Int J Soc Robot (2010) 2: 347–359
DOI 10.1007/s12369-010-0064-9

Review of Human Studies Methods in HRI and Recommendations

Cindy L. Bethel · Robin R. Murphy

Accepted: 21 June 2010 / Published online: 14 July 2010

Abstract This article provides an overview on planning, designing, and executing human studies for Human-Robot Interaction (HRI) that leads to ten recommendations for experimental design and study execution. Two improvements are described, using insights from the psychology and social science disciplines. First is to use large sample sizes to better represent the populations being investigated to have a higher probability of obtaining statistically significant results. Second is the application of three or more methods of evaluation to have reliable and accurate results, and convergent validity. Five primary methods of evaluation exist: self-assessments, behavioral observations, psychophysiological measures, interviews, and task performance metrics. The article describes specific tools and procedures for operationalizing these improvements, as well as suggestions for recruiting participants. A recent large-scale, complex, controlled human study in HRI using 128 participants and four methods of evaluation is presented to illustrate planning, design, and execution choices.

This material is based upon work supported by the National Science Foundation under Grant # 0937060 to the Computing Research Association for the CIFellows Project, a National Science Foundation Graduate Research Fellowship Award Number DGE-0135733, ARL Number W911NF-06-2-0041, IEEE Robotics and Automation Society Graduate Fellowship, and a Microsoft HRI grant.

C.L. Bethel (✉)
Department of Computer Science, Yale University, New Haven, CT 06511, USA
e-mail: cindy.bethel@yale.edu

R.R. Murphy
Center for Robot-Assisted Search and Rescue, Texas A & M University, College Station, TX 77843, USA
e-mail: murphy@cse.tamu.edu

Keywords Human-robot interaction · Experimental design · Human studies · Evaluation methods

1 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 [1]. This brings credibility and validity to scientific research that involves humans as subjects, as recognized in the psychology and social science fields. Two primary issues observed in HRI studies are the lack of significant-sized participant pools that closely represent the populations being studied, and the lack of three or more methods of assessment used to obtain convergent validity in HRI studies [12, 16, 17].

The focus until recently in HRI was on the development of specific robotic systems and applications while neglecting methods of evaluation and metrics. Some methods of evaluation have been adopted and/or modified from such fields as human-computer interaction, psychology, and social sciences [17]; 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, it will be important to develop accurate methods to assess how humans respond to robots, how they feel about their interactions with robots, and how they interpret the actions of robots. [1, 2, 6].

There are five primary methods of evaluation used for human studies in HRI: (1) self-assessments, (2) interviews, (3) behavioral measures, (4) psychophysiology measures, and (5) task performance metrics [2, 4, 6, 12, 17]. From the review of HRI literature, it appears the most common methods utilized in HRI studies are self-assessment and behavioral measures. There is limited research in the use of

psychophysiological measures, interviews, and task performance metrics. Each method has advantages and disadvantages; however disadvantages can be overcome by using more than one method of evaluation [2, 6, 17].

The design of quality research studies for use in HRI applications with results that are verifiable, reliable, and reproducible is a major challenge [1]. 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. describe the need for the development of common metrics as an open research issue in HRI [32]. They discuss an approach of developing common metrics for HRI; however this approach is oriented toward an engineering perspective and does not completely address the social interaction perspective. Both the engineering and social interaction perspectives require further investigation to develop metrics and methods of evaluation.

The article begins with a discussion of some related work on experimental designs and methods used in HRI. There is a brief summary of terminology and related information presented in Sect. 3. Next, the article details the process of planning and designing a human study in HRI in Sect. 4. This section covers how to determine the type of study to use, number of participants, and methods and measures of assessment (advantages and disadvantages). Additionally, there is a discussion of how to design a high-fidelity study site, select robots and other equipment, find assistants to conduct the study, recruit the required number of participants, develop contingency plans to deal with failures, and prepare Institutional Review Board (IRB) documents. In Sect. 5, illustrative examples are drawn from a recent large-scale, complex, controlled human study in HRI using 128 participants and 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. Section 6 summarizes three categories of recommendations for designing and executing large-scale, complex, controlled human studies in HRI using appropriate samples sizes with three or more methods of evaluation, and discusses how improving experimental design can benefit the HRI community.

2 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 is that 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 observed in HRI studies has been self-assessments. In general, most of the studies in HRI include some form of questionnaire; however in some cases, the researchers 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. in which they utilized self-assessments, unstructured interviews, and video observations from one camera angle [10]. 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 [21] in which they used self-assessment measures including a measure commonly used in psychology studies called the Positive and Negative Affect Schedule [34]. Video observation and coding were performed, though results were not presented. Their study included 20 participants in a controlled, laboratory setting. The study results 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 self-assessments [23]. The sample size for this study was 20 participants and the results were mixed. One hypothesis showed statistically significant results; however 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 larger studies in HRI to use psychophysiology measurements along with self-assessments was conducted by Kulić and Croft with a sample size of 36 participants [18]. Multiple psychophysiological signals were measured which is highly recommended for this type of study for reliability and validity in the results [2, 4, 15, 17, 20, 26, 28]. As a result of having a larger participant pool than previous studies, they found statistically significant results and were able to discover the best psychophysiology measures to determine valence and arousal responses from participants. The results may have been even more prominent with a larger sample size. Some of the psychophysiological signals that were concluded not to be good indicators of valence and arousal actually may have had a smaller effect size and may have shown different results with a larger sample size.

Mutlu et al. 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 [24]. The study relied

heavily on the use of self-assessments developed from their previous studies and adapted from psychology. 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 their conclusions.

