A Virtual Laboratory for Studying Long-term Relationships between Humans and Virtual Agents

Timothy Bickmore, Daniel Schulman
College of Computer & Information Science, Northeastern University
360 Huntington Ave, WVH202
Boston, MA, USA
+1 (617) 373-5477
{bickmore,schulman}@ccs.neu.edu

ABSTRACT

Longitudinal studies of human-virtual agent interaction are expensive and time consuming to conduct. We present a new concept and tool for conducting such studies—the virtual laboratory—in which a standing group of study participants interacts periodically with a computer agent that can be remotely manipulated to effect different study conditions, with outcome measures also collected remotely. This architecture allows new experiments to be dynamically defined and immediately implemented in the continuously-running system without delays due to recruitment and system reconfiguration. The use of this tool in the study of a virtual agent that plays the role of an exercise counselor for older adults is described, along with the results of an initial experiment into the effects of conversational variability on user engagement and exercise behavior.

1. INTRODUCTION

As computers interact with us in increasingly complex and human ways through robots, wearable devices, PDAs, and various other ubiquitous interfaces, the psychological aspects of our relationships with them take on an increasingly important role [3]. It is important to not only understand the nature of this phenomenon and its effects in work and leisure contexts, but also to develop strategies for constructing and managing these relationships, which directly impact productivity, enjoyment, engagement and other important outcomes of human-computer interaction.

Virtual agents are ideal platforms for exploring human-computer relationships, since their anthropomorphic appearance automatically cues social responses in users, and their nonverbal behavior can be used to communicate and assess the relational aspects of their user interactions. Examples of relational behavior that can be used by virtual agents include empathy, immediacy, and social chat. For many applications, such as in counseling, healthcare, education, and sales, relationships have been shown to lead to not only increased user satisfaction, but significant improvements in task outcomes as well [5].

Inherent in the notion of relationship is that it is a persistent construct; incrementally built and maintained over a series of interactions that can potentially span a lifetime. The problems that arise in maintaining user engagement, enjoyment, trust—and productivity (in work contexts)—over a long period of time are important and open issues in HCI and virtual agent research.

Unfortunately, conducting longitudinal evaluations of virtual agents is very difficult, tedious, costly and, of course, time consuming. The virtual agent software and infrastructure must be very robust to support many, many interactions without failing, and anomalous behavior (which can have negative long-term impacts on trust) must be avoided if at all possible. Study participant recruitment, retention, and compensation must be addressed in ways that are quite different from single session laboratory studies.

To address these needs, we have developed a "virtual laboratory" to support multiple, possibly concurrent, longitudinal studies of user interactions with a virtual agent. The laboratory is comprised of a virtual agent that is run on study participants' home computers as a network client, and a multi-user server consisting of a multi-threaded dialogue engine, relational database, and study administration interfaces (Figure 1).

In the rest of this paper we discuss related work, and briefly describe the virtual laboratory system for conducting longitudinal studies of virtual agents, in order to ground the subsequent discussion. We then describe a series of important issues that must be addressed when building systems for long-term use and longitudinal evaluation, using the virtual laboratory as an example. We then provide a brief primer on the statistical analysis of data from longitudinal studies, and present preliminary results from the first study we have conducted using the virtual laboratory before concluding.

2. RELATED WORK

Few longitudinal studies of virtual agents have been conducted to date. Bickmore developed a series of "relational agents" for health education and health behavior change interventions. The FitTrack system featured a virtual agent deployed on networked home computers that promoted walking among sedentary adults, and used a range of nonverbal (e.g., facial displays of empathy, proxemic cues, hand gestures) and relational (e.g., social dialogue, empathic exchanges, humor) behavior in its daily conversations with patients [3]. A one-month randomized pilot study, in which participants were asked to talk to the agent daily, was conducted to evaluate the effectiveness of this agent in increasing moderate-intensity physical activity levels among 101 sedentary adults [5]. Adults randomized to the agent program showed significant increases in number of days/week that they engaged in at least 30 minutes of moderate-intensity or more vigorous physical activity, relative to adults randomized to a usual

Cite as: A Virtual Laboratory for Studying Long-term Relationships between Humans and Virtual Agents, Timothy Bickmore, Daniel Schulman, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009),* Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 297–304

