Use of Large Sample Sizes and Multiple Evaluation Methods in Human-Robot Interaction Experimentation

Cindy L. Bethel
Department of Computer Science and Engineering
University of South Florida
Tampa, FL
cbethel@gmail.com

Robin R. Murphy
Center for Robot-Assisted Search and Rescue
Texas A & M University
College Station, TX
murphy@cse.tamu.edu

Abstract

This paper presents details on planning and designing human studies for Human-Robot Interaction. There is a discussion of the importance of using large sample sizes to better represent the populations being investigated in order to have a better chance of obtaining statistically significant results for small to medium effects. Coverage of the four primary methods of evaluation are presented: (1) self-assessments, (2) behavioral observations, (3) psychophysiological measures, and (4) task performance metrics. The paper discusses the importance of using multiple methods of evaluation in order to have reliable and accurate results and to obtain convergent validity. Recommendations for planning and designing a large-scale, complex human study are detailed as well as lessons learned from a recent study that was conducted using 128 participants, four methods of evaluation, and a high fidelity, simulated disaster site.

Introduction

Human-Robot Interaction (HRI) is a rapidly advancing area of research, and as such there is a growing need for strong experimental designs and methods of evaluation. This brings credibility and validity to scientific research that involves humans as subjects such as observed in psychology and the social sciences. Two primary issues observed in HRI studies is the lack of significant sized participant pools that closely represent the populations being studied and the lack of multiple methods of assessment used to obtain convergent validity in HRI studies (Kidd and Breazeal 2005), (Elmes, Kantowitz, and Roediger III 2006), (Johnson and Christensen 2004).

The focus until recently in HRI was on the development of specific robotic systems and applications instead of methods of evaluation and metrics. Some methods of testing and evaluation have been adopted and/or modified from such fields as human-computer interaction, psychology, and social sciences (Kidd and Breazeal 2005); however, the manner in which a human interacts with a robot is similar but not identical to interactions between a human and a computer or a human interacting with another human. As robots become more prevalent in daily life, it will be increasingly important to have accurate methods of evaluating how humans feel about their interactions with robots and how they interpret the actions of the robots (Bethel et al. 2007b).

There are four primary methods of evaluation used for human studies in HRI: (1) self-assessments, (2) behavioral measures, (3) psychophysiological measures, and (4) task performance metrics. The most common methods utilized in HRI studies are self-assessment and behavioral measures. There is limited research in the use of psychophysiological measures and task performance metrics. Each method has its advantages and disadvantages; however some of the disadvantages can be overcome by using more than one method of evaluation (Kidd and Breazeal 2005), (Bethel et al. 2007b).

The design of a quality research study for use in HRI applications that produce results that are verifiable, reliable, and reproducible is a major challenge. The use of a single method of measurement is not sufficient to interpret accurately the responses of participants to a robot with which they are interacting. Steinfeld *et al.* (Steinfeld et al. 2006) describe the need for the development of common metrics as an open research issue in HRI. They discuss an approach of developing common metrics for HRI; however this approach is oriented more toward an engineering perspective and does not completely address the social interaction perspective. Both the engineering and social interaction perspectives require further investigation in order to develop common metrics and methods of evaluation.

This paper provides examples of the experimental methods and design used for a recent large-scale, complex human study in HRI using 128 participants, four methods of evaluation, in a high fidelity, simulated disaster site. The focus of this study was to determine if humans interacting in close proximity with non-anthropomorphic robots would view interactions as more positively and calming when the robots were operated in an emotive mode versus a standard, non-emotive mode. Recommendations are provided on items to consider when developing this type of human study, how to plan and execute the study design, and some lessons learned while conducting this type of study.

The paper begins with a discussion of some related work on experimental designs and methods used in HRI. Next the paper details the process of planning and designing a human study in HRI in the Planning and Study Design section, which covers what type of study to use, how participant

levels are determined, methods and measures of assessment (advantages and disadvantages), designing a high fidelity study site, selection of robots and other equipment, finding assistants to run the study, methods to recruit the required number of participants, how to develop contingency plans, and how to plan for and deal with failures. The Additional Recommendations section presents problems that may arise and recommendations for improvements when conducting future studies. The Conclusions section summarizes recommendations for designing and conducting large-scale, complex human studies in HRI using appropriate samples sizes and multiple methods of evaluation.

