

Conducting Studies in Human-Robot Interaction

Cindy L. Bethel, Zachary Henkel, Kenna Baugus

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, such as those established in psychology and the other social sciences. Two primary concerns observed in HRI studies are (1) the lack of appropriate sample sizes that closely represent the populations being studied and has sufficient statistical power, and (2) the lack of three or more methods of assessment used to obtain convergent validity in HRI studies [21, 30, 32].

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 [32]; 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, in addition to how they may operate robots [1, 3, 9, 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 [3, 7, 9, 21, 32, 6, 49]. From the review of HRI literature, it appears the most common methods used in HRI studies are self-assessment, behavioral measures, and task performance metrics. There is limited research in the use of psychophysiological measures and interviews. Though

Cindy L. Bethel, Zachary Henkel, Kenna Baugus
Mississippi State University, Department of Computer Science and Engineering, P.O. Box 9637,
Mississippi State, MS 39762-9637, e-mail: c1b821, zmh68, kbb269@msstate.edu

using psychophysiological measures seems to be increasing in its use in HRI. A cautionary note, there has been an increase in the use of psychophysiological measures with the claims that this method can be used to identify specific emotions just from physiological measures, and physiology signals measure levels of arousal and not valence or how positive or negative a person is feeling. Therefore, it is not possible from the physiological signals alone to determine specific emotions such as a happy or to distinguish between surprise and anger, which are both high arousal emotions. Arousal levels are the intended method of measurement for physiological measures [12]. Each method of evaluation has advantages and disadvantages; however the disadvantages can be overcome by using more than one method of evaluation [3, 9, 32, 6].

The design of quality research studies for use in HRI applications with results that are verifiable, reliable, and reproducible is a major challenge [1, 6]. 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 [49]. This is still an issue today, though more researchers are dedicated to addressing this important issue to improve the quality of HRI research. Until recently, the discussion for the development of common metrics for HRI has been oriented toward an engineering perspective and this does not completely address the social interaction perspective. Both the engineering and social interaction perspectives require further investigation to develop standard and common metrics and methods of evaluation. There is a significant need for new methods and metrics to advance and validate the field of human-robot interaction [33]. There are some social science researchers that are working toward establishing methodological approaches to HRI [15]. There is a strong need to explore longitudinal studies in *real-world* settings moving studies away from lab settings with college-aged students and convenience samples [17, 33].

This chapter 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 Section 3. Next, is a discussion of the process of planning and designing a human study in HRI in Section 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, selection of robots and other equipment, find assistants to conduct the study, recruit the required number of participants, development of contingency plans to deal with failures, and preparation of Institutional Review Board (IRB) and/or ethics documents. In Section 5, illustrative examples are drawn from a 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 positive 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. Portions of this chapter come from two previous articles published by the lead author [5][6].

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 of evaluation was utilized in the studies. One issue observed with these studies is that sample sizes are often relatively small, and therefore may not have been representative of the population being investigated, which may have influenced the results. There needs to be an appropriate sample size used in order to have sufficient statistical power to make claims and generalize the results obtained from the data collected [6]. In some cases with smaller sample sizes statistical significance appears to be achieved; however the results then only apply to that particular population and do not generalize to a larger or different population.

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 or survey; however in some cases, the researchers add other methods of assessment such as video observations and coding (although the results are often not presented for this data), psychophysiology measurements, and/or task performance measures. A thorough evaluation of three years of human-robot interaction papers was conducted by Baxter *et al.* with recommendations provided by the authors for areas of improvement when conducting human studies in HRI. Their first recommendation was to not only have larger sample sizes, but to ensure that these samples are consistent with the domain and application area being explored (e.g., using children for child HRI, using elderly when investigating the impact of robots in nursing homes) [2]. It is important to not conduct studies using a convenience sample of college-aged students especially when investigating topics that are not relevant to that population. Their second recommendation is to take the experiments out of the laboratory controlled setting and move them into the real-world or *in the wild* in order to establish ecological validity [2]. The third trend in HRI studies is that studies are conducted typically in hours and at most a few days. It is important when determining the impact of robots, that longer-term interactions are observed and evaluated to have a better understanding of how people respond to robots once the novelty effect wears off [2, 17]. Another significant issue is how statistics are reported. Baxter *et al.* discuss the use of *p-values* as the standard for statistically significant results. It is important to also include the actual statistics, descriptive statistics, and also include the effect size calculations for all statistically significant results with a scale for the interpretation of the effect sizes [2, 6]. This information provides details on how much of an impact the statistically significant results may have. There is typically a test for effect size for most statistical tests and it may require some manual calculations, but should always be included with all statistically

significant results reported. The final point made in the paper by Baxter *et al.* is that studies should be designed and reported for future replication by the HRI community. There needs to be sufficient details provided so that others can replicate the procedures for any HRI study and it would be expected that those results should be consistent. This is rarely performed and needs to be more carefully considered.