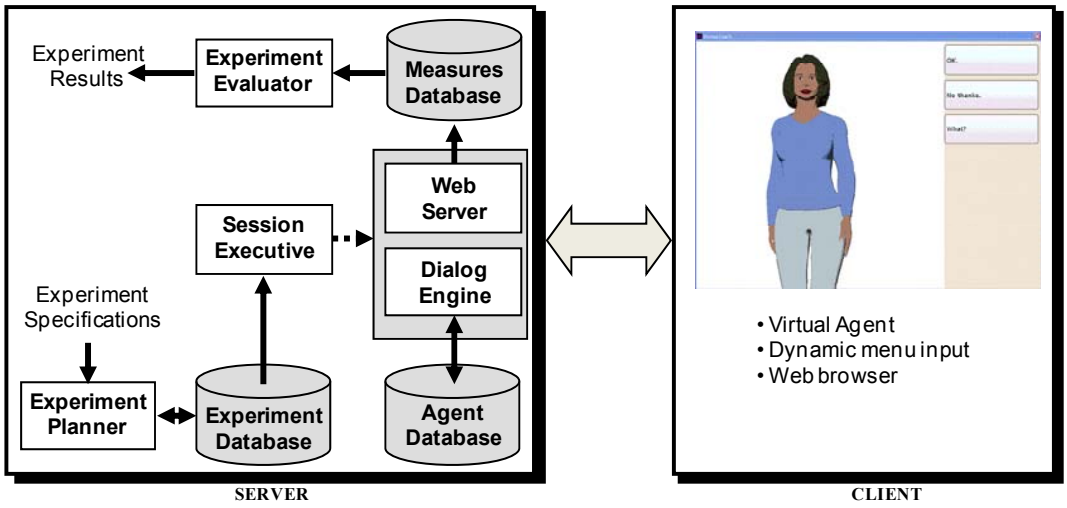

Figure 1. Virtual Laboratory Architecture

care program (standard print-based materials). In addition, those interacting with the fully relational agent showed significantly higher "therapeutic alliance" scores (measuring user trust in the agent) compared with users in a sub-group who interacted with a similar agent that did not use relational behavior. Users interacting with the relational agent also reported a significantly stronger desire to continue the intervention at the end of the month, compared to those interacting with the non-relational agent. However, relational behavior was not shown to mediate increases in physical activity.

A subsequent version of the FitTrack system was tailored for use with older low-income populations and evaluated in a two-month, daily contact, randomized pilot study involving 22 geriatrics patients (age range=63-85 years), 67% of whom had little or no previous computer experience. The virtual exercise counselor agent ran on stand-alone touch-screen PCs provided to users during the study. Following the two-month intervention, older adults randomized to the agent program arm showed significant improvements in physical activity levels relative to those randomized to a control arm (i.e., pedometers and print-based materials) [2].

Bickmore also reports a pilot study evaluation of a virtual agent that promoted antipsychotic medication adherence among a population of adults with schizophrenia [4]. This virtual agent ran on laptop computers provided to study participants for the 30-day, daily contact intervention. Preliminary results indicate that study participants talked to the agent on 65.8% of the available days. Self reported medication adherence (gathered through dialogue with the agent) was 97%, and desire to continue the intervention at the end of the month was rated at 4.0 on a scale of 1 (not at all) to 5 (very much).

There are also few examples of longitudinal evaluations in the Human-Computer and Human-Robot Interaction literature. Kidd developed a relational robotic agent to promote weight loss among obese patients. In a six-week trial involving 45 subjects (age 50.1+/10.6, BMI 25 to 42), participants randomized to use the robotic interface used their system to log diet information on 50.6 days, while those randomized to conduct the same interaction with a text-based computer interface used their system 36.2 days, and those using a paper diet log only recorded 26.7 days of data, $F(2; 30) = 11:51$; $p < 0:001$. Participants also rated the robotic relational agent higher on the working alliance inventory compared to the text-based computer, $t(17)=5.1$, $p<0.001$ [8]. There are also an increasing number of examples of longitudinal evaluation studies in HCI, mostly in the health and wellness domains (e.g., [6]).

3. THE VIRTUAL LABORATORY SYSTEM

The virtual laboratory incorporates a standing group of study participants who interact with a virtual agent on a regular (e.g., daily) basis for an indefinite period of time. The agent runs on their desktop computers and features a lightweight virtual agent and integrated web browser that obtains all daily content (web pages and dialogue) from a central server connected to the client over the Internet. A typical interaction consists of a brief 5-15 minute conversation with the agent, possibly including the display of web content, followed by a series of web form-based questionnaires to perform whatever self-report assessments are required for the current experiment being performed. The system was designed so that there is always default dialog content that will be used by the agent in the absence of any newly specified content.

The concept of the virtual laboratory is that it provides a persistent, on-going stream of user-agent interactions that can be perturbed—e.g., by changing dialogue or web content or agent behavior—and frequently assessed through self-report, questionnaire-based measures. This, in turn, allows new experiments of arbitrary duration to be dynamically specified and executed.

Figure 1 shows the virtual laboratory architecture. The client part of this architecture features a virtual agent, web browser, and user input windows (Figure 2). The server features the following components: an agent database for storing all user data and information about previous user-agent interactions; a measures database for storing all experimental results (e.g., from questionnaires remotely administered to users); an experiment database that contains specifications for all experiments to be run; a dialogue engine that manages conversational interaction between the agent and a user; a web server that provides users