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. It is essential to use three or more methods of evaluation to establish 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 utilizing study design techniques, which are the state of practice in the psychology and social science communities.

3 Terminology and Related Information for Conducting Human Studies in HRI

This section contains terminology and background information needed for planning, designing, and conducting human studies in HRI. The information presented will provide a general understanding of research methods and statistical terminology to form a very basic foundation. For a more in-depth understanding it is recommended that readers refer to research methods and/or statistical textbooks (e.g., [12, 14, 19, 30, 31]).

Alpha level: the probability of having a Type I error, which occurs when the null hypothesis is rejected and it is true.

Between-subjects design: participants are placed in different groups with each group experiencing different experimental conditions.

Confound: any extraneous variable that covaries with the independent variable and might provide another explanation for findings discovered in the study.

Contingency plans: plans that are developed for cases of failures or unexpected events that were not part of the original design or plan (robot failures, equipment problems, participants not showing up, etc.).

Control condition: one of the groups in an experimental design that does not receive the experimental condition being evaluated in a study.

Counterbalance: a procedure used in within-subjects designs that changes the order variables are presented to control for sequence effects.

Dependent variable: the behavior that is evaluated as the outcome of an experiment.

Effect size: the amount of variance in the dependent variable that can be explained by the independent variable. The amount of influence one variable can have on another variable. (Additional information and an example follow this terminology list)

Experimental condition: the group(s) in an experimental design that receive the experimental condition being evaluated in a study.

Independent variable: the variable(s) that are manipulated by the experimenter and is of interest.

Interaction: in a mixed-model factorial design, an interaction occurs when the effect of one independent variable that is manipulated depends on the level of a different independent variable.

Main effect: in a mixed-model factorial design, it is whether or not there is a statistically significant difference between different levels of independent variables.

Mixed-model factorial design: this type of design includes both between-subjects and within-subjects design components.

Objectivity: this occurs when observations can be verified by multiple observers with a high level of inter-rater reliability.

Power: the probability that the null hypothesis will be rejected when it is false. It is impacted by alpha level, effect size, and sample size.

Reliability: the consistency in obtaining the same results from the same test, instrument, or procedure.

Sample: a portion or subset of a population.

Type I error: occurs when the null hypothesis is rejected when it is true.

Type II error: failure to reject the null hypothesis when it is false. It occurs when there is failure to find a statistically significant effect when it does exist.

Validity: a method of evaluation (test, instrument, procedure) that measures what it claims to measure.

Within-subjects design: each participant is exposed to all levels of the independent variable(s).

Effect size information: The formula for calculating Cohen's $\hat{f}$ effect is [9]:

$$\hat{f} = \sqrt{df\left(\frac{F}{N}\right)}$$

where, df = the degree of freedom for the numerator (number of groups—1), F = the statistically significant result of an F-test, N = the total sample size.

The scale used to interpret Cohen's $\hat{f}$ effect is:

- 0.00–0.09 negligible effect
- 0.10–0.24 small effect
- 0.25–0.39 medium effect
- 0.40 + large effect

The following is an example of an effect size calculation using a significant main effect for arousal from the exemplar study:

The following F-test result is used to calculate the effect size: $F(1, 127) = 12.05$.

$$\hat{f} = \sqrt{1\left(\frac{12.05}{127}\right)} = 0.31$$

Based on Cohen's scale, this is a medium effect.

4 Planning and Experimental Design

A successful human study in HRI requires careful planning and design. There are many factors that need to be considered (see Fig. 1). When planning and designing a human study in HRI the following questions should be considered:

- 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, interviews, psychophysiological measures, task performance, etc.)?
- What type of environment and space is required to conduct the study (field, laboratory, virtual, etc.)?
- What type of robots will be used?
- How many different types of robots will be used?
- What type of equipment will be needed (computers, measurement equipment, recording devices, etc.)?
- How will contingencies and failures be handled?
- What types of tasks will the participants perform or observe (Study Protocol)?
- How will participants be recruited for the study?
- What type of assistance is needed to conduct the study?

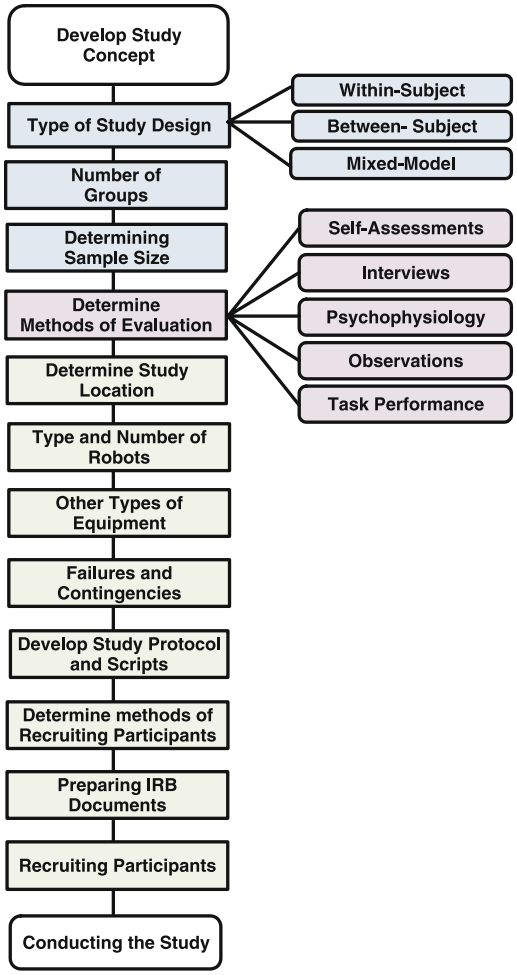

Fig. 1 The chronology of items required for planning, designing, and executing human studies in HRI

