Experimental Design for Human-Robot Interaction with Assistive Technology

Katherine M. Tsui and Holly A. Yanco
University of Massachusetts Lowell
Department of Computer Science
One University Avenue
Lowell, MA 01854
ktsui@cs.uml.edu
holly@cs.uml.edu

David Kontak and Linda Beliveau
Crotched Mountain Rehabilitation Center
Assistive Technology Unit
One Verney Drive
Greenfield, NH 03047
david.kontak@crotchedmountain.org
linda.beliveau@crotchedmountain.org

ABSTRACT

Experimental design for human subject research must consider many factors. Who is the end-user? Is this population accessible for user testing? How many conditions will be tested? How many participants are needed? What types of data can be collected? How should the data be analyzed?

This paper surveys experiments conducted in human-robot interaction (HRI), assistive technology (AT), and human-robot interaction with assisitve technology (HRI-AT). To illustrate complexities in experimental design in HRI-AT, we present a case study of our two experiments – a preliminary user trial with able-bodied subjects, and a recreation of the first experiment modified for cognitively impaired wheelchair users. Also, we discuss important points learned while conducting an evaluation with cognitively impaired wheelchair patients.

Categories and Subject Descriptors

H.5.2 [**Information Systems**]: Information Interfaces and Presentation—*User Interfaces*

1. INTRODUCTION

Experimental design for any human subject research has many facets to consider. There are a number of experiment types including controlled experiments, observational studies, and surveys. The type of participants in studies vary largely; for example, undergraduate college students, emergency responders, or pre-school children with developmental disabilities. The duration of a study can be a few minutes, hours, weeks, months, or years. Data collection can be both direct and indirect; for example, direct methods include pre- and post-experiment questionnaires, task completion time, measurement of cognitive workload, while indirect methods include post-hoc analysis and coding from video recordings.

Experimental design in assistive technology borrows heavily from clinical trials for medical devices, as they have an established protocol. The Good Clinical Practice Protocol requires clearly stated objectives, checkpoints, and types and frequency of measurement; detailed description of the proposed study; description of preventative biasing measures; expected duration of the trial and treatment regiment; "discontinuation criteria" for subjects or the partial/whole trial; and record keeping strategies [30].

Experimental design in human-robot interaction is not quite as well established as clinical trials. However, it borrows from human-computer interaction (HCI), computer supported cooperative work (CSCW), human factors, psychology, and social sciences. Drury, Scholtz, and Kieras applied GOMS (goals, operators, methods, selection rules) [3] analysis from HCI to human-robot interaction [6]. Drury, Scholtz, and Yanco employed the "think aloud" [8] method from HCI and coding from psychology [7]. Humphrey et al. used NASA-Task Load Index [13] from human factors [15].

Experimental design at the cross-section of human-robot interaction and assistive technology is more complex due to the unique abilities of the people, thus generalizations cannot be made. Experimental design must consider a patient's physical, cognitive, and behavioral ability. When executing a testing session however, the quality of the data and length of the session is dependent upon the patient's mood, attentiveness, and afflictions. Thus, observational studies are common and results are typically anecdotal.

This paper surveys experiments conducted in human-robot interaction (HRI), assistive technology (AT), and human-robot interaction with assisitve technology (HRI-AT). To illustrate the complexities of experimental design for HRI-AT, we present details of our experiments with a robot arm.

2. BACKGROUND AND RELATED WORK

An experiment is defined as "a test or procedure carried out under controlled conditions to determine the validity of a hypothesis or make a discovery" [5]. Controlled experiments are used to compare n number of conditions, as in an AB-style experiment. Users participate in all conditions in a within subjects study. In a between subjects study, users participate in only one condition as a new group of users is needed for each variable tested. Hypotheses answered with a controlled experiment require quantitative data for statistical significance, such as time to task completion. Controlled experiments are widely used in human subject research.

We borrow the term "observational study" from psychology and social sciences. Derived from [33], we define "observational study" as "a correlational research study that involves repeated observations over a long period of time, often months or years." A single condition per participant is tested for the duration of the study. Observational studies are also widely used in human subject research.

We borrow the term "clinical trial" from the medical field. Derived from [32], we define "clinical evaluation" as "a comparison test of medical device or treatment, versus other devices or the standard medical treatment for a patient's condition." Clinical evaluations, like controlled experiments, compare conditions. Typically, a control group serves as a baseline. Clinical evaluations have been largely used in assistive technology and rehabilitation robotics.

In this section, we present a survey of HRI, AT, and HRI-AT experiments. For each paper, we describe the experiment type, number of participants, type of participants, duration of study, task, data collected, and analysis techniques.

2.1 Human-Robot Interaction Experiments

Controlled experiments and observational studies are common types of experiments conducted in HRI. A number of HRI studies have been conducted in areas such as remote robot operation, robot appearance preference, and robot interaction as a teammate. A sampling of three HRI studies which represent the spectrum of methodology, type of users, number of users, and data collected are outlined below.

