceedings on HRI 2011 Workshop on Social Robotic Telepresence, Lausanne, March*, 2011, pp. 43–49.

[9] L. Tiberio, L. Padua, A. Pellegrino, I. Aprile, G. Cortellessa, and A. Cesta, "Assessing the Tolerance of a Telepresence Robot in Users with Mild Cognitive Impairment – A Protocol for Studying Users' Physiological Response," in *Proceedings on HRI 2011 Workshop on Social Robotic Telepresence, Lausanne, March*, 2011, pp. 23–28.

[10] S. Sabanovic, M. Michalowski, and R. Simmons, "Robots in the Wild: Observing Human-Robot Social Interaction Outside the Lab," in *Proceedings of the International Workshop on Advanced Motion Control*. Istanbul, Turkey: ACM, March 2006.

[11] E. Hutchins, *Cognition in the Wild*. MIT Press, 1995.

[12] T. W. Bickmore and R. W. Picard, "Establishing and Maintaining Long-Term Human-Computer Relationships," *ACM Transactions on Computer Human Interaction*, vol. 12, pp. 293–327, 2005.

[13] J. Nielsen, *Usability engineering*. San Diego, CA: Academic Press, 1993.

[14] J. Sauro and J. Lewis, *Quantifying the User Experience: Practical Statistics for User Research. Software Usability Scale (SUS)*. Chap. 8, pages 198-208. Morgan Kaufmann, 2012.

[15] M. Lombard, T. Ditton, and L. Weinstein, "Measuring Telepresence: The Temple Presence Inventory," in *Proceedings of the Twelfth International Workshop on Presence, Los Angeles, California (USA).*, San Francisco, 2009.

[16] Networked minds social presence inventory (scales only version 1.2). [Online]. Available: http://cogprints.org/6742/

[17] A. Terracciano, R. R. McCrae, and P. T. Costa, "Factorial and Construct Validity of the Italian Positive and Negative Affect Schedule (PANAS)," *European journal of psychological assessment official organ of the European Association of Psychological Assessment*, vol. 19, no. 2, pp. 131–141, 2003.

[18] H. D. J. Jutai, "Psychosocial impact of assistive devices scale (piads)," *Technology and Disability*, vol. 14, no. 3, pp. 107 – 111, 2002.

[19] D. Russell, L. A. Peplau, and C. E. Cutrona, "The Revised UCLA Loneliness Scale: Concurrent and Discriminant Validity Evidence," *Journal of Personality and Social Psychology*, vol. 39, pp. 472–480, 1980.

[20] J. E. J. Ware, M. Kosinski, and S. D. Keller, "A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity," *Medical Care*, vol. 34, no. 3, 1996.

[21] G. D. Zimet, N. W. Dahlem, S. G. Zimet, and G. K. Farley, "The Multidimensional Scale of Perceived Social Support," *Journal of Personality Assessment*, vol. 52, no. 1, pp. 30–41, 1988.

[22] J. A. Yesavage, T. L. Brink, T. L. Rose, O. Lum, V. Huang, M. Adey, and V. O. Leirer, "Development and Validation of a Geriatric Depression Screening Scale: a Preliminary Report," *Journal of Psychiatric Research*, vol. 17, no. 1, pp. 37–49, 1983.

[23] M. Heerink, B. J. A. Kröse, V. Evers, and B. J. Wielinga, "Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model," *I. J. Social Robotics*, vol. 2, no. 4, pp. 361–375, 2010.

[24] G. Cortellessa, M. Scopelliti, L. Tiberio, G. Koch Svedberg, A. Loutfi, and F. Pecora, "A Cross-Cultural Evaluation of Domestic Assistive Robots," in *Proceedings of AAAI Fall Symposium on AI in Eldercare: New Solutions to Old Problems*, 2008.