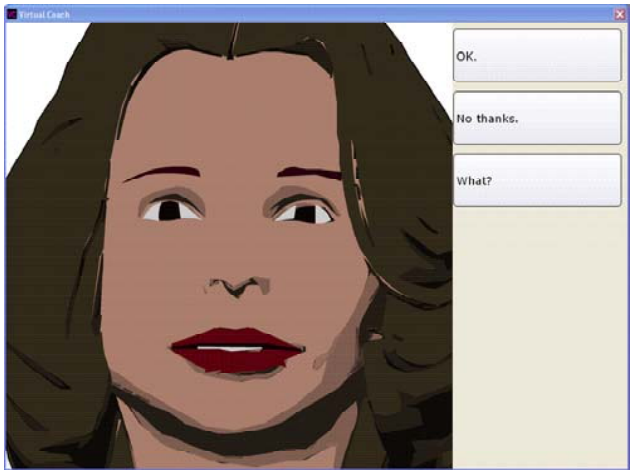

Figure 2. Virtual Agent Interface

with web content (e.g., multimedia educational material and study questionnaire forms); the dialogue engine parameters to instantiate for a particular user on a particular day; an experiment planner that schedules requested experiments; and an experiment evaluator that produces data files and web-based summaries of experimental results.

3.1 RADIUS Dialogue Engine

The virtual laboratory system can be used with a variety of dialogue engines which manage the interaction between a user and the virtual agent. In previous systems, we have used dialogue engines based on augmented transition networks, in which a fragment of dialogue is specified as a finite state machine, generally with agent utterances as states, and possible user responses as state transitions [5]. The hierarchical structure of many dialogues [7] is modeled by using hierarchies of state machines, in which a state transition may cause another state machine to be executed as a sub-dialogue, before continuing on to the following state. This model is conceptually simple, and has been shown to be easy to understand for domain experts without extensive knowledge of programming or computational linguistics. In previous projects, domain experts have been able to author dialogue in this formalism with relatively little training or assistance. However, while this approach has worked well for the creation and use of content, we have had difficulty with the *reuse* of content across projects. For example, within the domain of health behavior change, we have identified several dialogue fragments that are generally useful, but often require small changes for a particular behavior change intervention. In practice, this leads to the cutting and pasting of dialogue fragments, with the resulting well-known associated software engineering problems of maintaining multiple copies of the same code.

Other researchers have used task-decomposition-based planners to model collaborative dialogue, such as in COLLAGEN [12]. In these models, a dialogue is a task to be performed jointly between the agent and the user. Tasks may be either atomic (e.g., a turn of dialogue), or they may be performed by decomposition into a sequence of simpler tasks. A "recipe" specifies a sequence of tasks that achieves a particular complex task (goal) [10]. There may be multiple recipes that achieve the same goal, possibly with preconditions that restrict when each recipe may be used.

For the virtual laboratory, we have developed a new dialogue engine—RADIUS (relational agent dialogue system)—which subsumes both augmented transition network-based and task-decomposition-based models of dialogue. In contrast to more complex systems, such as COLLAGEN, RADIUS models a recipe as a state machine, in which agent utterances are states, and user utterances are state transitions. A state transition may invoke a sub-task by specifying a goal, which will cause the dialogue engine to find an appropriate recipe and execute it, before continuing to the next state. In practice, this provides increased modularity and reuse with only a small increase in complexity for authors. Dialogue may still be written as state machines. However, when modifications are required in order to reuse a dialogue fragment, this may be implemented by providing additional recipes for those portions of dialogue.

3.2 Test Domain: Physical Activity Promotion for Older Adults

Although the virtual laboratory can be used in any virtual agent application domain, we have been using physical activity promotion for older adults as our initial domain. This domain has enormous potential: it is intrinsically valuable to the study participants, it allows us to combine task talk with social conversation in our experiments, and the use of pedometers provides us with an objective measure of task success. Participation in moderate amounts of physical activity has important health benefits to everyone, including beneficial effects on risk factors for disease, disability, and mortality, yet, a substantial proportion of the U.S. adult population remain underactive or sedentary [9]. Older adults are in particular need of physical activity interventions: only 12% of adults over 75 get the minimum level of physical activity currently recommended by the Centers for Disease Control and Prevention, and 65% report no leisure time physical activity [9]. In addition, the older adult population comprises a very diverse group—in age and gender, physical and mental ability, racial and ethnic background, and computer experience. Our current protocol uses Omron HJ-720ITC pedometers that can store and automatically upload their data to users' home computers and from there to the virtual laboratory server, providing behavioral data every day a participant is in the study.

A typical daily interaction with the exercise counselor will last 10 minutes and consist of the agent walking on the screen, greeting the user, engaging in small talk, acquiring the user's pedometer readings, providing feedback (including plots of steps and goals over time), setting goals for the next day, and a farewell exchange, after which a series of self-report questionnaires are displayed for the user to complete.

4. DESIGNING VIRTUAL AGENTS FOR LONG-TERM USE

In this section we outline several important considerations in running longitudinal studies, and methodologies and architecture features for addressing them.

4.1 Persistence

In order to conduct more than one "intelligent" interaction with users, virtual agents must remember something about their past encounters with them. At a minimum, the fact that the agent has interacted with a given user before, and perhaps the number

and/or duration of such interactions must be remembered between sessions. Persistent memory should ultimately be represented as an episodic store recording details of all past interactions with users. A useful middle ground is to record specific facts that can be referenced in future conversations. Examples in the physical activity coaching domain include remembering the name of a user's walking buddy or favorite walking location, as well as purely social (off-task) facts, such as the user's favorite television program and whether they had any big plans for the upcoming weekend. In addition, if there is any possibility that more than one user can interact with a virtual agent, some form of user identification must be used so that the correct persistent memory is retrieved and used [13].