2.1.1 Controlled experiments

Keyes et al. conducted two experiments to evaluate camera placement in remote robot operation [16]. Two unique robots and respective camera views were evaluated by nineteen subjects each. Camera configuration was randomized between overhead versus forward facing camera view and switchable forward-rear versus forward facing only camera view. Users operated the remote robot in a known maze from a start point, randomized from three locations; they were instructed to visit three checkpoints. Data collected included number of collisions, time to checkpoints, time to task completion, command logs, and pre- and post-experiment surveys. From the command logs, time spent using a given camera view and the number of switches were also extracted. Significance was determined using paired t-tests on task completion time, number of cameras, and number of collisions. An anecdotal analysis was also given.

2.1.2 Observational studies

Kozima and Nakagawa conducted a longitudinal study of typically-developing preschool children interacting with Keepon [17]. The interaction between Keepon and twenty seven three-year-old children lasted for one year, composed of twenty five three-hour sessions. Each three-hour sessions was comprised of ninety minutes of free interaction and ninety minutes where Keepon supplemented the teaching activities. Observations included the child's interactions, emotional display, and utterances to or about Keepon; the presented results were anecdotal.

Forlizzi conducted an ethnographic study of floor cleaning as a family activity [11]. Each member of six families was interviewed about household cleaning specifics including products, processes, and frequency. The homemaker of each family completed twelve entries in a visual diary of events that make the floor dirty; they also documented the products and services used for floor cleaning, in addition to their mood. Then the families were divided into those with elders and those without and each homemaker was given a vacuum. A Roomba Discovery vacuuming robot was given to one family with elders and two families without. A Hoover Flair! Upright vacuum was given to two families with elders and one family without. Again, the homemaker completed twelve entries in a visual diary and described if their given vacuum assisted. Follow-up interviews were conducted with each member of the families. The interviews, notes, and visual diaries yielded categories of functionality, aesthetics, and symbolism. The data was coded and analyzed using these categories. The presented results were anecdotal.

2.1.3 Discussion

HRI evaluation must deal with non-deterministic robot behavior and relative platform fragility. Thus, repeatability and the large number of trials in the HCI domain are not easy to come by in HRI. However, experiments with able-bodied people can be run with any number of participants, from six expert subjects ranging to over one hundred subjects for general users. The length of an experiment can range from five minutes to a few hours, given proper compensation. Thus, AB-style experiments are commonly conducted given the accessibility to participants.

Quantitative data can also be derived from observational studies. For example, in Kozima and Nakagawa's study interaction time with Keepon per child and eye gaze could be directly measured. As seen in Keyes et al.'s study, qualitative data can be found in controlled experiments through post-experiment surveys.

2.2 Assistive Technology Experiments

Taking a cue from the medical field, clinical trials are commonly found in assistive technology and rehabilitation robotics. However, observational studies can provide a feasibility trial for a developing assistive device on which a future clinical trial can be based. Controlled experiments provide more rigor than an observational study and can illuminate the adoptability of an assistive device. A number of assistive technology studies have been conducted, and a sampling of four studies are outlined below. They show the spectrum of methodology, number of end-users, data collected, and statistical analysis.

2.2.1 Controlled experiments

Davis et al. conducted a longitudinal study of children with impaired communication [4]. Of the twelve children, ages five through eleven, ten were either diagnosed with autism or displayed autistic behavior. The children played a "fill in the blank" story computer game, called TouchStory. Each child participated in twelve TouchStory units over a period of five months. There were two phases of this trial. The non-adaptive phase established a routine of using TouchStory, where the number of story types was fixed. In the adaptive phase, the number of story types varied based on the child's previous success to create an enjoyable, challenging experience. Experimenter's notes were taken for each session, in addition to the automated logs generated by TouchStory. Percentage correct over all trials for each child and story type was extracted from the logs. Significance of the children's narrative comprehension task scores was determined using Spearman rank correlation [20].

2.2.2 Observational studies

In the Communication by Gaze Interaction project, Bates et al. conducted a case study of four-end users [1]. The users consisted of two people with ALS (Lou Gehrig's disease), one person with cerebral palsy, and one person who had suffered a stroke. Each user spent time evaluating the eye-tracking system, using it to type and communicate. Observational notes about each user's session were collected. Only qualitative results were presented in the paper.

2.2.3 Clinical evaluations

Mayogitia et al. conducted user evaluation of a stair-climbing aid [21]. Twenty-eight elderly people participated in a portion of the experiment. A testing session consisted of clinical tests, biomedical tests, and interviews. The set of clinical tests included ascent time, descent time, balance rating, and health questionnaire, and their order of administration was randomized. The biomedical tests recorded the stability of the user with and without the stair-climbing aid using accelerometer and gyroscope sensors. Ten categories were generated from assessing the interview questions, the data was coded based on these themes, and agreed upon by two independent assessors. The presented results are anecdotal with discussion including the user's first impressions, device's appearance, device's ease of use, and the user's confidence in the device. Additionally, occupational therapists provided a summary of pros and cons of the aid.