Survey of Human Studies for HRI

This section summarizes a representation of previous human studies conducted in HRI that employ at least one of the various methods of evaluation and in some cases more than one method was utilized in the studies. One issue observed with these studies are the sample sizes are relatively small and therefore may not have been representative of the population being investigated which may have influenced the results.

The most commonly used method of evaluation in HRI studies has been self-assessments. In general, most of the studies in HRI include some form of questionnaires; however in some cases, the researchers will add other methods of assessment such as video observations and coding, psychophysiology measurements, and/or task performance measures. One of the more comprehensive studies was performed by Dautenhahn *et al.* (Dautenhahn et al. 2006) in which they utilized self-assessments, unstructured interviews, and video observations from one camera angle. The study included 39 participants from a conference venue and 15 participants that were involved in a follow-up study in a controlled laboratory environment. In this study, the researchers were able to obtain statistically significant results.

Another study that incorporated multiple methods of evaluation was performed by Moshkina and Arkin (Moshkina and Arkin 2005) in which they used self-assessment measures including a measure commonly used in psychology studies called the Positive and Negative Affect Schedule (PANAS) (Watson, Clark, and Tellegen 1988). Additionally, video observation and coding were performed though results were not presented. Their study included 20 participants in a controlled, laboratory setting. The results for this study were mixed and may have been more conclusive had a larger sample size been used.

A study conducted by Mutlu *et al.* used both task performance measures and multiple self-assessments. The sample size for this study was 20 participants and the results were mixed. One hypothesis showed statistically significant results; however some of the other items of interest were not statistically significant. The results may have been different had a larger participant pool been utilized. The use of larger samples sizes makes it possible for smaller effects to be discovered with significance.

One of the largest studies to date in HRI using psychophysiology measurements along with self-assessments was conducted by Kulić and Croft (Kulić and Croft 2006) with a sample size of 36 participants. Multiple psychophysiological signals were measured which is highly recommended for this type of study for reliability and validity in the results (Bethel et al. 2007a), (Itoh et al. 2006), (Kidd and Breazeal 2005), (Liu, Rani, and Sarkar 2006), (Picard, Vyzas, and Healey 2001), (Rani et al. 2004). As a result of having a larger participant pool than other previous studies, statistically significant results were found in addition to the ability to determine the best psychophysiology measures to use for determining valence and arousal responses from participants. The results may have been even more prominent with an even larger sample size. Additionally, some of the psychophysiological signals that were concluded to be not good indicators of valence and arousal may have actually had a smaller effect size and may have shown different results with a larger sample size.

Mutlu *et al.* (Mutlu et al. 2006) conducted a study using two groups with the first having 24 participants and the second having 26 participants for a total sample size of 50. The study relied heavily on the use of self-assessments developed in other studies and adapted from psychology. In this study participants competed or cooperated with other participants or with an ASIMO robot in playing a game. The study found that several of the human-human interaction scales were not useful in human-robot interaction activities. The results may have been different had a larger sample size been utilized which was also mentioned in the conclusions for this paper.

It is clear from previous studies conducted to date in HRI that standards need to be established for conducting reliable and quality studies where methods of measurement can be validated for use by the HRI community. Making sure multiple methods of evaluation are used is essential in establishing study validity. Additionally, it is important to determine the appropriate sample size necessary to obtain statistically significant results. This can be accomplished with careful planning and using study design techniques already in use in the psychology and social science communities.

Planning and Study Design

A successful human study of any type requires careful planning and design. There are many factors that need to be considered. When planning a human study the following are some common questions to consider:

- What type of study will be conducted (within-subjects, between-subjects, mixed-model, etc.)?
- How many groups will be in the study?
- How many participants per group will be required?
- What methods of evaluation will be used (self-assessments, behavioral measures, psychophysiological measures, task performance, etc.)?
- What types of tasks will the participants perform or observe (Study Protocol)?
- What type of environment and space is required to conduct the study (field, laboratory, virtual, etc.)?
- What type of equipment will be needed (robots, computers, measurement equipment, recording devices, etc.) ?

Figure 1: The Robots: Inuktun Extreme-VGTV (left) and iRobot Packbot Scout (right).

- What type of assistance will be needed to conduct the study?
- How will participants be recruited for the study?
- How will contingencies and failures be handled?