In the virtual laboratory system, persistent memory is kept in two forms. First, the date each participant was enrolled in the laboratory is recorded, in addition to the date they were enrolled in any particular study treatment condition. Second, all dialogue engines currently share a single data structure representing longitudinal information about each user in the form of updateable attribute-value pairs that are retrieved from the database at the start of each conversation and saved back to the database at the conversation's completion. Use of this latter form of persistent storage is explicitly scripted in dialogue recipes.

4.2 Reliability

Reliability is another essential feature for a virtual agent that will support long-term interactions with users. Not only are many people going to be interacting with the agent, but they will be doing so on a regular basis for a long period of time, so the software used must be as robust as possible to support the hundreds or thousands of expected interactions without continual support from research staff. Even so, research staff or system administrators must be available to resolve problems as quickly as possible when they arise, so as to not jeopardize the validity of an on-going study.

The virtual laboratory was developed using many components that had been successfully used on prior studies, and all software is thoroughly tested by our development team before it is fielded. We also provide participants with phone and email support should they encounter any problems. In addition, one of the tenets of reliable software is that there should be multiple recovery procedures in place to provide seamless or "graceful" degradation when things fail. When a software exception occurs, the agent tells the user "Sorry, I have to run now." walks off the screen ending the interaction, and an alert is logged in the user database for a member of research staff to investigate. We also ensure that default dialogue content is provided in as many situations as possible in case authors forget to cover all situations.

4.3 Client-Server Architecture

The use of home-based or mobile systems are crucial for high retention rates in longitudinal studies, since requiring participants to return to a lab even a few times can significantly increase drop-out rates. However, the use of such distributed systems can require extensive effort to update should bugs or other required changes be identified once a system is deployed, and the likelihood of such issues arising increases with the length of the study. In addition, these systems must be complete before they are sent out into the field, even though some parts of them may not be used for weeks or months after the start of the study. All of these issues can be addressed through the use of a central server that provides the content and logic for a virtual agent, with a client, for example running on participants' home computers, simply providing the interface (web-based agents are an example). This enables bug fixes and updates to be immediately propagated to all participants, and allows a study to begin even before all of the content has been developed.

The virtual laboratory uses a thin client agent interface that is run on study participants' home computers as a PC application (Figure 2). Upon startup, it connects to the server, administers a login sequence, then begins the interaction with the user. Communication between the virtual agent client and the server is performed via XML messages that specify interface actions (agent, browser, and menu displays) and user responses (menu selections, browser actions, pedometer data).

4.4 Dynamically Updatable Software

Ideally, the server should be configured so that it can be updated without having to continually stop and restart it, preventing participants from talking to the virtual agent while updates are being made.

The virtual laboratory server is modular and component-based, so that content updates (including dialogue content, surveys, web pages, and even the dialogue engine itself) can be updated while the server is running. To this end, the server is based on the OSGi framework, which is a dynamic component model for Java [1]. OSGi manages software in bundles (archive files) and tracks and resolves dependencies among bundles that are dynamically loaded and unloaded. OSGi also maintains a registry of services provided by bundles, and this mechanism is used to look up the dialogue engine and content to use for a given user for a given conversation, typically indexed by the number of days a user has been interacting with the system. Although a variety of dialogue engines can be used, all engines currently share a single data structure representing longitudinal information about each user: the set of attribute-value pairs described above. Study conditions are effected by setting attribute values in this common data structure which either cause the dialogue indexing service to return different dialogue engines and/or scripts, or cause the appropriate dialogue engine to function differently, depending on how the attributes are set.

In addition, the RADIUS dialogue engine described above uses a hierarchical model of dialogue, with dialogue recipes specified by the conversational goal they satisfy and preconditions. For example, the top-level "HaveAConversation" recipe may specify an initial sub-goal labeled "DoGreetingExchange" that is satisfied by two sub-dialogues, "InitialGreetingExchange" and "SubsequentGreetingExchange", with the selection between them governed by applicability conditions (tests on the user model, in this case the number of conversations a user has had with the agent). This effectively provides run-time linking of dialogue segments, which further enables agent content to be updated while the server is running. For example, a new greeting recipe "FinalGreetingExchange" may be developed once a study is underway, intended to be used on the final day a participant talks to the agent.

4.5 Interaction Content (Recipe) Re-Use

Virtual agents that hold many conversations with users typically require a large amount of dialogue content, even if any given

conversation is very short in duration. Although text generation techniques hold promise for ultimately providing unbounded, procedurally-generated dialogue, the state of the art for most virtual agents is hand-scripted dialogue, albeit organized into various formalisms. In order to support the authoring of this volume of content, dialogue recipes should be designed for re-use across conversations. To this end, they should provide variability in the surface forms that utterances take, and provide context-dependent behavior. Returning to the greeting dialogue discussed above, "SubsequentGreetingExchange" could produce the agent utterances "Good morning, Bob." or "Good afternoon, Sally." based on time of day and the current user's given name.