Housman et al. conducted user evaluation of arm rehabilitation in stroke patients using the Therapy Wilmington Robotic Exoskeleton (T-WREX) [14]. Eleven stroke survivors exercised with T-WREX for an hour, three times per week for eight weeks. A control group of twelve patients also exercised with a physical therapist for the same duration. Blood pressure readings and pain ratings were taken before and after each session. Time working directly with a therapist was recorded. After eight weeks, the groups switched to allow for subjective comparison. Clinical tests administered evaluated functional arm movement, quality of affected arm use, range of motion, and grip strength. Also, a quantitive survey was administered to evaluation patient satisfaction; patients indicated their therapy preference for exercising and recording progress. Paired and non-paired t-tests were used to determine significance on the Fugl-Meyer score [12], Motor Activity Log, and survey responses.

2.2.4 Discussion

In all types of assistive technology experiments, the end-user is directly involved. To ensure a successful assistive device or therapy, the end-user and relevant staff (caregiver, nurse, occupational therapist, physical therapist) must be involved in the design and testing. The power of the study increases with the trial iterations. A small number of end-users are needed in an observational study where the data is largely qualitative. More end-users are needed for a controlled experiment where the data has more quantitative data. In a clinical trial, a large number of end-users are needed in a clinical trial to provide a control group; qualitative results supplement quantitative findings.

2.3 HRI-AT Experiments

At the cross-section of human-robot interaction and assisitve technology is HRI-AT. The types of experiments conducted largely inherits from HRI. However, as with AT, end-user evaluation is more prevalent; able-bodied subjects provide an upper bound of expected performance. A number of HRI-AT studies have been conducted in areas such as autism therapy, stroke therapy, and eldercare. Six studies are outlined below, which represent the spectrum of methodology, number of end-users, data collected, and statistical analysis.

2.3.1 Controlled experiments

Tijsma et al. conducted experiments of human-robot interface with the Manus Assistive Robotic Manipulator in both a lab setting and field evaluation [26]. In the lab evaluation, sixteen able-bodied subjects participated in a 2x2 experiment (conventional mode switching versus their new mode and Cartesian mode versus "pilot" mode). The participants executed two tasks: picking up an upside-down cup and placing it right-side-up in another; and picking up a pen and placing it in the same cup. The experimental conditions were balanced using two Latin squares [2]. A third task, placing a block into a box of blocks, was used to investigate the center of rotation of the gripper (conventional versus alternative). Data collected included the number of mode switches, task time, and Rating Scale of Mental Effort [34]. Factorial ANOVA was applied for statistical significance for the first two tasks, and standard ANOVA on the third task.

In Tijsma et al.'s field trial, four end-user participants were recruited; however, the interface was successfully integrated with only two participants' wheelchair joysticks. The participants executed three tasks: picking up an upside-down cup and placing it right-side-up in another and picking up a pen and placing it in the same cup; putting two square blocks in a box of blocks; and retrieving two pens out of sight. A baseline experiment was comprised of the first task in Cartesian mode and the second task in the conventional center of rotation; the third task was not part of the baseline evaluation. Due to fatigue, the participants were only able to perform one trial per experimental condition. Data collected included the number of mode switches, task time, Rating Scale of Mental Effort (at 5, 10, 20, and 40 minutes), and survey responses. Field study results were anecdotal due to the small sample size and insufficient data.

2.3.2 Observational studies

Wada et al. conducted a longitudinal study of the therapeutic effects of Paro at an elderly day service center [31]. Twenty three elderly women, age seventy three to ninety three, volunteered or were selected to participate. The women interacted with Paro for twenty minute blocks for five weeks, one to three times per week, in groups of eight or less. Data collected included a self assessment of the participant's mood (pictorial Likert scale of 1 (happy) to 20 (sad)) before and after the interaction with Paro; questions from the Profile of Mood States questionnaire to evaluate anxiety, depression, and vigor (Likert scale of 0 (none) to 4 (extremely)); urinary specimens; and comments from the nursing staff. Wilcoxson's sign rank sum test [20] was applied to the mood scores to determine significance.

Wada et al. also examined the effects of Paro on the nursing staff with respect to burnout. Over a period of six weeks, four female and two male staff members participated in the burnout scale questionnaire once per week. Friedman's test was used to determine statistical significance on the total average score of the burnout scale.

Kozima et al. have also used Keepon to studied social interactions in children with developmental disorders [18]. A longitudinal study was conducted for over eighteen months with a group of children, ages two to four, at a day-care center. Keepon was placed in the playroom. In a three-hour remedial session, the children could play with Keepon during free play. During group activities, Keepon was moved to the corner. The paper details a case study of two autistic children in an anecdotal fashion. The first case describes the emergence of a dyadic relationship of a girl with Kanner-type autism over five months with Keepon. The second case describes the emergence of a interpersonal relationship between a three-year old girl also with Kanner-type autism, her mother or nurse, and Keepon over eighteen months.