Type of Study and Number of Groups

The first step in designing a human study is to determine what question(s) are being investigated. This will assist the researcher in determining how many groups are needed and whether the study should be a within-subjects, between-subjects, or a mixed model approach. The most comprehensive approach is the within-subjects design in which participants experience all of the factors being manipulated. This type of study is the most comprehensive, increases statistical power, and reduces error variance; however, participants may experience a concept known as habituation in which participants' responses are reduced due to repetitive presentation of the same or similar stimuli. In a between-subjects design participants are placed into only one group and experience one set of factors being investigated. The results between groups are then compared. A mixed-model factorial design will include some factors or variables as within-subjects and some are set up as between-subject.

An example of a mixed-model factorial design is the study that was just completed, where the between-subjects factor was the operating mode (standard versus emotive) of the robots. Participants were randomly assigned to experience either the standard or emotive mode of robot operation. The within-subjects factor was robot (Inuktun Extreme-VGTV and iRobot Packbot Scout) in which every participant experienced both robots (See Figure 1). The order the robots appeared was counterbalanced, and operating mode assignments were balanced for age and gender.

Determining Sample Size

Determining the appropriate sample size appears to be a challenge in human studies in HRI. An a priori power analysis can be conducted to estimate the number of participants necessary for a study. A power analysis is a statistical calculation that can be performed to determine the appropriate number of participants needed for obtaining accurate and reliable results based on the number of groups in the study, alpha level, expected effect size, and a certain level of statistical power. There are typically tables in the appendices of most statistical books that will provide power analysis values. Additionally, there is software available online that will assist with this type of calculation (e.g., G*Power 3 software located at http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/).

A power analysis was conducted for the study discussed in this paper and based on using two groups, power of .80, a medium effect size of .35, and an alpha = .05 the calculation resulted in two groups of 64 participants for a total of 128 participants (Stevens 1999). From a preliminary analysis of the data the effect sizes were small to medium sized and had there not been such a large sample size used for the study, the results may not have been statistically significant; however that was not the case. In the self-assessment data, statistically significant results were obtained for the main effect of arousal and a three-way interaction was obtained for valence (Bethel, Salomon, and Murphy 2009).

Methods of Evaluation

There are four primary methods of evaluation used in human studies in HRI: (1) self-assessment, (2) observational or behavioral measures, (3) psychophysiology measurements, and (4) task performance metrics. Each of these methods have advantages and disadvantages; however most problems can be overcome with the use of multiple methods of evaluation. For results to be valid, accurate, and reliable at least two different forms of evaluation should be used in order to obtain convergent validity (Kidd and Breazeal 2005), (Bethel et al. 2007b), (Bethel et al. 2007a).

The use of self-assessments is one of the most commonly used methods of evaluation in HRI studies. Self-assessment measures include paper or computerized psychometric scales, questionnaires, and/or surveys. With this type of method, participants provide a personal assessment of how they felt or their motivations related to an object, situation, or interactions. Self-assessments can provide valuable information but there are often problems with validity and corroboration. Participants may not answer the questions based on how they are feeling at the time but rather respond based on how they feel others would answer the questions or in a way they think the researcher wants them answered. Another issue with self-assessment measures is that observers are unable to immediately and directly corroborate the information provided by participants (Elmes, Kantowitz, and Roediger III 2006). Participants may not be in touch with what they are feeling about the object, situation, and/or interaction and therefore may not report their true feelings. Also, the responses could be influenced by participants' mood and state of mind on the day of the study (Elmes, Kantowitz, and Roediger III 2006), (Johnson and Christensen 2004). For these reasons, it is important to perform additional types of measurements such as behavioral, task performance and/or psychophysiological measures to add another dimension of understanding of participants' responses and performance in HRI studies (Bethel et al. 2007b).

Behavioral measures are probably the second most common method of evaluation in HRI studies and sometimes included along with psychophysiological evaluations and participants' self-assessment responses for obtaining conver-

gent validity. Johnson and Christensen (Johnson and Christensen 2004) define observation as "the watching of behavioral patterns of people in certain situations to obtain information about the phenomenon of interest." The "Hawthorne effect" is a concern with observational studies. It is a phenomenon in which participants know that they are being observed, and it impacts their behaviors (Elmes, Kantowitz, and Roediger III 2006), (Johnson and Christensen 2004). For this reason, psychophysiological measures can assist with obtaining a better understanding of participants' underlying responses at the time of the observations. The benefit of behavioral measures is that researchers are able to record the actual behaviors of participants and do not have to rely on participants to report accurately their intended behaviors or preferences (Elmes, Kantowitz, and Roediger III 2006), (Bethel et al. 2007b). Video observations are often recorded later coded for visual and auditory information using two or more independent raters (Burke et al. 2004).