2.1 HRI Study Examples

Dautenhahn *et al.*: One of the more comprehensive studies performed by Dautenhahn *et al.* utilized self-assessments, unstructured interviews, and video observations from one camera angle [18]. 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 [6].

Moshkina and Arkin: Another study that incorporated multiple methods of evaluation was performed by Moshkina and Arkin [37] in which they used self-assessment measures including a measure commonly used in psychology studies called the Positive and Negative Affect Schedule [52]. 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 [6].

Mutlu *et al.*: A study conducted by Mutlu *et al.* used both task performance measures and self-assessments [39]. 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 [6].

Kulić and Croft: 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 [34]. Multiple psychophysiological signals were measured which is highly recommended for this type of study for reliability and validity in the results [3][7][29][32][36][43][45]. 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 [6]. A major issue is the use of physiological measures for determining valence. These signals are intended to measure arousal, and are not intended to determine specific emotions or valence of a participant. There is a growing trend to use these measures for determining specific emotions through the use of machine learning, and this is not what these measures are intended for or accurately measure. Facial coding could be used to determine the actual emotional expression of the participant, but it is really not possible to identify specific emotions with any accuracy

through the use of physiological measures alone (e.g., heart rate, respiration, skin conductance response, etc.) [7, 9].

Mutlu *et al.*: They 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 [40]. 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.

Fisicaro *et al.*: More recently, Fisicaro *et al.* [23] performed a preliminary evaluation of plush social robots in the context of engaging people with neuro-developmental disorders. Eleven medium-low intellectual functioning participants listened to a story told by a human speaker one week and told by the ELE plush elephant robot the next week. Behavioral measures of the time looking at the speaker and the number of times the participant shifted focus away from the speaker indicated that the ELE robot speaker was more engaging than the human speaker. Though this study was conducted on a small sample, it illustrates that a larger study of this type of interaction should be investigated in the future. It is still quite common to use smaller sample sizes, which limits the impact and generalizability of the results of the research.

Vogt *et al.*: [51] examined the effects of a social robot added to a tablet computer application designed to teach new English vocabulary words to young children. In total 194 children participated in the study, which involved seven sessions taking place over a three-week time period. Depending upon condition, participants experienced vocabulary lessons administered by the tablet application alone, the tablet application and a robot using iconic gestures, the tablet application and a robot not using iconic gestures, or were part of a control group that danced each week rather than learning English vocabulary. Performance was measured using two translation tasks and a comprehension task within two days of the final session and again within two weeks of the final session. Statistical comparisons between groups indicated vocabulary gains for all groups (except the control condition) and found that the presence of a social robot did not significantly affect the level of vocabulary gains achieved by participants. This is a good example of using a larger sample size with appropriate participants. It is also good that the study was conducted over multiple sessions versus a single interaction. This reduces the impact of any potential novelty effects. In this case, having a larger sample likely would not provide any additional knowledge and indicates the importance of having an appropriate sample size for the research.

Conti *et al.*: [16] recruited fifty-two students from the same kindergarten and visited the children's school for three sessions throughout a three-week period. The sessions were held in a classroom with their teacher to insure the children felt comfortable.

During the first session, children were asked to design a robot and given a blank piece of paper, colored pencils, and encouragement to discuss robots with their peers. Once the drawing was completed, each child participated in an interview

and answered open-ended questions about the characteristics he or she attributed to the robot. The second session consisted of interacting with the Softbank Robotics humanoid robot, Nao. In the study, the robot introduced himself, danced and played music, and told the children a story. In the third and final session, children were again asked to design and draw a robot, then completed the same interview process as the initial session.