Robins et al. studied the effect of exposure to a robot doll, Robota, over a long period of time on social interaction skills of autistic children [24]. Four children, ages five through ten, were selected by their teacher to participate in this longitudinal study. Over a period of several months, the child interacted with the robot doll as many times as possible in an unconstrained environment. Trials lasted as long as the child was comfortable and ended when the child wanted to leave or was bored. In the familiarization phase, the robot doll danced to pre-recoded music. In the learning phase, the teacher showed the child that the robot doll would imitate his movements. Free interaction was similar to the learning phase without the teacher. A post-hoc analysis of video footage of interaction sessions yielded eye gaze, touch, imitation, and proximity categories. All video data was coded on one second intervals using these four categories. An extension study investigated the preference of robot doll appearance (pretty versus plain).

Scassellati also investigated human-robot interaction with autistic children [25]. In a pilot experiment, seven children with autism and six typically-developing children watched a robotic face change shape and make sounds. The robot functioned in two modes: scripted and teleoperated. The session began with the script where the robot face "woke up," asked some questions, then "fell asleep." Then the operator manually controlled the robot face. Data collected included social cues such as gaze direction. Eye gaze was analyzed in each frame to determine the primary location of focus. The focus points were used train a linear classifier used to generate predictive models.

Michaud et al. conducted an exploratory study of low-functioning autistic children with a sixty centimeter tall humanoid robot [23]. Four autistic children, age five, participated in a seven week study. Each child played with Tito three times per week for five minutes. In a session, the robot asked the child to imitate actions including smiling, saying hello, pointing to an object, moving their arms, and moving forwards and backwards. The child's favorite toy was placed in the room with the robot. Data collected included video and automated interaction logs. The interactions were categorized into shared attention, shared conventions, and absence of shared attention or conventions; all video data was coded using twelve second windows. The coding was completed by two evaluators with a confidence of 95%.

2.3.3 Discussion

In HRI-AT, the results tend to generally be more qualitative due to the uniqueness of the patients within a population. The data from a session may be skewed due to the patient's mood, their anxiety level, pain, sleepiness, etc. The Profile of Mood States questionnaire can be used in self evaluation, but largely, it is the subjective notes of an observer that capture the patient's unusual behaviors and feelings. However, quantitative analysis is still possible using measures such as coding and interaction time.

3. HRI-AT: CASE STUDY

Our research focuses on providing methods for independent manipulation of unstructured environments to wheelchair users utilizing a wheelchair mounted robot arm

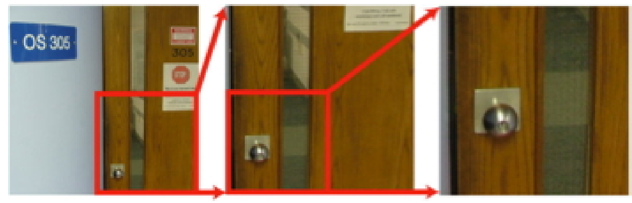

Figure 1: The user "zooms in" on the doorknob using progressive quartering. The red box indicates the selected region which contains the desired object.

for manipulation. We hypothesize that a vision-based interface is easier to use than a menu-based system. With greater levels of autonomy, less user input is necessary for control. Thus, by explicitly designating the end goal, the end user population can be expanded to include low functioning wheelchair users. Our target audience is disabled people with cognitive impairments.

We conducted a preliminary experiment using able-bodied subjects as an evaluation baseline in August 2006. We conducted field trials with cognitively impaired wheelchair users in August and September 2007. To lend power to our evaluation with our target population, we conducted a hybrid observational evaluation; that is, we conducted a controlled experiment with four conditions, and ran the experiment for eight weeks. The subjects participated as frequently as possible with sessions ranging from one encounter to eight.

3.1 Hardware

The Manus Assistive Robotic Manipulator (ARM) is a wheelchair mounted robotic arm manufactured by Exact Dynamics [9]. It has a two-fingered gripper end-effector and is a 6+2 degree of freedom unit with joint encoders. A user may manually control the Manus ARM by accessing menus keypad, joystick, or single switch. A vision system with two cameras has been added to improve user interaction with the Manus ARM. A pan-tilt-zoom camera at the shoulder provides the perspective of the wheelchair occupant for the interface. A small camera mounted within the gripper provides a close up view for the computer control.

3.2 Single Switch User Interface

3.2.1 Interface Design

We assumed that single switch scanning was the lowest common denominator for all patients in our target audience as there are many options for switch sites, including hands, head, mouth, feet, upper extremities, lower extremities, and mind [19]. Thus, we created a visual interface with text-based prompts which used a single switch [27]. A conceptual flow diagram is shown in figure 1.

3.2.2 User Trials

Participants. Twelve able-bodied participants (ten men, two women) were recruited for an AB-style alternating condition experiment; details in [27]. Participants were asked demographic information in the pre-experiment questionnaire used to uncover skill biases. In the post-experiment questionnaire, participants were asked about their experiences in an open-ended fashion and in Likert scale ratings.