Another method of evaluation that is gaining more popularity in studies is the use of psychophysiological measures. The primary advantage for using psychophysiological measurements is that participants cannot consciously manipulate the activities of their autonomic nervous system (Kidd and Breazeal 2005), (Picard, Vyzas, and Healey 2001), (Rani et al. 2004), (Itoh et al. 2006), (Kulić and Croft 2006), (Liu, Rani, and Sarkar 2006). Additionally, psychophysiological measures offer a non-invasive method that can be used to determine the stress levels and reactions of participants interacting with technology (Picard, Vyzas, and Healey 2001), (Rani et al. 2004), (Itoh et al. 2006), (Kulić and Croft 2006), (Liu, Rani, and Sarkar 2006). Psychophysiological measurements can complicate the process because the results are not always straightforward and confounds can lead to misinterpretation of data. There is a tendency to attribute more meaning to results because of the tangible nature of the recordings. Information needs to be obtained from participants prior to beginning a study to help reduce these confounds (e.g., health information, state of mind, etc.). Multiple physiological signals should be used in order to find correlations in the results (Bethel et al. 2007a).

The use of task performance metrics is evolving and becoming more common in HRI studies, especially where teams are being evaluated and/or more than one person is interacting with one or more robots (Steinfeld et al. 2006), (Olsen and Goodrich 2003), (Burke et al. 2004), (Mutlu, Hodgins, and Forlizzi 2006). These metrics are designed to measure how well a person or team performs or completes a task or tasks. This is essential in some HRI studies and should be included with other methods of evaluation such as behavioral and/or self-assessments.

The study presented in this paper utilized four methods of evaluation (self-assessments, psychophysiology measurements, video-recorded observations, and a structured audio-recorded interview) so that convergent validity may be obtained to determine the effectiveness of the use of non-facial and non-verbal affective expression for naturalistic social interaction in a simulated disaster application. Multiple self-assessments were used in this study. Some of the assessments were adopted and/or modified from existing scales used in psychology, the social sciences, and other HRI researchers. The assessments were given to the participants prior to any interactions and after each robot interaction. There were five different psychophysiological signals recorded as part of this study: (1) EKG, (2) Skin Conductance Response, (3) Abdominal Respiration, (4) Thoracic Respiration, and (5) Blood Volume Pulse, using Thought Technologies ProComp5 Infiniti system (http://www.thoughttechnology.com/pro5.htm). Multiple signals were used for obtaining reliable and accurate results. Correlations will be conducted between the different signals to determine the validity of participants' responses. Video recorded observations were obtained from four different camera angles (face view-including the upper torso, overhead view, participant view, and robot view) using night vision and IR devices. A visual summary of this study can be viewed in a video format in (Bethel, Bringes, and Murphy 2009). After the interactions were complete each participant was interviewed in a structured interview format that was audio recorded. Participants were required to read and sign IRB approved informed and video/audio recording consent forms. They did have an option to deny publication of their video/audio recordings and three participants elected not have their recordings published. It is important to clearly note this in all of their files and related documents for their privacy and protection.

No one method of evaluation is sufficient in and of itself to evaluate any interaction; therefore it is important to include multiple methods of evaluation in a comprehensive study to gain a better understanding of Human-Robot Interaction. Within a single method of evaluation there should be multiple measures utilized. For example, in self-assessments, more than one assessment should be used for validity purposes; in behavioral studies, observations should be obtained from more than one angle or perspective; for psychophysiological studies more than one signal should be obtained for validity and correlation; and task performance should be measured in more than one way. This will help to ensure a comprehensive study with reliable and accurate results which can be validated.