After analyzing the children's responses and drawings, the researchers found that after meeting the robot, children included more colors in their designs. They observed a ten percent decrease in reports of aggressive behaviours performed by the robot when asked "What is the robot doing?" and an overall positive reaction and decrease of distress regarding meeting a robot. The researchers emphasized that an early introduction to robots can facilitate positive perceptions of robots and technology. Because of this being a vulnerable population, this study represents a larger sample size. The study may have more impact with more participants; however this demonstrates the benefit of larger samples and repeated measures over time versus a single interaction.

Chien *et al.*: [13] recruited twenty-four younger (20 to 24 years) and twenty-four older (59 to 86 years) participants to compare implicit and explicit attitudes towards assistive robots. The study measured explicit attitudes using self-report questionnaires and implicit attitudes via an implicit association test (IAT) [27].

Prior to interacting with a Zenbo home robot, participants completed self-report surveys that included: Negative Attitudes toward Robots Scale (NARS) [41], Technology Acceptance Model (TAM) [19], and Subjective Technology Adaptivity Inventory (STAI) [31]. After interacting with the robot, participants completed these measures again and completed an implicit association test (IAT) that examined pairing the concept of robot or human with negative or positive words. IATs measure the ability of and time taken for a participant to correctly categorize a stimulus that appears on a computer screen. Shorter categorization times indicate a stronger association between the two concepts. By asking participants to respond as quickly as possible and balancing categorization conditions, social desirability biases in participant responses can be avoided.

The results indicated that although both younger and older participants had similar increases in explicit measures of positive attitudes towards assistive robots after interacting with a robot, on average older participants had more implicit negative associations towards assistive robots than younger participants. The sample size is larger than several HRI studies, but may have benefited from having an even larger population.

2.2 Summary

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 using study design techniques that 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., [21][26][28][35]).

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 d effect (for statistically significant t -tests) is [14]:

$$d = \frac{M_1 - M_2}{\sigma}$$

where, d is the effect size index for t -tests of means in standard unit, M_1 , M_2 are the population means expressed in raw (original measurement) units, and σ is the standard deviation of either population (since they are assumed equal).

The scale used to interpret Cohen's d effect is:

- 0.00 - 0.19 negligible effect
- 0.20 - 0.49 small effect
- 0.50 - 0.79 medium effect
- 0.80 + large effect

The formula for calculating Phi ϕ effect (for statistically significant Chi-Square χ^2 tests) is:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

where, χ^2 is the statistically significant result of a Chi-square test and n = the total number of observations.

The scale used to interpret a ϕ effect is:

- 0.00 - 0.09 negligible effect
- 0.10 - 0.29 small effect
- 0.30 - 0.49 medium effect
- 0.50 + large effect

The formula for calculating Cohen's $\hat{f}$ effect (for statistically significant F tests) is [14]:

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

where, df = the degree of freedom for the numerator (number of groups - 1), F = the statistically significant result of an F-test (e.g., from an ANOVA statistical 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?

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.

Within-Subjects Design: 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 their responses are reduced due to repetitive presentation of the same tasks or the robot performs in a similar manner reducing the novelty effect the robot over time. 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) [47].

Between-Subjects Design: 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 [35]. 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 or male participants versus female participants. 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 because participants