Task and Trials. The participant was instructed to move the Manus ARM from its folded position towards a

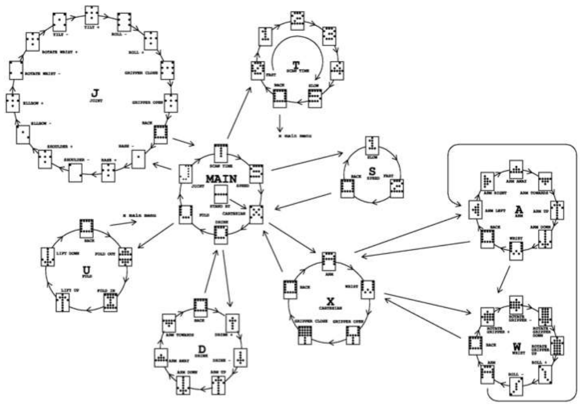

Figure 2: The Manus ARM is controlled using menu hierarchies. The single switch menu, shown above, is two or three layers deep depending upon the desired functionality. Also, the timing component inherent to single switch applications is indicated as a clockwise cycle and current state is depicted on a small LED matrix. (Courtesy of Exact Dynamics.)

specified target. This positioning task was repeated six times. The entire process took approximately ninety minutes per participant, including training on each condition and pre- and post-experiment questionnaires.

Two conditions were tested: menu control (commercial) and computer control (our interface). An equal number of start conditions were generated prior to all user testing and the control condition was alternated for each of the remaining runs. The user participated in three runs per condition to counteract any learning effect. The user was trained in each condition until he/she felt comfortable.

The input device was kept constant across conditions. The single switch menu was used for menu control. For computer control, the user pressed the switch to "zoom in" on the desired object. Six of eight possible targets were chosen at random prior to all experiments for all twelve sequences.

Data Collection and Analysis. Data was collected from questionnaires (pre- and post-experiment), video of the Manus ARM movement, video from over the participant's shoulder, task completion time, trial failures, participant technique, distance between the gripper camera and the center of the desired object.

The number of clicks executed by participants was recorded in a post-hoc analysis of the video footage. Workload was defined as the number of clicks during the run divided by the run time, which is the average clicks per second. From the observer notes, the distance to time ratio was extracted, which provides a means of cost analysis: moving X distance takes Y time. Paired t-tests were used to show statistical significance in workload and cost analysis for the computer control versus manual control.

The post-experiment survey asked both open ended and Likert scale rating questions, and solicited for interface improvement suggestions. Participants were asked which feature they liked the most and why, and similarly for their least favorite. They were asked to rate their experience with each interface using a Likert scale from 1 to 5, where 1 indicates most positive. The survey data yielded both qualitative and quantitative (albeit subjective) results. The interface experience rating served as a cross-check for the open ended interface preference.

Figure 3: The user selects an object from the scene using our flexible interface. The fixed camera view selection is shown on the left. The moving camera view is shown center and right, using joystick and touch screen, respectively. (Best viewed in color.)

3.3 Flexible Interface

The single switch interface was an initial prototype. However, the underlying assumption of using single switch scanning as a lowest common denominator across populations was incorrect. Single switch scanning inherently involves a timing component, thus is cognitively intensive and not conducive for use with our target population of cognitively impaired wheelchair users. The interface was reworked to use a touch screen or joystick as input.

3.3.1 *Interface Design*

Accessibility is a challenge for people with disabilities. Differences in cognitive ability, sensory impairments, motor dexterity, and behavioral and social skills must be taken into account when designing interfaces for assistive devices [29]. Our current work proposes that flexible interfaces tuned for the individual, instead of completely custom built solutions, may benefit a larger number of people [28]. The assistive technology director at Crotched Mountain identified touch screens and joysticks as input devices that covered a large portion of the patients.

Using these input devices, the user selects an object in the scene shown on the touch screen. There are two modes for object selection: fixed camera and moving camera. When the shoulder camera is in a fixed position, the user can touch the object on the screen or move the mouse-emulating joystick to the object, shown in figure 3 (left).

When the shoulder camera is able to be moved, the user places a "plus sign" in the center of the screen over the object by directing the camera. This mode is similar to if an object of interest lies in the periphery and we move our head to focus on it. To move the "plus sign" over the object using a joystick, the user moves to joystick in the desired direction, up, down, left, or right, shown in figure 3 (center). Moving the camera via touch screen is similar: the user presses blue buttons shown on the sides of the touch screen to move the camera up, down, left, and right, centering on the desired object, shown in figure 3 (right).

3.3.2 *End-User Trials*

The preliminary experiment from the single switch interface provided a baseline for our research. Our goal was to recreate the experiment with our target population. However, most participants did not have the cognitive ability to perform the baseline evaluation of the menu interface. Further, due to physical and cognitive abilities, not all participants could use all versions of the flexible interface. Thus, a traditional AB-style alternating condition experiment was deemed inappropriate. We adapted the experiment to a hybrid observational evaluation study. During an eight week period, eight users interacted with the robot arm as often as possible, like [24].

Participants. Participants from the Crotched Mountain School and Brain Injury Center were recruited in a different manner than in the able-bodied experiment. The director of the Assistive Technology Unit played the central role in recruitment of participants. Given all of the students and residents, the director evaluated their overall physical dexterity, cognitive ability, visual ability, and access method. Candidates were invited to participate in the experiment. Eight participants consented: three were their own guardians, and the remaining five required parental or guardian approval.