4.1 Type of Study and Number of Groups

The first step in designing a human study in HRI is to determine the research question(s) and hypotheses. From the research questions the researcher can determine how many groups are needed and whether the study design should be a within-subjects, between-subjects, or a mixed-model factorial approach. The most comprehensive approach is the within-subjects design in which every participant experiences all of the experimental conditions being investigated. The within-subjects design requires less time to conduct the study because all experimental conditions can be performed in one session and there is no waiting to run different participants through different experimental conditions as observed in between-subjects designs. Within-subjects designs require fewer participants, the variables remain constant across the different experimental conditions, it increases statistical power, and reduces error variance. However, within-subject designs are prone to confounds from demand characteristics, in which participants make inferences about the purpose of the study in an effort to cooperate. Participants may also experience a concept known as habituation or practice effect in which participants' responses are reduced due to repetitive presentation of the same tasks or the robot performs in a similar manner reducing the novelty effect of the robot. Participants are likely to be impacted by side effects of different events that can occur during a study (e.g., robots breaking or behaving in ways that were not anticipated) [30].

In a between-subjects design, participants experience only one of the experimental conditions. The number of experimental groups depends on the number of experimental conditions being investigated [19]. A common reason for using a between-subject design would be the participants themselves may dramatically differ such as experiments that may evaluate human-robot interactions for typically developing children versus children diagnosed with autism. In these situations, the participants can be classified in only one of the groups. The results between the groups are then compared. Between-subject designs are typically cleaner be-

cause participants are exposed to only one experimental condition and typically do not experience practice effects or learn from other task conditions. The time to run the participant through one condition is less than in a within-subjects design where the participants experience all experimental conditions. The between-subjects design reduces confounds such as fatigue and frustration from repeated interactions with the robot. A limitation of the between-subject design is that results are compared between the groups which can result in substantial impacts from individual differences between the participants of the different groups. Therefore, it makes it more difficult to detect differences in responses to the robots and Type II errors can be more prevalent [19].

A mixed-model factorial design uses both between-subjects and within-subjects designs. This can be useful when there are one or more independent variables being investigated with a between-subjects design and the other variables are explored through a within-subjects approach in the same study. In this type of design, the variables being investigated have different levels, such as there may be two types of robots used in the interactions. It allows the investigator to explore if there are significant effects from each individual variable, but this type of design allows for the exploration of the interaction effects between two or more independent variables [19]. The limitations previously mentioned for within-subjects and between-subjects designs can apply in the mixed-model factorial design as well and must be considered. Section 5.1 provides an example of a mixed-model factorial design.

4.2 Determining Sample Size

Determining the appropriate sample size appears to be a challenge in human studies in HRI. An *a priori power analysis* is a statistical calculation that can be performed to determine the appropriate number of participants needed to obtain accurate and reliable results based on the number of groups in the study, the alpha level (typically $\alpha = .05$), the expected or calculated effect size, and a certain level of statistical power (commonly 80%). There are power analysis tables in the appendices of most statistical textbooks (e.g., refer to Appendix C in [33]) that will provide group size values. Additionally, there is software available online that will assist with this type of calculation (e.g., G*Power3; http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/). Section 5.2 presents examples for calculating sample size using statistical tables and the G*Power3 software.

4.3 Methods of Evaluation

There are five primary methods of evaluation used in human studies in HRI: (1) self-assessment, (2) observational or behavioral measures, (3) psychophysiology measurements, (4) interviews, and (5) task performance metrics. Each of these methods has advantages and disadvantages; however most problems can be overcome with the use of three or more appropriate methods of evaluation. For results to be valid and reliable, use at least three different credible forms of evaluation in a process known as triangulation to obtain convergent validity [2, 4, 6, 17, 19, 27]. Due to the fact that no one method of evaluation is without problems, researchers should not rely solely on the results of one source of evaluation. The use of two sources of evaluation may result in conflicting results; however using three or more methods of evaluation it is expected that the results for two of the three or more methods should support each other adding validity and reliability to the findings. If three or more methods of evaluation are in conflict, reconsider the research question(s) and possibly structure in a different way.

It is important to use credible measures that are appropriate for the study and participants being investigated. For example, if the researcher is evaluating anxiety and stress levels of participants interacting with a robot, then they may want to use validated self-assessments, video observations, and psychophysiology measures such as respiration rate, electrocardiography (EKG), and skin conductance response. These measures would be appropriate for the study and these are credible measures for assessing these types of responses. The use of multiple methods of evaluation that are not credible or appropriate will result in data that is not meaningful to answer the hypothesized research questions.

Self assessments are among the most commonly used methods of evaluation in HRI studies; however obtaining validated assessments designed for HRI studies can be a challenge. Self-assessment measures include paper or computer-based psychometric scales, questionnaires, or surveys. With this 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 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 corroborate the information provided by participants immediately and directly [12]. 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 to self-assessments and other measures could be influenced by participants' mood and state of mind on the day of the study [12, 16]. For these reasons, it is important to perform additional types of measurements such as behavioral, interviews, task performance

and/or psychophysiological measures to add another dimension of understanding of participants' responses and performance in HRI studies [2, 6].