Study Protocol

Another important phase of the planning and study design process is the development of the study protocol. The protocol involves determining exactly how the study will proceed from start to finish once a participant arrives. It is a detailed description of instructions that will be provided to the participant, what assessments will be done and in what order, what tasks the participant will perform, the timing of events, recording of information, and how and where the data and personal information will be handled and stored. This is necessary for completing the IRB paperwork required for human studies and for privacy and protection. Trial runs of experiments should be conducted until the study can be run smoothly from start to finish. This is the only way to determine where problems can and likely will occur. Planning is terrific but until an actual trial run of the experiment is conducted there is no way of knowing for sure where the problems are in the design that has been developed and there are

always unexpected problems! Systems do not always work out as expected and until several trial runs of the protocol are conducted there is no way to make sure all the bugs are worked out for the process to run smoothly. It is also important that once the actual study begins with participants that the study protocol is discussed with them as part of the instruction process as well as providing this information as part of the informed consent form participants will sign.

Study Location and Environment

A major factor to consider when planning any study is where the study will be conducted: in the field, laboratory, virtual environment, or online. For a study to be successful, the environment should realistically reflect the application domain and the situations that would likely be encountered so that participants respond in a natural manner. In some cases, it is just not practical or possible to place participants in the exact situation that is being investigated, so it is important to closely simulate that situation and/or environment. It is also important to consider lighting conditions, temperature, designing the setting to appear as close as possible to the actual setting by including props, and making sure the integrity of the site is preserved by not allowing potential participants to see it prior to the start of the study through the use of draping or other means of privacy. If psychophysiology studies are being conducted it is extremely important especially if using skin conductance as a measure that temperature is controlled in the study environment (Bethel et al. 2007a).

For the study presented in this paper, the application domain was Urban Search & Rescue, which required a confined space environment that simulated a collapsed building (See Figure 2). The study was conducted in the dark with the robots equipped to carry IR devices for recording in that environment. Participants were placed in a confined space box with a cloth cover to simulate a trapped environment. Actual rubble was brought into the lab to give the look and feel of a building collapse. The robots were all pre-programmed so that the movements would be consistent and reproducible for all participants in either a standard or an emotive mode of operation. The medical assessment path developed and traveled by the robots was based on video observations of experiments conducted by Riddle *et al.* (Riddle, Murphy, and Burke 2005), (Murphy, Riddle, and Rasmussen 2004) with emergency responders and medical personnel on how they would operate a robot to conduct a medical assessment of a trapped victim. Ideally, it would have been better to conduct the study in a real disaster or even a training exercise; however due to overall practicality and the psychophysiology measures, the study had to be conducted in a temperature controlled environment.

Robots and Other Equipment

Another consideration when designing a human study in HRI is the selection of robots for the study. The use of more than one type of robot provides a mechanism to detect if the concepts being investigated can be generalized or if they are specific to a particular robot. The results are more meaningful if they can be extended to more than one specific robot. This is often difficult to do with the cost of robots; however it does add another dimension to the study and increases the usefulness to the HRI and robotics community.

Figure 2: Confined space simulated disaster site.

Determining what equipment will be used in an HRI study impacts the success and results of this type of study. Whenever possible equipment choices should be redundant. Equipment failures are common and it is important to make sure that there are contingency plans in place in times of failure. If performing video observation or behavioral studies it is important that cameras are synchronized and that plenty of batteries and tapes/CDs/DVDs are readily available. For psychophysiological studies it is helpful to keep on hand multiple sensors in case of failure which seems to be common due to the sensitive nature of the equipment itself. This can make the difference between a productive, successful, and organized study and one that produces stress, delays, and sometimes failure.

Conducting the Study

In most cases, to run a successful human study requires assistance in addition to the principal investigator. This is especially true when running a large-scale, complex study with a significant sample size and multiple methods of evaulation. Finding research assistants can be a challenge for some researchers, especially when economic times are tough and there may not be funding available to pay for assistants. One option available is to contact the Honors College or Program if your university or institution has this type of program. These students typically desire research experience and often are willing to volunteer their time for the experience and knowledge they can gain. Depending on the study, often students can be easily trained to assist and do not necessarily need to be in the field. Psychology and pre-medical students often need a certain amount of volunteer hours and assisting in a research study can fulfill these requirements. It is important to make sure the volunteers understand the need for reliability and attention to detail. It is recommended that whenever possible schedule an additional person to be available in case of emergencies or when plans do not move forward as expected.