The virtual laboratory supports dialogue recipe re-use through the dynamic sub-goaling mechanism in RADIUS and use of the attribute-value pairs stored in the persistent user model. We have also developed several design patterns for re-entrant dialogue scripts. One pattern involves asking the user about something, remembering the value, then asking the user on a subsequent conversation if the value has changed. For example, the "DoWeatherChat" recipe initially asks the user about the current weather conditions, then on subsequent conversations asks them if it is "still nice out" or "still cloudy", etc. Qualitative feedback from past study participants has indicated that users like these relatively trivial references to past conversations, because it gives them a sense of continuity with the agent.

5. CONDUCTING LONG-TERM STUDIES WITH VIRTUAL AGENTS

5.1 Participant Recruitment, Retention, and Compensation

The recruitment, retention and compensation of participants in long-term studies present significant challenges. Planning of fixed duration studies involves estimation of the participant attrition rate so that enough participants are recruited in order to satisfy the power analysis requirements at study completion. For example, if—based on prior experience—you assume that 20% of your participants will withdraw during your study ("lost to follow up"), and your statistical power analysis indicates that you need a total of 50 subjects to guarantee your desired power, then you need to recruit at least 63 participants ((1-0.2)xNumberRecruited=50). Compensation is usually pro-rated based on number of study tasks completed in order to motivate participants to adhere to the study protocol. Explicit rules for when and how study participants will be contacted during the study (e.g., if they fail to check in within a specified time interval) must be specified in order to minimize demand effects and other confounds, as are rules for when participants will be dropped from a study for non-adherence.

In the virtual laboratory system, the definition of experiments, as well as the assignment of study participants to experimental conditions (both within- and between-subjects) is handled through a web-based administrative interface. This same interface is used to enroll participants into and withdraw participants from the system. In addition to maintaining adherence to any one study, we are also concerned with retention of participants in the virtual laboratory framework. At the time they first start, participants are told they can stay in the virtual laboratory system up to four years or until they withdraw or miss 14 consecutive daily interactions, at which time they are dropped. Participants are contacted after missing 5 days and again after 10 and 12 days in an attempt to keep them in the laboratory. In addition, participants are paid monthly, with compensation based on the number of complete interactions they have conducted with the virtual agent. Overall compliance with this protocol (conducting daily interactions) has been 79% after 16 weeks and over 1,500 interactions.

5.2 Automated Experiment Administration and Data Collection

Administration of and data collection from longitudinal study participants can be greatly simplified by limiting the number of contacts you have with them during the study. For example, a typical smoking cessation intervention may only collect data from study participants at baseline, and at six, twelve and eighteen months into the study. Even this number of contacts, however, can be burdensome with more than a handful of participants. In addition, more frequent collection of data provides the opportunity to use more powerful statistical analysis techniques, as well as to obtain a richer variety of data than would otherwise be possible. These issues can be addressed by automating all aspects of experiment administration and data collection.

The virtual laboratory provides a web-based interface to administrative functions, including enrollment of new participants, assignment of participants to study conditions, and report generation, including payment schedules and a "slackers report" that indicates which participants have not talked to the virtual agent for 5 or more days and should be contacted by a member of the research staff. In addition, all study data, including pedometer steps and self-report questionnaire responses, are automatically collected during each virtual agent interaction and stored in a study measures database.

6. DATA ANALYSIS FOR LONGITUDINAL STUDIES

Longitudinal studies can provide far richer information about change over time than non-longitudinal studies. However, study design and data analysis become correspondingly more complex.

6.1 Stopping Conditions

A key design choice of any longitudinal study is the stopping conditions which determine when data collection from a participant is complete. The standard approach in longitudinal studies is to specify a fixed time duration between measurements, and a fixed number of measurements before a participant is said to complete the study. Alternative approaches are to withdraw a participant from a study once their outcome measurements appear to have been following a stable trajectory for some period of time, or when some other stopping condition is satisfied, such as those offered by sequential analysis techniques [17]. In contrast to well-known statistics for analyzing a one-time test of hypotheses following all data collection, sequential analysis is a family of methods designed for repeated or continuous analysis. These methods produce stopping conditions which determine when sufficient data has been collected to either accept or reject the hypotheses, and correct for the fact that multiple hypothesis tests are being performed. Sequential analysis is most commonly applied for clinical trials where ethical considerations dictate that a study should be halted as soon as possible so that all participants can receive the most effective treatment. These methods are appealing for this type of longitudinal study, but there are caveats: First, the application of sequential analysis to longitudinal data

adds substantial complexity [14]. Second, these methods are primarily designed to produce stopping conditions for a study as a whole, not for individual participants.

The virtual laboratory system is designed so that participants take part in a series of experiments sequentially, until they choose to leave the system or become inactive. The stopping conditions therefore determine when a participant should be switched into a different experiment, and the key consideration is making the most efficient use of the participant pool. If the duration of a study is too short, it will fail to collect sufficient data to produce significant results, while longer study durations will decrease the number of experiments a participant can take part in, and increase the likelihood that they will drop out of the system.