Behavioral measures are the second most common method of evaluation in HRI studies, and sometimes are included along with psychophysiological evaluations and participants' self-assessment responses for obtaining convergent validity. Johnson and Christensen define observation as "the watching of behavioral patterns of people in certain situations to obtain information about the phenomenon of interest" [16]. The "Hawthorne effect" is a concern with observational as well as self-assessment studies. It is a phenomenon in which participants know that they are being observed, and it impacts their behaviors [12, 16]. For this reason, psychophysiological measures can assist with obtaining an 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 need to rely on participants to report accurately their intended behaviors or preferences [2, 6, 12]. Video observations of human-robot interactions are often recorded and later coded for visual and/or auditory information using two or more independent raters [8]. Interpreting audio and video data does require training to provide valid, accurate, and reliable results. There are multiple approaches to interpreting this type of data, which is beyond the scope of this article (refer to [14, 19, 27, 30, 31]).

Psychophysiology measures are gaining popularity in HRI studies. The primary advantage for using psychophysiological measurements is that participants cannot consciously manipulate the activities of their autonomic nervous system [15, 17, 18, 20, 26, 28]. Also, psychophysiological measures offer a minimally-invasive method that are used to determine the stress levels and responses of participants interacting with technology [15, 18, 20, 26, 28]. Psychophysiological measurements can complicate the process because the results may not be straightforward and confounds can lead to misinterpretation of data. There is a tendency to attribute more meaning to results due to the tangible nature of the signals. Information needs to be obtained from participants prior to beginning a study to reduce these confounds (e.g., health information, state of mind, etc.). Multiple physiological signals should be used to obtain correlations in the results [2, 4].

Interviews which are closely related to self-assessments are another method of evaluation. Interviews can be structured in which the researcher develops a series of questions that can be close-ended or open-ended; however the questions are given in the same order to every participant in the study. The interview can be audio and/or video recorded for evaluation at a later time. Unstructured interviews are used less frequently in research studies. In unstructured interviews, the questions are changed and developed based on participants responses to previously presented questions. It is an adaptive process and more difficult to have consistency and to evaluate in research studies. Interviews often provide additional information that may not be gathered through self-assessments; however there are numerous issues that may arise from using interviews. Response style of participants can influence responses to interview questions. There are three types of response styles, (1) response acquiescence—participants answer in the affirmative or yea-saying, (2) response deviation—participants answer in the negative or nay-saying, and (3) social desirability—participants provide what they perceive as socially acceptable responses. It can be a challenge to obtain responses that are reflective of participants' true behaviors [12]. Another issue related to interviews is that participants that volunteer for the research study may not answer interview questions in a manner consistent with those participants that are not volunteers. Some of these challenges can be overcome by using other methods of evaluation to obtain convergent validity among different measures.

Task performance metrics are becoming more common in HRI studies, especially in studies where teams are evaluated and/or more than one person is interacting with one or more robots [8, 11, 23, 25, 32]. 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.

No single method of measurement is sufficient to evaluate any interaction; therefore it is important to include three or more 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 credible assessment should be used for validity purposes. In behavioral studies, obtain observations from more than one angle or perspective. For psychophysiological studies use more than one signal to obtained validity and correlation. Measure task performance in more than one way. This ensures a comprehensive study with reliable and accurate results that can be validated. Compare the findings from three or more of these measures to determine if there is meaningful support in at least two or more of the evaluations. Normalize the means for each type of measurement to a meaningful, common scale and then perform a separate analysis on this data. Conduct a correlation analysis to determine if there are positive correlations in the results. Essentially it is important to interpret the meaning of the results discovered and determine if there are commonalities in the results of two or more of the methods of evaluation.

4.4 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 successful study, the environment should reflect realistically 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 important to consider lighting conditions, temperature, and design the environment to appear as close as possible to the actual setting by including props and/or sound effects. Use draping or other means of privacy to ensure the integrity of the site is preserved prior to the start of the study. In psychophysiology studies, if skin conductance measures are used, it is extremely important that the temperature is controlled in the study environment [2, 4].

4.5 Type and Number of Robots

Another consideration when designing a human study in HRI is the selection of robots for the study. The selection of a robot needs to be congruous with the application and domain being investigated. It is important to select robots that have been used or would be expected in the type of tasks/applications being examined in the research 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.

4.6 Other Equipment

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.

When performing video observation or behavioral studies, the first step is to determine the number of different perspectives needed for the study. It is important to obtain multiple viewing angles because each perspective may contain unique information and gives a more comprehensive record of the events that occur in a study. It is important that cameras are synchronized and extra batteries and tapes/CDs/DVDs are readily available. There should be at least one or two extra cameras available in case of equipment failure. It is also advisable not to reuse tapes if at all possible because it can impact the integrity of the recordings. To preserve data and prevent mishaps, it is important to off-load recordings quickly to a more stable media.

For psychophysiological studies, it is necessary to determine if the equipment needs to be connected to a stationary computer or if the participant will be mobile. There are limited options available for ambulatory psychophysiology equipment. It is recommended to keep on hand multiple sensors in case of failure, which seems to be common due to the sensitive nature of this type of equipment. This can make the difference between a productive, successful, and organized study and one that produces stress, delays, and sometimes failure.

4.7 Failures and Contingencies

Even with careful planning, failures and problems are likely to occur. It is imperative to plan for as many potential failures as can be anticipated. 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 to prevent delays in the study. It is also recommended that there be redundancy in personnel as well. Develop a call list for participants and essential personnel who might be available on short notice to fill a timeslot where a participant or assistant does not arrive as scheduled. It is common to expect approximately 20% of scheduled participants not to appear for their appointment. When calculating the number of participants required for a study this number should be increased to take into account the likelihood that some participants will miss their appointment and to account for any possible data failures.