All participants had medium to high cognitively functional ability, minor vision impairment, and were able to operate a joystick or touch screen (finger isolation for pointing only). Details about the participants' visual abilities, cognitive abilities, behavioral abilities, and computer access devices can be found in [28].

Like [4], each participant was given an initial interface settings profile by the experimenters. A profile contains parameters for access method (joystick or touch screen) based on the participant's accessibility and behavioral abilities; fixed camera view or moving based on their level of cognition; cursor size based on their visual acuity; cursor speed if applicable based on their physical dexterity; and dwell length if applicable also based in their physical dexterity.

Task. We believe that the task places appropriate demands on the user. Tasks should be demanding enough to be interesting and stimulating without the negative result of being frustrating and discouraging. This is often referred to as the "just right challenge." For this experiment, matching was decided to be an appropriate task. The participant would be asked to match the flash card of an object with the object displayed on the screen in front of them.

The objects used in this experiment had real-world qualities so that the task was not deemed trivial by the participants and to satisfy educational requirements of the school. The set of objects contained a yellow mug, a blue cup, a teal stencil, a green book, a purple plush bear, a tan plush bear, a clay pot, an orange egg salt shaker, a purple gift wrapping bow, and various colored and textured balls.

People with cognitive impairments who understand the physical instance of an object, say an apple, may not be able to understand a picture of the apple. Further, those who understand the picture may not be able to understand the word "apple." Thus, abstraction can be a limiting factor. For this experiment, all participants are able to understand the actual object and also the photo of the object. To remove the possible confounding factor of semantics (e.g. "the purple bear"), an individual photograph of the object was used as a flash card. The flash card served as the primary prompt, and verbal semantics served as a secondary prompt.

According to Fitts' Law, time to move to a target area is the function of the distance to the target and the target size [10]. Thus, the size of the object directly affects the difficulty of the task. A smaller object will be more difficult to select, and conversely, a larger object will be easier to select. This allowed the experimenter to tune the difficulty to engage the user without altering the user's interface parameters.

The objects from which the participants would select were placed on a three-shelved bookshelf in one of four scenarios, according to difficulty. Size, color, and reflectivity of the objects were taken into account when designing the scenarios. The scenarios were generated by the experimenters prior to commencing user testing. The task level could be adjusted by prompting the participant to select larger or smaller objects, or the configuration could change entirely.

Trials. The experiments were conducted by assistive technology technicians at Crotched Mountain. The initial profile was used during the participant's first session. The profile was iteratively adjusted as needed each session.

At the start of a session, the experimenter placed the robot arm on the participant's right-hand side and the touch screen in front. The previous session was briefly reviewed. The experimenter then described the interface to be used in the current session, method for object selection, and arm movement. The participant was then trained on the interface. Training was necessary to minimize the learning effect.

Once comfortable, the trial began. The participant was shown a flash card of the desired object and prompted and encouraged as necessary. The participant then used the interface to select the object. A correct selection would yield a red rectangle around or on the object and a "ding" played. Otherwise, a "Please, try again!" prompt sounded. The participant made three object selections per session, which would maximize their attention to task and minimize exhaustion and behavioral issues. At the conclusion of a session, the experimenter administered a survey and removed the robot arm and touch screen. The experiment ran for eight weeks and the participants interacted with the robot arm as frequently as possible.

Data Collection and Analysis. Data was collected from video, manual logs, post-session questionnaires, and computer generated log files. Qualitative data included the post-experiment questionnaire administered at the end of each user session and the observer notes. The questionnaire posed open ended questions about which interface the user liked most to date, which interface they liked least to date, and suggestions for improving the interface. The observer notes contains relevant notes about the session, including length of reorientation.

Quantitive data included trial run time, attentiveness rating, prompting level, close up photos of the object selected, and computer generated log files. For each trial, the run time was recorded, specifically the time from object prompt to participant selection, the time from the Manus ARM movement to the object visually confirmed, and the time for the Manus ARM to complete folding. The experimenter, who is an assistive technologist professional, rated the user's prompting level per trial based on the FIM scale, where 0 indicated "no prompting needed" and 5 indicated "heavy prompting needed" [22]. The experimenter also rated the user's attentiveness to the task on a Likert scale, where 0 indicated "no attention" and 10 indicated "complete attention." Two separate scales were used because it is not necessarily the case that a person who requires high levels of prompting is unmotivated to complete the task. Paired t-tests were run on object selection time, prompting level, and attention level to determine statistical significance.

3.4 Future Work

Additional eight-week periods of user testing at Crotched Mountain will occur over the next year. We will be expanding the participant pool to include users with low cognitive function. We will evaluate improvements in our flexible interface with emphasis on cognitive workload.

4. DISCUSSION

User trials at the cross-section of human-robot interaction and assistive technology require careful experimental design. Based on our user testing experiences, we offer the following HRI-AT recommendations:

Involve the end-user. Involve the end-user as soon as possible in the design process. The interface style changed from the indirect selection (single switch scanning) to direct selection (using touch screen or joystick) due to an incorrect assumption. By involving the end-user and, in the case of assistive technology, their staff (caregiver, nurse, occupational therapist, physical therapist, etc.), a clear understanding of the users' desires and abilities grounds the project in reality.