6.2 Outcome Evaluation

A number of statistical frameworks have been developed in recent decades which offer powerful tools for analyzing longitudinal data. Two that are commonly used are generalized estimating equations [18] and linear mixed models (also referred to as hierarchical linear models [16]). Linear mixed models focus on modeling individual change over time, and can estimate how much of the variability in observed outcomes is due to differences between subjects. Generalized estimating equations are most commonly used to estimate population-average effects, but require fewer distributional assumptions. Compared to simpler and more familiar approaches, such as a repeated-measure ANOVA, linear mixed models have several advantages. They are more efficient when analyzing unbalanced data, in which varying numbers of measurements are available for different participants, and in which the times of measurements may vary.

7. INITIAL STUDY: DOES DIALOGUE VARIABILITY MATTER?

One surprising finding in the longitudinal studies of the FitTrack system was that, even though dialogue scripts had been authored to provide significant variability in each days' interaction, most participants found the conversations repetitive at some point during the month, and because of this many lost motivation to follow the agent's advice [2; 5]. As one participant in the second study put it, "It would be great if Laura could just change her clothes sometimes." This repetitiveness was more than an annoyance; some subjects indicated that it negatively impacted their motivation to exercise (e.g., "In the beginning I was extremely motivated to do whatever Laura asked of me, because I thought that every response was a new response.").

Our first longitudinal study using the virtual laboratory is thus to evaluate the impact of perceived agent repetitiveness on retention and adherence to a health behavior change intervention. The study had a between-subjects with two treatments: VARIABLE and NONVARIABLE. We designed two parallel sets of dialogue scripts to promote walking as a form of exercise (following Bickmore [3]). The scripts were functionally identical, except that in the NONVARIABLE condition, the agent used the exact same dialogue structure and language in every situation (e.g., contingent positive reinforcement was always given as "Congratulations. Looks like mission accomplished on the exercise.") and the agent's appearance and setting are never changed. In contrast, in the VARIABLE condition, one of five different dialogue structures are randomly selected each conversation to guide the overall flow of the interaction, and every agent utterance has multiple surface forms, of which one is selected randomly during each conversation (e.g., "Looks like you met your exercise goal of 5,000 steps. Great job!", "Looks like you got your walking in and met your goal of 5,000 steps!", etc.). In addition, one of five different background scene images was randomly selected and displayed behind the agent at the start of each conversation.

7.1 Participants

Twenty-four participants (17 female, 7 male, aged 55 to 75) enrolled in the virtual laboratory system and took part in the initial study. All participants were required to be 55 or older, and to have access to an internet-connected personal computer. Participants were screened at recruitment for eligibility. Participants were required to be able to start a physical activity program, assessed using the PAR-Q questionnaire [15], and participants who were already regularly engaging in regular moderate exercise (30 minutes or more, 5 days a week) were excluded.

7.2 Measures

Steps walked per day were measured with Omron HJ-720ITC pedometers. Participants were prompted once per day to connect their pedometer to the computer so that the step count could be automatically downloaded. The pedometers store up to 6 weeks of step counts, so that information was not lost if a participant did not interact with the system on a particular day.

At the end of each daily interaction, participants completed two single-item questionnaires, which measured their desire to continue using the system ("How much would you like to continue working with Karen?"), and the perceived repetitiveness of the interactions (manipulation check; "How repetitive are your conversations with Karen?"). Both used a 5-point Likert scale, ranging from "not at all" to "very much".

7.3 Procedure

Participants underwent a short intake procedure, which took place in our laboratory, at which time they were randomly assigned to one of the two study conditions. Participants received brief instruction in the use of the pedometer and in interaction with the virtual agent, and participated in a sample interaction. Following intake, the participant retention and compensation procedures described above were used. Participants had up-to-daily interactions with the virtual agent. The researchers did not contact participants except to follow the retention procedures if a participant did not interact with the agent for several days.

7.4 Results

Of the 24 participants, 10 were randomized to the VARIABLE condition, and 14 to NONVARIABLE. To date, participants have been interacting with the system between 40 and 120 days (mean 82.25), and 3 subjects from each group (6 in total) dropped out before the time of this analysis.

Figure 3 shows the primary outcomes from the study.

In order to examine the trends in participant behavior over time, we analyzed the data using linear mixed modeling. All analysis was performed using R 2.7.2 [11] with the "nlme" package.