Recruiting Participants

Recruiting participants is a challenge that most human studies face in any field including HRI. That may be a significant reason why most of the studies conducted to date did not have large sample sizes. There are several methods of recruitment available, and they should all be implemented for a successful study. Flyers are a good method of recruitment on campus with the added bonus of some type of incentive to participate (e.g., door prizes, payment for participation, extra credit in courses). In some cases, the psychology department may have research participation requirements and a system for advertising research studies on campus. This is a terrific method of recruitment if it is available. Using word of mouth to friends, family, and associates to attract non-traditional participants is also beneficial.

Another challenge is having participants show up once they are scheduled. It is important to maintain contact with participants to remind them of their time to participate. It is encouraged that researchers schedule appointment times for their convenience and for the participants. It is more likely a participant will show up if they have a specific time and additionally it is recommended to allow appropriate time between participants to account for time delays in the study or in case the participant is running late. Even with all this planning things happen and participants do not show up but that time can be used for other duties.

Failures and Contingencies

Even with careful planning, failures and problems seem to always occur. It is imperative that planning is done for failures of all types. Robots can fail, cameras can fail, computers and sensors can fail; therefore it is important whenever possible to have redundancy in all necessary equipment. It needs to be available immediately so that delays do not occur in the study. It is also recommended that there be redundancy in personnel as well. There can also be a call list developed for participants that might be available on short notice to fill a timeslot where a participant does not show up. It is common to expect approximately 20% of scheduled participants to not appear for their scheduled appointment. When calculating the number of participants required for a study this number should be increased to take into account the likelihood of participants that will not show up for their appointment and also to account for any possible data failures or problems.

In the study presented in this paper, the "no show" percentage was much lower at around 8% and equipment failures did occur. There were focus problems with two of the video cameras which impacted some of the video data; however as soon as the problem was noticed the cameras were swapped out and the problem quickly resolved. The EKG sensors had multiple failures and backup sensors were on-hand to correct the problem fairly quickly. The primary failure that ended the study and resulted in the cancellation of 18 participants was the failure of the one robot for which there was no redundancy; however the goal number of 128 participants was attained.

Additional Recommendations

Performing a large-scale, complex human study in HRI has many pitfalls and rewards. Even with the most careful planning and study design it becomes apparent through the course of the study that changes could be made to improve the study. An example from the study that was presented throughout this paper was the design and development of the simulated disaster site. It was high fidelity and based on real-world knowledge; however it would have been more realistic had the confined space box been more confining. The box was designed based human factors standards for designing spaces to accommodate 95% of the population. In the case of this study most of the population of participants were much smaller than average and the space was truly confining to a small portion of the participants. To increase the feeling of confinement a blanket or heavy cover should be utilized in the future. Additionally, a soundtrack playing in the background with sounds from an actual disaster or a training exercise would have improved the fidelity of the site and the experiences of the participants.

Making sure there are contingencies for equipment cannot be stressed enough. The study experienced a one week delay due to the failure of an EKG sensor which was essential to the psychophysiology portion of the study. Planning ahead and having extra sensors could have prevented delays and the loss of participants that could not be rescheduled. Following that experience extra sensors were ordered and kept on hand and they were needed. Video cameras had autofocus problems and they were not observed until the video data was being offloaded. Also one video camera was moved between the two different robots and accidentally the zoom was activated making some of the robot view video data not usable. It is important to always check and double check equipment settings and constantly verify all equipment is working properly so that no data is lost or determined to be unusable.

Synchronizing multiple cameras can be a challenge. In the case of the study presented the interactions were all conducted in the dark. Turning video cameras on before the lights were turned off and turning the lights back on before shutting off the cameras made a good synchronizing point for multiple cameras. Another technique would be the use of a sound that all cameras would detect through built-in microphones.

It would be recommended to conduct pilot studies of all the assessments to make sure that they are understandable and testing exactly what was expected. In the study conducted, some of the questions were confusing to the participants and will not be considered as part of the data analysis. It is important to note the questions that participants find confusing and/or they request further explanation. In the case of one assessment the valence and arousal questions were easily interpreted; however the questions relating to the dominance dimension were often misunderstood. That dimension will not be included as part of the data analyses. The questions associated with the dominance dimension of the assessment will need to be reworded and then validated; however the valence and arousal portions have been validated for future HRI studies and will be made available.