Define the user population. People with disabilities have a wide range of physical and cognitive impairments with wide range in severity. Generalizations cannot be made over this population, thus it is imperative to choose a well defined user population. In the case of this experiment, we chose users with mild cognitive impairments and minor vision problems which resulted in a participant population of eight. In future experiments, we will add users with more cognitive impairments to determine the bounds of usability of our robot arm system.

Additionally, consider the number of participants necessary for statistical significance in a controlled experiment. When evaluating with end-users, a smaller sample size may be acceptable. If the population is too small for statistical analysis, the experiment may need to be changed to an observational study. Even then, it is difficult to make generalizations because every person is unique.

Accesibility. Accessibility is more than just a person's access device, it is the ability of a person to understand and manipulate a device. Accessibility is a significant challenge for people with disabilities. They may have a large range of strengths and challenges. These differences include differences in cognitive ability, sensory impairments, motor dexterity, and behavior and social skills [29].

Devices and interfaces designed to assist this population must take into account these differences. There are always alternative ways to accomplish a given task. Compensatory techniques can be used in all areas of accessibility. For example, procedural modification, such as increased prompting or feedback, may assist with cognitive challenges. Vision impairments may be supplemented by using auditory feedback, increased size/contrast, or tactile prompts. Alternative physical access may use voice command or other adapted interfaces, such as special mounting to compensate for reduced range of motion, the use of foot or head controls, or single switch systems.

Our flexible interface allowed for two commonplace access methods, a touch screen and joystick. Additionally, the auditory channel was used to provide feedback. Given a selection, a "ding" played if the object was decipherable to the system, and a "Please, try again!" played otherwise.

Mounting. Fast setup and break down time are essential to good user testing. Often a participant will lose motivation or feel frustrated if the setup time takes too long. Since the session length averaged thirty minutes, we wanted to mount as generically as possible.[1] Thus, the ARM was mounted to the wheelchair with casters and clamps. However, unforeseen issues did occur with the joystick and touch screen placement. The joystick was initially placed on a lap tray, instead of mounting the joystick to the participant's wheelchair. Some users had shallow laps and therefore could not use the lap tray easily. Eventually, a small, adjustable height side tray was used to hold the joystick. Still, some power wheelchair users had difficulty positioning the joystick because their wheelchair joystick was already placed in its optimal position on their dominant hand. This would result in having two abutting joysticks. Most users were able to reposition their arm and hand to operate the non-driving joystick. However, there was one person who was not able to participate in the user trials because he did not have adequate motor skills in his right arm to accommodate the awkwardly placed non-driving joystick.

Appropriately challenging task. A task that is too easy will bore a user; one that is too difficult is frustrating. Either can cause the user to feel disdain towards the experiment or underlying system. For our system, it was determined that not all users would be able to use the original menu hierarchy to control the robot arm. Therefore, the baseline was removed as an interface option. Using the original menu hierarchy would likely cause frustration, and the users might dislike the robot arm system overall.

Motivation and interest. It was imperative not to bore or overwhelm the participants with the task because we will be working with this patient population in the future. The matching task was suitable because it was game-like and the level of difficulty was customizable to the participant. If necessary, the generic objects could be replaced with personal items, thereby increasing their motivation.

The participants' reasons for being in the study differed. Some thought the robot arm was fun and were excited about using new technology. Others wanted to help with the ongoing research. Many asked questions about the inner workings of the robot arm and about the system developers. It is imperative that the users not feel obligated to continue in the study. After each session, the participants were always asked if they would like to come again to use the robot arm.

Qualitative and quantitative data. When conducting experiments in HRI-AT, it is imperative to collect both qualitative and quantitative data. Qualitative data can be derived, for example, from interviews with the participant, observer notes about a trial, and caretaker observations of the participant. Quantitative data is more difficult to obtain. Ease of use and user workload may be derived from task completion time and commands issued. Coding may also provide quantitative data.

5. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (IIS-0534364). Dr. Aman Behal of the University of Central Florida is a collaborator on this research. Special thanks to all the Crotched Mountain participants, Tae-Young Park of Crotched Mountain, and Philip Thoren, Mark Micire, Munjal Desai, Harold Bufford, and Jeremy Badesa (all of UMass Lowell).

[1]The duration of a session in future trials will be longer; we expect sessions to last multiple hours or days. At this point, we will interface with the patients' drive joysticks and fully mount the robot and touch screen to their wheelchairs.

6. REFERENCES

[1] R. Bates, M. Donegan, H. O. Istance, J. P. Hansen, and K. J. Raihia. *Designing Accessible Technology*, chapter 8, pages 77–84. Springer-Verlag, 2006.

[2] J. Bradley. Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design. *Journal of the American Statistical Association*, 53(282):525–528, 1958.

[3] S. Card, T. Moran, and A. Newell. *The Psychology of Human-computer Interaction*. Erlbaum, 1983.

[4] M. Davis, K. Dautenhahn, C. Nehaniv, and S. Powell. *Designing Accessible Technology*, chapter 11, pages 101–114. Springer-Verlag, 2006.