Chronology of Planning, Designing, and Executing a Human Study in HRI

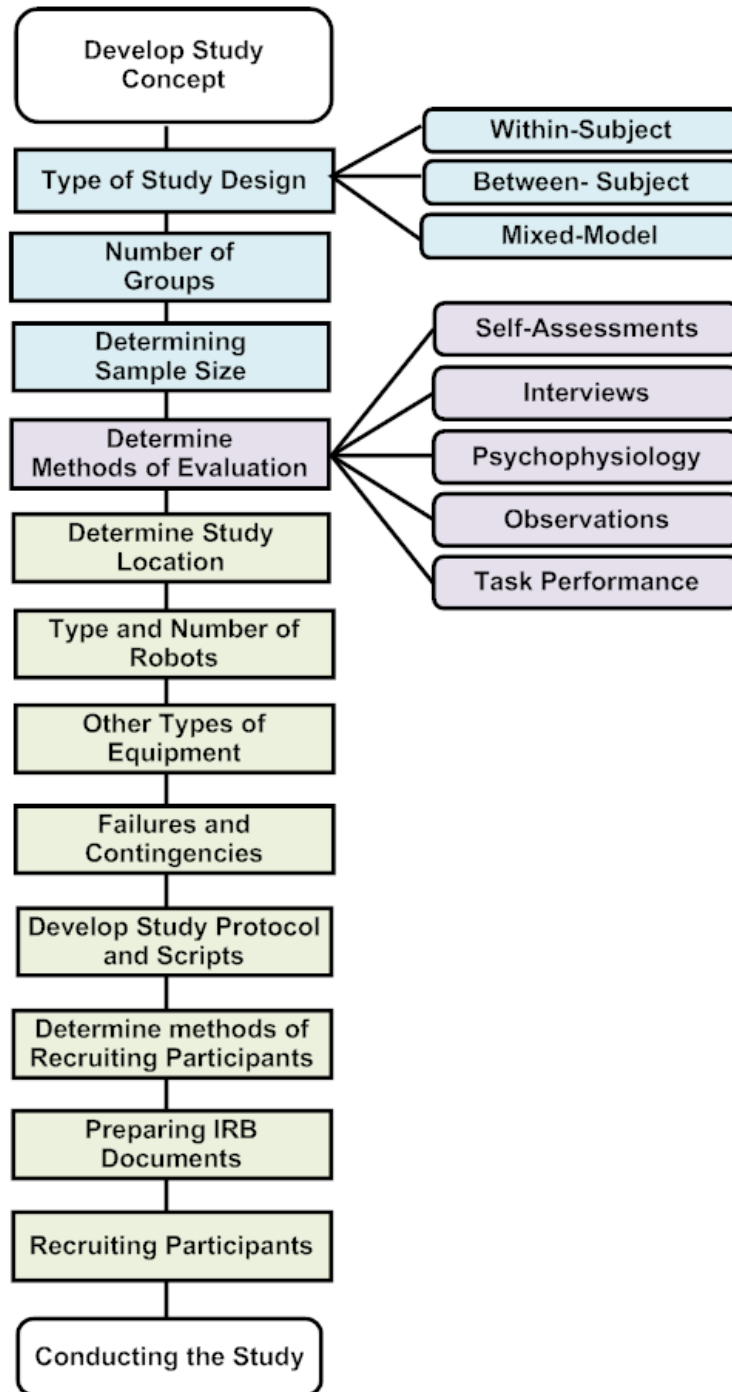


Fig. 1 The chronology of items required for planning, designing, and executing human studies in HRI.

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 [35].

Mixed-Model Factorial Design: 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 [35]. 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. Subsection 5.1 provides an example of a mixed-model factorial design.

4.2 Determining Sample Size

Determining the appropriate sample size is often a challenge in human studies in HRI, though it is important to determine when you are in the planning phase for a study. 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 [50]) that will provide group size values. Additionally, there is software available online that will assist with this type of calculation (e.g., G*Power3.1 – <http://www.pscho.uni-duesseldorf.de/abteilungen/aap/gpower3/>). Subsection 5.2 presents an example for calculating sample size using statistical tables and the G*Power3.1 software. When calculating the sample size, if you do not know what your effect size is expected to be based on results from prior studies, then a common approach is to use a medium effect size for the power analysis (in the example case that was 0.25).

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 [3][7][9][32][35][44]. 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 them in a different way.

When planning a human study, it is important to use credible measures that are appropriate for the study and for the 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, interviews and/or psychophysiology measures such as respiration rate, electrocardiography (EKG/ECG), and skin conductance response. Each of these measures would be appropriate for this type of study and these are credible measures for assessing participants' levels of stress and anxiety. The use of multiple methods of evaluation that are not credible or appropriate will result in data that is not meaningful to answer the hypotheses and research questions.

Self assessments are among the most commonly used method of evaluation in HRI studies; however obtaining validated assessments designed for HRI studies can be a challenge. There are several surveys available in the HRI community but very few have been validated. There are surveys available in the psychology literature that are validated that may be useful, but if they are modified then they need to be validated again. 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 [21]. 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 [21][30]. Another issue with self-assessments is that participants will complete them after performing tasks

and may not recall exactly how they felt during the actual interaction. 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 [3][9].

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" [30]. 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 [21][30]. 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 [3][9][21]. Video observations of human-robot interactions are often recorded and later coded for visual and/or auditory information using two or more independent raters [11]. 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 chapter (refer to [25][35][44][47][48]). Behavioral measures and observations are often collected but rarely is the data analyzed and published. It is a very tedious and time consuming process that takes significant personnel to complete. If it is analyzed, the resulting data is usually very rich and more informative than using only self-assessment data. Unfortunately many researchers never end up processing this data.