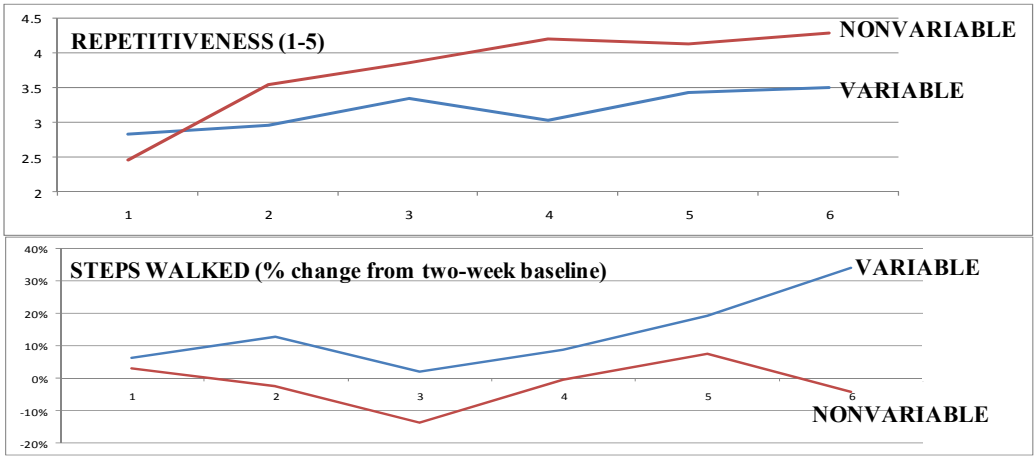

Figure 3. Results from Variability Study (daily data averaged by week)

7.4.1 Perceived Repetitiveness

Table 1 shows the result of fitting a linear mixed model, with perceived repetitiveness as the outcome variable, and including fixed effects of study day, study condition, and their interaction, and random effects on intercept and slope. Parameters were estimated using restricted maximum likelihood.

Inspection of the data showed that most participants tended to give the same answer for several days consecutively. Therefore, we modeled the within-subject residuals as a first-order autoregressive process, in order to account for autocorrelation. A likelihood ratio test ($\chi^2(1)=218.65$, $p<.0001$) showed that this produces a significantly better fit to the data.

There was a near-significant interaction between study day and condition ($p=0.051$); The average participant in the NONVARIABLE condition reported an increase in perceived repetitiveness of approximately 0.018 per day (on a 5-point Likert scale). There was a large amount of between-subject variability in the intercept, which corresponds to perceived repetitiveness on day 0 (SD=1.301), and a smaller amount of between-subject variability in the slope (SD=0.020).

Table 1. LMM estimates of effects of study day and condition on perceived repetitiveness. (Condition 0=NONVARIABLE, 1=VARIABLE)

Parameter	Value	Std. Error	p-value
Intercept	3.100	0.421	0.000
Day	0.010	0.007	0.142
Condition	-0.015	0.554	0.979
Day*Condition	0.018	0.009	0.051

7.4.2 Desire to Continue

We observed a ceiling effect on this outcome measure; the mean response was 4.67 (on a 5-point Likert scale). Due largely to these issues, little useful information, and no significant results, were observed using linear mixed modeling.

7.4.3 Performance Relative to Goals

Every time a participant talks to the agent, they are asked to negotiate a goal for the number of steps they will walk each day until their next conversation. As a measure of performance, we analyzed the difference between steps walked and the goal, for each day on which a participant had an interaction with the agent and negotiated a goal. Table 2 shows the result of fitting a linear mixed model, including fixed effects of study day, study condition, and their interaction, and random effects on intercept and slope. Parameters were estimated using restricted maximum likelihood.

There was a significant interaction between study day and condition ($p<0.01$); The average participant in the NONVARIABLE condition reported a decrease in performance of approximately 45 steps per day. There was a large amount of between-subject variability in performance on day 0 (SD=1296.77), a smaller amount of variability in the slope (SD=27.97), and a very large amount of within-subject variability in performance on any arbitrary day (SD=2653.1).

Table 2. LMM estimates of effects of study day and condition on performance relative to goals. (Condition 0=NONVARIABLE, 1=VARIABLE).

Parameter	Value	Std. Error	p-value
Intercept	541.42	483.71	0.263
Day	-5.12	11.93	0.668
Condition	1105.32	646.14	0.102
Day*Cond'n	-45.77	17.12	0.008

7.5 Discussion

While these results are preliminary, we can see that there is indeed a negative effect of removing dialogue variability from this intervention: Participants reported significantly worse performance relative to their daily walking goals over time, and also reported a trend towards greater perceived repetitiveness over time.

We observed some methodological issues with the daily subjective assessments we used in this study (desire to continue, and perceived repetitiveness). There was a substantial ceiling effect on desire to continue, and on both measures, participants had a strong tendency to give the same response for several consecutive days. We conjecture that these issues may be largely the result of "question fatigue" in participants, who are asked to answer an identical question every day, and quickly adopt a

strategy of either selecting the same answer every day, or selecting random answers. An alternative approach would be to give a longer, multi-itemf questionnaire at greater intervals. Researchers planning future longitudinal studies should carefully consider the appropriateness of daily assessments.

8. GENERAL DISCUSSION

The virtual laboratory represents a new concept, tool, and system for conducting longitudinal studies of interaction between humans and virtual agents. The virtual laboratory concept separates the methodological issues of participant recruitment, retention, and compensation from the essential research questions and design of a study. The virtual laboratory tools provide reusable software and content to ease the implementation of longitudinal studies. And finally, the virtual laboratory system provides a stable and persistent pool of participants, to reduce the up-front effort required to recruit participants for these studies.