[5] Dictionary.com. Experiment. In *http://dictionary.reference.com/browse/experiment*, December 2007.

[6] J. Drury, J. Scholtz, and D. Kieras. Adapting GOMS to model human-robot interaction. *ACM SIGCHI/SIGART Human-Robot Interaction*, pages 41–48, 2007.

[7] J. Drury, J. Scholtz, and H. Yanco. Applying CSCW and HCI Techniques to Human-Robot Interaction. *Proc. CHI 2004 Workshop on Shaping Human-Robot Interaction*, pages 13–16, 2004.

[8] K. Ericcson and H. Simon. Verbal reports as data. *Psychological Review*, 87:215–251, 1980.

[9] Exact Dynamics. Assistive robotic manipulator. In *http://www.exactdynamics.nl/*, September 2007.

[10] P. Fitts and R. Deninger. SR compatibility: correspondence among paired elements within stimulus and response codes. *Journal of Experimental Psychology*, 48(6):483–92, 1954.

[11] J. Forlizzi. How robotic products become social products: An ethnographic study of robotic products in the home. In *Human Robot Interaction Conference*, pages 129–136, March 2007.

[12] D. Gladstone, C. Danells, and S. Black. The Fugl-Meyer Assessment of Motor Recovery after Stroke: A Critical Review of Its Measurement Properties. *Neurorehabilitation and Neural Repair*, 16(3):232, 2002.

[13] S. Hart and L. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*, 1:139–183, 1988.

[14] S. J. Housman, V. Le, T. Rahman, R. J. Sanchez, and D. J. Reinkensmeyer. Arm-training with t-wrex after chronic stroke: Preliminary results of a randomized controlled trial. In *IEEE 10th International Conference on Rehabilitation Robotics*, June 2007.

[15] C. Humphrey, C. Henk, G. Sewell, B. Williams, and J. Adams. Assessing the scalability of a multiple robot interface. *ACM SIGCHI/SIGART Human-Robot Interaction*, pages 239–246, 2007.

[16] B. Keyes, R. Casey, H. A. Yanco, B. A. Maxwell, and Y. Georgiev. Camera placement and multi-camera fusion for remote robot operation. In *IEEE International Workshop on Safety, Security, and Rescue Robotics*, August 2006.

[17] H. Kozima and C. Nakagawa. Longitudinal child-robot interaction at preschool. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, pages 27–32, March 2007.

[18] H. Kozima, C. Nakagawa, and Y. Yasuda. Interactive robots for communication-care: a case study in autism therapy. In *ROMAN*, pages 341–346, August 2005.

[19] M. L. Lange. Switch assessment: Determining type and location. In *Canadian Seating & Mobility Conference*, 2006.

[20] R. Langley. *Practical Statistics Simply Explained*. Courier Dover Publications, 1971.

[21] R. E. Mayagoitia, S. Kitchen, J. Harding, R. King, and A. Turner-Smith. *Designing Accessible Technology*, chapter 13, pages 127–134. Springer-Verlag, 2006.

[22] MedFriendly. Medfriendly.com: Functional independence measure. In *http://www.medfriendly.com/functionalindependencemeasure.html*, September 2007.

[23] F. Michaud, T. Salter, A. Duquette, H. Mercier, M. Lauria, H. Larouche, and F. Larose. Assistive technologies and child-robot interaction. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, March 2007.

[24] B. Robins, K. Dautenhahn, R. te Boekhorst, and A. Billard. Robots as assistive technology – does appearance matter? In *IEEE International Workshop on Robot and Human Interactive Communication*, September 2004.

[25] B. Scassellati. Quantitative metrics of social response for autism diagnosis. *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pages 585–590, 2005.

[26] H. A. Tijsma, F. Liefhebber, and J. L. Herder. Evaluation of new user interface features for the manus robot arm. In *Ninth International Conference on Rehabilitation Robotics*, pages 258–263, June 2005.

[27] K. Tsui and H. Yanco. Simplifying wheelchair mounted robotic arm control with a visual interface. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, pages 247–251, March 2007.

[28] K. Tsui, H. Yanco, D. Kontak, and L. Beliveau. Development and evaluation of a flexible interface for a wheelchair mounted robotic arm. In *Human-Robot Interaction Conference*, March 2008.

[29] University of Texas Austin. Accessibility and disability defined. In *http://www.utexas.edu/learn/accessibility/disability.html*, December 2007.

[30] US Food and Drug Administration. Guidance for industry, E6 good clinical practice: consolidated guidance. *Federal Register*, 10:25691–709, 1997.

[31] K. Wada, T. Shibata, T. Saito, and K. Tanie. Effects of robot-assisted activity for elderly people and nurses at a day service center. *IEEE*, 92(11):1780–1788, 2004.

[32] Wikipedia.org. Clinical trial. In *http://en.wikipedia.org/wiki/Clinical_trial*, January 2008.

[33] Wikipedia.org. Longitudinal study. In *http://en.wikipedia.org/wiki/Longitudinal_study*, January 2008.

[34] F. Zijlstra. *Efficiency in work behaviour: A design approach for modern tools*. PhD thesis, Delft University, 1993.