4.8 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, how the data and personal information will be handled, and where this information will be stored for security purposes. This is necessary for completing the IRB paperwork required for human studies and for maintaining privacy.

Trial runs of experiments should be conducted until the study can be executed smoothly from start to finish. This is the only way to determine where problems can and likely

will occur. Systems and study designs do not always execute as expected and until several trial runs of the protocol are performed there is no way to ensure that all the problems are resolved for the process to run smoothly. It is important that once the study begins with participants that the study protocol is discussed with each participant as part of the instruction process as well as providing this information as part of the informed consent form participants will sign.

4.9 Methods of 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 in HRI do not have large sample sizes. It is important to recruit participants that will appropriately represent the population being studied. The novelty effect of robots in some cases is not enough to entice participants to be involved in a study. 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. Establish relationships with management of other participant pools or databases. These are excellent sources of recruitment on college campuses; however limiting participation to these sources will often bias your participant pool. In most cases the population of interest may not include college educated participants, and the results of studies using just these sources of participants will not generalize. Therefore, it is important to explore other methods of recruitment such as word of mouth to family, friends, and acquaintances. It is also possible to contact other resources for permission to solicit participants, such as a local mall for testing the general public, and kindergarten through 12th grade educational institutions for recruiting children (the use of children requires informed consent from the parents and informed assent from the children to participate). These methods are more involved, but can serve as rich sources of recruitment.

4.10 Preparing Institutional Review Board Documents

The next step in planning, designing, and executing a human study is the preparation of the Institutional Review Board documentation (this is applicable to all human studies conducted in the United States, there may be similar requirements in other countries). The Institutional Review Board (IRB) is a committee at each university. Its members are from diverse backgrounds and they review proposals for research conducted with human subjects. The IRB was established to review the procedures of proposed research projects, to determine if there are any known risks or benefits to participants in the study, methods of recruiting participants, and how the participants' confidentiality will be maintained. A part of the IRB application is the informed consent, or in the case of studies with children, parental informed consent/child assent documents. This document is provided and explained to each participant prior to being involved in any research study and includes the study protocol, informed consent/assent, permissions to audio and/or video record the study, any risks or benefits to the participants, and how confidentiality of the data will be maintained [30]. Also included in the informed consent is a statement that participants can terminate participation in the study at any point without any penalty and they will still receive any incentives provided in the study. It is important to include this type of language in any informed consent form provided to participants for ethical reasons. The IRB reviews these documents and provides approval to proceed with the study, requests revisions to the study documents, or can deny approval of the study. Typically there is training required by the personnel involved with the study but the requirements can vary by institution; therefore it is important to investigate all the requirements of the institution(s) that will be involved in the study.

4.11 Recruiting Participants

Once IRB approval is received, the next step in the study process is the actual recruitment of participants and ensuring they follow through once recruited. A significant challenge in most human studies is participants attendance once they are scheduled. Schedule researchers, assistants, and participants for mutually convenient times. It is important to remind participants of their appointment time. It is more likely a participant will show up if they have a specific time and also it is recommended to allow adequate time between participants to account for time delays in the study or in case a participant is running late. Even with planning, problems occur and participants do not show up, but this time can be used to process data.

4.12 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 three or more methods of evaluation. 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 research assistants. One option available is to contact the Honors College or Program if the university or institution has this type

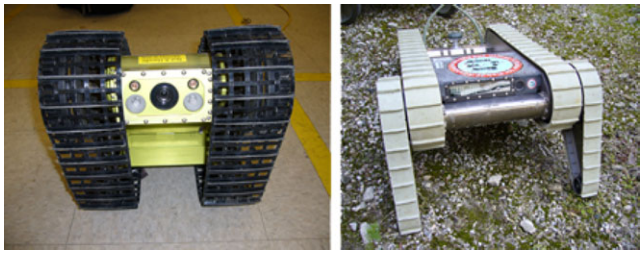

Fig. 2 The Robots: Inuktun Extreme-VGTV (left) and iRobot Packbot Scout (right)

of program. These students typically desire research experience and often are willing to volunteer their time for the experience and knowledge they may gain. Depending on the study, often students can easily be trained to assist and do not necessarily need to be in the field of study. 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 ensure the volunteers understand the need for reliability and attention to detail. Whenever possible, schedule an additional person to be available in case of emergencies or when plans do not proceed as expected. Volunteer research assistants each will typically provide between five and ten hours per week; therefore it is necessary to consider their availability when designing the study and the timeline. It is also advisable to schedule assistants for data processing as well as assisting with conducting the actual experiments. These recommendations may not be applicable to all institutions or studies.

5 An Exemplar Study

This section presents examples from a recent HRI study on determining study design, sample size, methods of evaluation, study location, and how failures and contingencies should be handled. This was a large-scale, complex, controlled study involving 128 participants responding to two different search and rescue robots (Inuktun Extreme-VGTV and iRobot Packbot Scout) operated in either the standard or emotive modes. The experiment was conducted in the dark in a high-fidelity simulated disaster site, using four different methods of evaluation.

5.1 Type of Study and Number of Groups

This study was a mixed-model factorial design, in which the between-subjects factor was the robot operating mode (standard versus emotive) and the within-subjects factor was robot, the Inuktun Extreme-VGTV and the iRobot Packbot Scout (See Fig. 2). This design was selected because there were four conditions and that was too many for a within-subjects design. We did not want to expose participants to both the emotive and standard operating modes in the same study or they would likely determine the purpose of the experiment. Participants were randomly assigned to one of two groups (standard operated or emotive operated). Every participant experienced both robots within their assigned group. The order in which the robots appeared was counterbalanced (e.g., Inuktun first or Packbot first), and operating mode assignments were balanced for age and gender.