Conclusions

Planning and designing a human study for HRI can be challenging; however with careful planning many of these challenges can be overcome. There are two main improvements that need to be made in human studies conducted in HRI and those are (1) having large sample sizes to appropriately represent the population being studied and so that small to medium effects can be obtained with statistically significant results; and (2) the use of multiple methods of evaluation to establish reliable and accurate results that will have convergent validity. We have the following recommendations for planning, designing, and conducting human studies in HRI:

1. Determine the most appropriate type of study for the hypotheses being investigated using either a within-subjects, between-subjects, or mixed-model design.
2. Perform an a priori power analysis to estimate the appropriate number of participants required for the study in order to have a better opportunity to obtain statistically significant results that are valid, reliable,and accurate. This can be accomplished through power analysis tables or freely available software. An a priori power analysis is based on the number of groups in the study, the effect size, the alpha level, and the desired statistical power. It is recommended to add a few more participants to the estimated number to account for problems with data, participant cancellations that cannot be rescheduled, and participants that do not show up.
3. Determine the best methods of evaluation for the hypotheses being investigated, but it is recommended that at least two or more methods should be utilized in order to obtain convergent validity in the study. The results from a study with multiple methods of evaluation are viewed as more reliable and accurate. Additionally, by incorporating multiple methods of evaluation it will overcome the inherit problems found with every method of evaluation. No one method is sufficient for accurately measuring participants' responses.
4. Develop a written study protocol of all instructions, assessments, participant tasks, timing of events, coordination of data collection, and activities. This will used when preparing IRB paperwork, creating instructions for participants, and preparing informed consent documents.
5. Perform multiple test runs of the planned study protocol until all glitches and problems have been discovered and resolved and there is a smooth running system in place.
6. Design an environment or study space that closely reflects the real-world that is being tested for more natural participant responses. If a real-world environment is not possible for testing, then make sure the test environment is as high fidelity as possible.
7. Whenever possible performing the study with more than one type of robot will help with generalizing results across different robot types versus results that are specific to a particular robot.
8. Make sure that there is redundancy in all equipment that is required for the study because failures are common.
9. A good source of recruiting quality volunteer research assistants is from an Honors College or Program if available at the university or institution. Additionally, pre-medical and psychology students often have volunteer hours requirements and will volunteer their time to fulfill these requirements.
10. Recruiting participants to reach the estimated number required by the power analysis requires multiple methods of contact such as flyers posted across campus; word of mouth to friends, family, and associates; offering incentives such as door prizes, pay for participation, and extra credit in courses for participation; signing up for research study participant pools through the psychology department on campus if offered.
11. Always prepare for and expect equipment failures, participants and/or research assistants not showing up at their designated times, or just about anything else that might unexpectedly go wrong. Having backup plans in place such as redundant equipment, double schedule volunteer assistants or have an on-call list, and develop an on-call list for participants that are available on short notice.
12. Always allow time for study delays, participants running late, and/or equipment failures that may cause cancellation of participants and delay of the overall study.

Conducting human studies can be challenging and also very rewarding. Careful planning and design can make the experience more positive and successful. Following the above recommendations should improve the chances of having a successful study with accurate and reliable statistically significant results. Through the use of appropriate sample sizes and multiple methods of evaluation, convergent validity should be obtained.

Acknowledgments

This work is supported in part under a National Science Foundation Graduate Research Fellowship – Award Number DGE – 0135733, ARL Number W911NF-06-2-0041, and the Microsoft Rescue Buddy Research Project. Special thanks go to Kristen Salomon, Jennifer Burke, John Ferron, and my lab assistants – Brian Day, Christine Bringes, Megan Brunner, Andrea Vera, Leslie Salas, Stephanie Smith, Kimberlee Fraser, Cherisse Braithwaite, and Caitlin Howell for their contributions and assistance.

References

Bethel, C. L.; Salomon, K.; Burke, J. L.; and Murphy, R. R. 2007a. Psychophysiological experimental design for use in human-robot interaction studies. In *The 2007 International Symposium on Collaborative Technologies and Systems (CTS 2007)*. Orlando, FL: IEEE.

Bethel, C. L.; Salomon, K.; Murphy, R. R.; and Burke, J. L. 2007b. Survey of psychophysiology measurements applied to human-robot interaction. In *16th IEEE International Symposium on Robot & Human Interactive Communication*.