8.1 Limitations and Open Issues

Carryover effects represent one of the most significant methodological concerns for the virtual laboratory concept. Participants' experience in previous studies may influence their behavior in future studies. If not handled correctly, participant history may become a confounding variable in future studies, leading to invalid results. Fortunately, the system also provides a complete history of all studies each participant has experienced. We plan to make use of this information to mitigate the effects of participant history. First, the system will allow study designers to exclude participants who have experienced previous studies that are expected to cause strong effects. Second, participant history can be explicitly included (as a covariate) in statistical analysis.

A related issue involves the use of multiple concurrent studies: the virtual laboratory could allow a participant to be enrolled in multiple studies simultaneously, to make maximum use of the available participants. However, the other studies a participant is experiencing may become a potential confounding variable. As above, a study designer can decide when two studies conflict, and either exclude participants, or else explicitly consider the possible effects of multiple concurrent studies.

8.2 Future Work

Currently, the virtual laboratory relies on a thin client application that must be installed on participant's home computers. A purely web-based client is under development, which not only obviates the need for installation, but will allow new animations and even new characters to be dynamically provided to participants.

To address the repetitiveness problem, techniques for maintaining user engagement through agent variability will be developed, including: methods for dynamically integrating new authored dialogue fragments into a larger, running, dialog system; methods for subtly varying agent behavior, for example, in response to a randomly-generated "mood"; and the dynamic extraction of information from the Internet (e.g., weather, sports scores) and incorporation into social dialog generated by the agent.

9. ACKNOWLEDGMENTS

Thanks to Thomas Brown, Jenna Zaffini and the rest of the Relational Agents Group at Northeastern for their help. This work was supported by NSF CAREER IIS-0545932.

10. REFERENCES

[1] OSGi - The Dynamic Module System for Java www.osgi.org.

[2] Bickmore, T., Caruso, L., Clough-Gorr, K. and Heeren, T. 2005. "It's just like you talk to a friend" - Relational Agents for Older Adults. Interacting with Computers 17, 711-735.

[3] Bickmore, T., Gruber, A. and Picard, R. 2005. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Education and Counseling *59*, 21-30.

[4] Bickmore, T. and Pfeifer, L. 2008. Relational Agents for Antipsychotic Medication Adherence In Proceedings of CHI'08 Workshop on Technology in Mental Health, Florence, Italy.

[5] Bickmore, T. and Picard, R. 2005. Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer Human Interaction 12, 293-327.

[6] Consolvo, S., Mcdonald, D., Toscos, T., Chen, M.Y., Froehlich, J., Harrison, B., Klasnja, P., Lamarca, A., Legrand, L., Libby, R., Smith, I. and Landay, J.A. 2008. Activity Sensing in the Wild: A Field Trial of UbiFit Garden. In Proceedings of CHI '08.

[7] Grosz, B. and Sidner, C. 1986. Attention, Intentions, and the Structure of Discourse. Computational Linguistics 12, 175-204.

[8] Kidd, C.D. 2007. Engagement in Long-Term Human-Robot Interaction. PhD Thesis in Media Arts & Sciences, MIT, Cambridge, MA.

[9] Laporte, R., Adams, L., Savage, D., Brenes, G., Dearwater, S. and Cook, T. 1984. The spectrum of physical activity, cardiovascular disease and health: an epidemiologic perspective. American Journal of Epidemiology 120, 507-571.

[10] Pollack, M.E. 1990. Plans as Complex Mental Attitudes. In Intentions in Communication, P. Cohen, J. Morgan and M. Pollack Eds. MIT Press, Cambridge, MA, 77-102.

[11] R Development Core Team. 2008. R: A Language and Environment for Statistical Computing, at http://www.R-project.org.

[12] Rich, C. and Sidner, C.L. 1997. COLLAGEN: When Agents Collaborate with People. In Proceedings of Autonomous Agents '97, Marina Del Rey, CA, 284-291.

[13] Schulman, D., Sharma, M. and Bickmore, T. 2008. The Identification of Users by Relational Agents. In Proceedings of the Autonomous Agents and Multi Agent Systems (AAMAS), Estoril, Portugal2008.

[14] Spiessens, B., Lesaffre, E., Verbeke, G., Kim, K. and Demets, D.L. 2000. An overview of group sequential methods in longitudinal clinical trials. Stat Methods Med Res 9, 497-515.

[15] Thomas, S., Reading, J. and Shephard, R. 1992. Revision of the Physical Activity Readiness Questionnaire (PAR-Q). Canadian J Sports Sci. *17*, 338-345.

[16] Verbeke, G. and Molenberghs, G. 2001. Linear Mixed Models for Longitudinal Data. Springer.

[17] Whitehead, J. 1997. The Design and Analysis of Sequential Clinical Trials, 2.Rev.Ed. Wiley.

[18] Zeger, S., Liang, K. and Albert, P. 1988. Models for Longitudinal Data: A Generalized Estimating Equation Approach. Biometrics 44, 1049-1060.