5.2 Determining Sample Size

A power analysis was conducted for this study and based on using two groups, power of 0.81, a medium effect size of 0.25, and $\alpha = 0.05$ the calculation resulted in two groups of 64 participants for a total of 128 participants based on Table C.1 on page 384 of [33]. The same sample size was calculated using the G*Power3 software [13]. In this example, the test family was the F-test, the statistical test was the MANOVA: repeated measures, within-between interaction, and the type of power analysis was—A priori: compute required sample size-given α, power, and effect size. The effect size was based on a medium effect using Cohen's $\hat{f} = 0.25$, $\alpha = 0.05$, power was set at 0.80, the number of groups was 2 (standard versus emotive), and the number of measurements was also 2 for the two robots (Inuktun versus Packbot) resulting in a calculated sample size of 128 participants (See Fig. 3 with the input values and sample size highlighted with red boxes). Based on the analysis of the data the effect sizes were small to medium depending on the analyses performed and had there not been such a large sample size used for the study, some of the results may not have been statistically significant. In the self-assessment data, statistically significant results were obtained for the main effect of arousal and a three-way interaction was significant for valence [5]. If there is existing data available, then the effect size can be calculated using Cohen's $\hat{f}$ effect for any significant F-tests (Refer to Sect. 3) and used as input in the a priori power analysis [9].

5.3 Methods of Evaluation

This study utilized four methods of evaluation (self-assessments, video-recorded observations, psychophysiology measurements, and a structured audio-recorded interview) so that convergent validity could 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 studies. The assessments

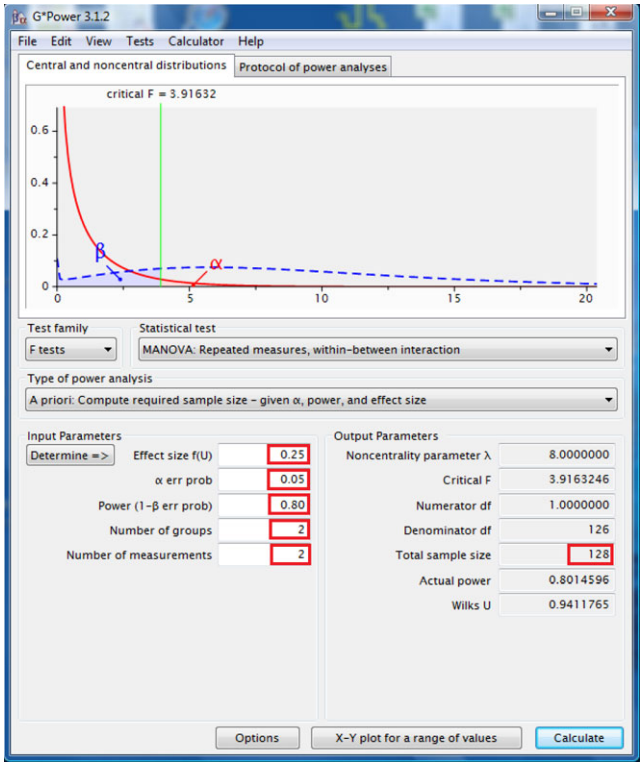

Fig. 3 G*Power3 example using the data for the exemplar study

were given to the participants prior to any interactions and after each robot interaction. It is recommended to conduct pilot studies of all the assessments to ensure that they are understandable and testing exactly what was expected. In this study, some of the questions were confusing to the participants and were not considered as part of the data analyses. It is important to note the questions that participants found confusing and/or required further explanation. In the case of one assessment, the Self-Assessment Manikin (SAM) [7] the valence and arousal questions were easily interpreted; however the questions relating to the dominance dimension were often misunderstood. That dimension was not included as part of the data analyses. The questions associated with the dominance dimension of the SAM assessment will need to be reworded and then validated; however the valence and arousal portions of the SAM assessment have been validated for future HRI studies and are available [2].

There were five different psychophysiological signals recorded as part of this study: (1) EKG, (2) Skin Conductance Response, (3) Thoracic Respiration, (4) Abdominal Respiration, and (5) Blood Volume Pulse, using the Thought Technology ProComp5 Infiniti system (http://www.thoughttechnology.com/pro5.htm). Five signals were used for obtaining reliable and accurate results. Correlations were conducted between the different signals to determine the validity of participants' responses and there was support between the heart rate variability and respiration rates. There

Fig. 4 Confined space simulated disaster site

was also support in the findings for heart rate variability, respiration rates, and the self-assessment data to provide validity in the results of this study.

Videotaped 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 infrared devices. When recording video observation data, synchronizing multiple cameras can be a challenge. In the case of the this study 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 is to use a sound that all cameras can detect through built-in microphones. A visual summary of this study can be viewed in a video format in [3]. 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 prior to participating in the study. They were given the option to deny publication of their video and audio recordings and three participants elected to deny publication of their recordings. It is important to note this denial in all files and related documents for their protection.

5.4 Study Location and Environment

The application domain for the study was Urban Search & Rescue, which required a confined-space environment that simulated a collapsed building (See Fig. 4). 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 with the robots exhibiting either standard or emotive behaviors. The medical assessment path, traveled by the robots,

was developed from video observations of experiments conducted by Riddle et al. with emergency responders and medical personnel based on how they would operate a robot to conduct a medical assessment of a trapped victim [22, 29]. Ideally, it would have been better to conduct the study in a real disaster or even a training exercise; however due to practicality and physiological measures, the study was conducted in a temperature-controlled environment.