Bethel, C. L.; Bringes, C.; and Murphy, R. R. 2009. Non-facial and non-verbal affective expression in appearance-constrained robots for use in victim management: Robots to the rescue! In *4th ACM/IEEE International Conference on Human-Robot Interaction (HRI2009)*.

Bethel, C. L.; Salomon, K.; and Murphy, R. R. 2009. Preliminary results: Humans find emotive non-anthropomorphic robots more calming. In *4th ACM/IEEE International Conference on Human-Robot Interaction (HRI2009)*.

Burke, J. L.; Murphy, R. R.; Riddle, D. R.; and Fincannon, T. 2004. Task performance metrics in human-robot interaction: Taking a systems approach. In *Performance Metrics for Intelligent Systems*.

Dautenhahn, K.; Walters, M.; Woods, S.; Koay, K. L.; Nehaniv, C. L.; Sisbot, A.; Alami, R.; and Siméon, T. 2006. How may i serve you?: A robot companion approaching a seated person in a helping context. In *1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI2006)*, 172–179. Salt Lake City, UT: ACM Press, New York, NY, USA.

Elmes, D. G.; Kantowitz, B. H.; and Roediger III, H. L. 2006. *Research Methods in Psychology*. Belmont, CA: Thomson-Wadsworth, 8th edition.

Itoh, K.; Miwa, H.; Nukariya, Y.; Zecca, M.; Takanobu, H.; Roccella, S.; Carrozza, M. C.; Dario, P.; and Atsuo, T. 2006. Development of a bioinstrumentation system in the interaction between a human and a robot. In *International Conference of Intelligent Robots and Systems*, 2620–2625.

Johnson, B., and Christensen, L. 2004. *Educational Research Quantitative, Qualitative, and Mixed Approaches*. Boston, MA: Pearson Education, Inc., 2nd edition.

Kidd, C. D., and Breazeal, C. 2005. Human-robot interaction experiments: Lessons learned. In *Proceeding of AISB'05 Symposium Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction*, 141–142.

Kulić, D., and Croft, E. 2006. Physiological and subjective responses to articulated robot motion. *Robotica* Forthcoming:15.

Liu, C.; Rani, P.; and Sarkar, N. 2006. Affective state recognition and adaptation in human-robot interaction: A design approach. In *International Conference on Intelligent Robots and Systems (IROS 2006)*, 3099–3106.

Moshkina, L., and Arkin, R. C. 2005. Human perspective on affective robotic behavior: A longitudinal study. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, 2443–2450.

Murphy, R. R.; Riddle, D.; and Rasmussen, E. 2004. Robot-assisted medical reachback: a survey of how medical personnel expect to interact with rescue robots. In *13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2004)*, 301–306.

Mutlu, B.; Osman, S.; Forlizzi, J.; Hodgins, J. K.; and Kiesler, S. 2006. Task structure and user attributes as elements of human-robot interaction design. In *15th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2006)*. University of Hertfordshire, Hatfield, UK: IEEE.

Mutlu, B.; Hodgins, J. K.; and Forlizzi, J. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *2006 IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS'06)*. Genova, Italy: IEEE.

Olsen, D. R., and Goodrich, M. A. 2003. Metrics for evaluating human-robot interactions. In *Performance Metrics for Intelligent Systems Workshop*.

Picard, R. W.; Vyzas, E.; and Healey, J. 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(10):1175–1191.

Rani, P.; Sarkar, N.; Smith, C. A.; and Kirby, L. D. 2004. Anxiety detecting robotic system - towards implicit human-robot collaboration. *Robotica* 22(1):85–95.

Riddle, D. R.; Murphy, R. R.; and Burke, J. L. 2005. Robot-assisted medical reachback: using shared visual information. In *IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2005)*, 635–642. Nashville, TN: IEEE.

Steinfeld, A.; Fong, T.; Kaber, D.; Lewis, M.; Scholtz, J.; Schultz, A.; and Goodrich, M. 2006. Common metrics for human-robot interaction. In *1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. Salt Lake City, Utah, USA: ACM Press.

Stevens, J. P. 1999. *Intermediate Statistics A Modern Approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 2nd edition.

Watson, D.; Clark, L. A.; and Tellegen, A. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology* 54(6):1063–1070.