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 exemplar study 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 had smaller body sizes than average and the space was truly confining to only a small portion of the participants. To increase the feeling of confinement, a blanket or rough heavy plastic or fabric that would crinkle or make audible sounds 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. Without these changes the results were statistically significant; however the impact and effect might have been greater if the environment was more realistic.

5.5 Failures and Contingencies

In this study, the "no show" percentage was much lower than the expected, at approximately 8%; equipment failures did occur. Making sure there are contingencies for equipment cannot be stressed enough. This 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 who could not be rescheduled. Following that experience extra sensors were ordered and kept on hand and they were needed. Video cameras had auto-focus problems that were not noticed until the video data was being off-loaded. 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 unusable. It is important always to double check equipment settings and verify all equipment is working properly so that no data is lost or determined to be unusable. 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.

6 Conclusions

Planning, designing, and executing 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 larger sample sizes to appropriately represent the population being studied, and so that small to medium effects can be determined with statistically significant results; and (2) the use of three or more methods of evaluation to establish reliable and accurate results that will have convergent validity. From our experiences in completing a large-scale, complex, controlled human study in HRI we present recommendations that fall into three categories: (A) Experimental Design Recommendations, (B) Recommendations for Study Execution, and (C) Other Recommendations.

6.1 Experimental Design Recommendations

These recommendations are presented to assist with the planning and design of large-scale, complex human studies in HRI. They will assist researchers with the development of a comprehensive experimental design that should provide successful study results.

1. Determine the most appropriate type of study for the hypotheses being investigated using a within-subjects, between-subjects, or mixed-model factorial 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 of obtaining statistically significant results that are valid, reliable, and accurate. This can be accomplished through power analysis tables or available software.
3. Determine the best methods of evaluation for the hypotheses being investigated; however, it is recommended to utilize three or more methods to obtain convergent validity in the study.
4. Design a study environment that closely reflects the real-world that is being tested for more natural participant responses. When conducting psychophysiology studies using skin conductance response, a temperature-controlled environment is essential.
5. If the goal of the research is to generalize results to different robots, perform the study with more than one type of robot.

These recommendations offer guidelines and practical tips to determine the best approach to design a comprehensive study in robotics. This will increase the probability that results will be statistically significant.

6.2 Recommendations for Study Execution

The following recommendations are provided to facilitate the execution of the study experimental design. These recommendations will assist in revealing potential flaws in the experimental design so that corrections can be implemented resulting in a smooth running, efficient study. However, even with the best designs you can expect equipment failures, participants and assistants arriving late or not at all, and other pitfalls. The key is to have contingency plans in place and anticipate worst case scenarios because they do occur.

1. Develop a written study protocol of all instructions, assessments with ordering, participant tasks in order of execution, timing of events, coordination of data collection, and any associated activities. This study protocol document will be used when preparing IRB paperwork, creating instructions for participants, and preparing informed consent documents.
2. 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.
3. Make sure that there is redundancy in all equipment that is required for the study and that backup equipment is always ready to use because failures are common.
4. Always prepare for the unexpected with contingency plans in place to handle equipment failures, participants and/or research assistants not arriving at their designated times, or other events.
5. Always allow time for study delays, participants arriving late, or equipment failures that may cause the cancellation of participants and delay of the study.

6.3 Other Recommendations

The following are recommendations for the recruitment of participants and volunteer research assistants and are based on our experiences and may not apply to all researchers and universities. They were excellent resources for our particular study and we are aware of similar programs available at many United States and European universities and institutions.

- Recruit quality volunteer research assistants 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 are willing to volunteer.
- Recruit participants through the use of flyer's 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 and/or other departments on campus if offered.
- Recruit the general public by requesting permission to post flyers at local malls, stores, or applicable agencies to represent the target population.
- Obtaining permission to recruit children from local schools, museums, and organizations.

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 three or more methods of evaluation, convergent validity should be obtainable. Readers are directed to [12, 16, 33] or other research methods books for further reference.

6.4 Impact of Valid Human Studies on HRI and the Use of Robots in Society

The area of Human-Robot Interaction is an emerging field and as such it is essential to use good research methods and statistical testing when conducting human studies. The use of appropriate sample sizes and three or more methods of evaluation can provide validity and credibility to the human studies that are performed associated with HRI. This will improve the overall field, but also will result in stronger public acceptance of robots. The public will be more likely to accept robots in their homes, schools, work environments, and as entertainment if they know that the use of these robots has been thoroughly tested for safety and effectiveness using good experimental methodology. Additionally, the engineering community will be able to use the information obtained from well conducted user studies to design and build better robots.

Acknowledgements Special thanks go to John Ferron, Kristen Salomon, Jennifer Burke, Jodi Forlizzi, Dewey Rundus, Larry Hall, AAAI Spring Symposium 2009, and the volunteer research assistants (Brian Day, Christine Bringes, Megan Brunner, Andrea Vera, Leslie Salas, Stephanie Smith, Kimberlee Fraser, Cherisse Braithwaite, and Caitlin Howell) for their assistance.

References

1. Bartneck C, Kulic D, Croft E, Zoghbi S (2008) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 2009(1):71–81
2. Bethel CL (2009) Robots without faces: non-verbal social human-robot interaction. Dissertation, University of South Florida
3. Bethel CL, Bringes C, Murphy RR (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), San Diego. ACM, New York

4. Bethel CL, Salomon K, Burke JL, Murphy RR (2007) Psychophysiological experimental design for use in human-robot interaction studies. In: The 2007 international symposium on collaborative technologies and systems (CTS 2007). IEEE, Orlando
5. Bethel CL, Salomon K, Murphy RR (2009) Preliminary results: Humans find emotive non-anthropomorphic robots more calming. In: 4th ACM/IEEE international conference on human-robot interaction (HRI2009), San Diego, CA
6. Bethel CL, Salomon K, Murphy RR, Burke JL (2007) Survey of psychophysiology measurements applied to human-robot interaction. In: 16th IEEE international symposium on robot and human interactive communication, Jeju Island, South Korea. IEEE, New York
7. Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25:49–59
8. Burke JL, Murphy RR, Riddle DR, Fincannon T (2004) Task performance metrics in human-robot interaction: taking a systems approach. In: Performance metrics for intelligent systems, Gaithersburg, MD
9. Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Earlbaum, Hillsdale
10. Dautenhahn K, Walters M, Woods S, Koay KL, Nehaniv CL, Sisbot A, Alami R, 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). ACM Press, New York, pp 172–179
11. Elara MR, Wijesoma S, Acosta Calderon CA, Zhou C (2009) Experimenting false alarm demand for human robot interactions in humanoid soccer robots. Int J Soc Robot 2009(1):171–180
12. Elmes DG, Kantowitz BH, Roediger HL III (2006) Research methods in psychology, 8th edn. Thomson-Wadsworth, Belmont
13. Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*power 3: A flexible statistical power analysis program for social, behavioral, and biomedical sciences. Behav Res Meth 39(2):175–191
14. Goodwin CJ (2003) Research in psychology-methods and design. Wiley, Hoboken
15. Itoh K, Miwa H, Nukariya Y, Zecca M, Takanobu H, Roccella S, Carrozza MC, Dario P, 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, Beijing, China, pp. 2620–2625
16. Johnson B, Christensen L (2004) Educational research quantitative, qualitative, and mixed approaches, 2nd edn. Pearson Education, Boston
17. Kidd CD, 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, Hatfield, Hertfordshire, pp. 141–142
18. Kulić D, Croft E (2006) Physiological and subjective responses to articulated robot motion. Robotica 15(1) 13–27. doi:10.1017/S0263574706002955
19. Lazar J, Feng JH, Hochheiser H (2010) Research methods in human-computer interaction. Wiley, New York
20. Liu C, Rani P, 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), Beijing, China, pp. 3099–3106
21. Moshkina L, Arkin RC (2005) Human perspective on affective robotic behavior: a longitudinal study. In: IEEE/RSJ international conference on intelligent robots and systems (IROS 2005), pp. 2443–2450
22. Murphy RR, Riddle D, 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), pp. 301–306
23. Mutlu B, Hodgins JK, 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), IEEE, Genova, Italy
24. Mutlu B, Osman S, Forlizzi J, Hodgins JK, 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). IEEE, University of Hertfordshire, Hatfield
25. Olsen DR, Goodrich MA (2003) Metrics for evaluating human-robot interactions. In: Performance metrics for intelligent systems workshop
26. Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans Pattern Anal Mach Intel 23(10):1175–1191
27. Preece J, Rogers Y, Sharp H (2007) Interaction design-beyond human-computer interaction, 2nd edn. Wiley, New York
28. Rani P, Sarkar N, Smith CA, Kirby LD (2004) Anxiety detecting robotic system—towards implicit human-robot collaboration. Robotica 22(1):85–95
29. Riddle DR, Murphy RR, Burke JL (2005) Robot-assisted medical reachback: using shared visual information. In: IEEE international workshop on robot and human interactive communication (RO-MAN 2005), IEEE, Nashville, TN, pp. 635–642
30. Schweigert WA (1994) Research methods & statistics for psychology. Brooks/Cole Publishing Company, Pacific Grove, CA
31. Shaughnessy JJ, Zechmeister EB (1994) Research methods in psychology. McGraw-Hill, New York
32. Steinfeld A, Fong T, Kaber D, Lewis M, Scholtz J, Schultz A, 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, New York
33. Stevens JP (1999) Intermediate statistics a modern approach, 2nd edn. Erlbaum, Mahwah
34. Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the Panas scales. J Pers Soc Psychol 54(6):1063–1070

Cindy L. Bethel is an NSF Computing Innovation Fellow in Computer Science at Yale University. In 2009, she received her Ph.D. in Computer Science and Engineering from the University of South Florida. She was an NSF Graduate Research Fellow and IEEE Robotics and Automation Society Graduate Fellow. Her research focuses in the areas of human-robot interaction, affective computing, robotics, and artificial intelligence. Cindy L. Bethel received a B.S. in Computer Science from the University of South Florida in 2004.

Robin R. Murphy received a B.M.E. in Mechanical Engineering, a M.S. and Ph.D. in Computer Science in 1980, 1989, and 1992, respectively, from Georgia Tech. She is the Raytheon Professor of Computer Science & Engineering at Texas A&M. Her research interests are artificial intelligence, human-robot interaction, and heterogeneous teams of robots. In 2008, she was awarded the Al Aube Outstanding Contributor award by the AUVSI Foundation, for her insertion of ground, air, and sea robots for urban search and rescue at the 9/11 World Trade Center disaster, Hurricanes Katrina and Charley, and the Crandall Canyon Utah mine collapse. She is an associate editor for IEEE Intelligent Systems, a Distinguished Speaker for the IEEE Robotics and Automation Society, and has served on numerous boards, including the Defense Science Board, USAF SAB, NSF CISE Advisory Council, and DARPA ISAT.