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 [29][32][34][36][43][45]. Also, psychophysiological measures offer a minimally-invasive method that are used to determine the stress levels and arousal responses of participants interacting with technology [29][34][36][43][45]. 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 [3][7][9]. An issue of concern that is becoming more frequent is the use of data mining and machine learning techniques with the physiological signals and researchers are claiming that they can use these signals to identify specific emotions. The physiological signals such as heart rate and respiration rate are designed to measure levels of arousal. They are not going to indicate valence aspects of emotion. The results are often inconclusive in the literature and this is because it is not what these signals are intended to

measure. They will provide only the level of arousal a participant is experiencing at the time of interactions. Another issue with these signals are that they only measure what a person feels and if the interaction does not provide significant arousal then the signals will not provide any definitive insights. These are signals that cannot be manipulated consciously.

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 same 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 and details 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 and feelings [21]. 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. The data from interviews, like behavioral measures are often difficult, tedious, and time consuming to analyze. Even though interview data is collected, it is often not analyzed. It also takes training in transcribing, coding, and analyzing the data. If the time is taken to perform the analysis, then the result is often rich data that gives additional meaning to the research questions. When performing research with children, the interview is helpful in obtaining accurate responses from them. Young children are not able to respond to survey questions that are rated on a scale. They are capable of responding to yes/no questions and will provide additional knowledge and information through interviews.

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 [11][20][39][42][49]. 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. Task performance measures are useful in determining if technology is helping to improve the performance of a specific task. It typically measures aspects like time of completion for a task and/or the number of errors or problems encountered. It is an excellent method for evaluating how well teams of people work together and how they work with technology. Task

performance metrics are considered objective measures that have concrete values associated with them.

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 used. 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 obtain 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 as it relates to the research question(s). 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 or negative 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 [3][7].

In the example study, provided in Section 5, placing participants in an actual search and rescue site of a building collapse was not practical or possible. Therefore a simulated disaster sight was created that contained rubble, was dark, and the temperature was kept on the cooler side. Participants were placed in a confined space similar to what might be experienced in an actual disaster situation.

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. Sometimes you may not have access to a particular robot, but it is important to try and get a robot that would work well in the environment and situation being investigated. For example, you would not likely put a Nao humanoid robot, that is small and has difficulty with mobility, into a search and rescue site.

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 communities.

4.6 Other Equipment

Determining what equipment will be used in an HRI study impacts the success and results of the study. Whenever possible, equipment choices should be redundant. Unfortunately, equipment failures are more common than you want to believe and it is important to make sure that there are contingency plans in place in times of failure.

Video Observation Studies: When performing video observation or behavioral studies, the first step is to determine the number of different perspectives to be recorded 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 and gives you some redundancy in case there is a problem with a particular recording. It is important that cameras are synchronized and extra batteries and SD cards/tapes are readily available. There should be at least one or two extra cameras available in case of equipment failure that are available to swap out if needed. It is also advisable not to reuse tapes if at all possible because it can impact the integrity of the recordings. This is less of a problem with current video cameras as they typically now use SD cards for recordings. To preserve data and prevent mishaps, it is important to off-load recordings quickly to a more stable media. It is recommended that all data be backed up to multiple locations and media.

Psychophysiological Studies: 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 in which the participant is mobile. Currently, there are more options available, but they are often expensive if you are using re-

search grade systems. Another option is to use smart watches, as some will now offer recording capabilities, and are becoming more accurate in the recordings. It depends on what is being measured if that is an option. In a study I conducted recently, heart rate data was compared between an Apple watch and a QardioCore system and the heart rates were identical (<https://www.getqardio.com/qardiocore-wearable-ecg-ekg-monitor-iphone/>). This was a very limited study, and a more formal and extensive study should be conducted, but it is expected to be a less expensive option for recording heart rate data.

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. Physiological equipment is becoming more stable and reliable, but it is still prone to issues.

4.7 Failures and Contingencies

Even with careful planning, failures and problems are likely to occur when conducting studies. 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. Even with contingency plans, problems will occur and it is important to be as prepared as possible to deal with these problems to avoid delays and risks to your data collection.

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 Institutional Review Board (IRB) or Ethics Committee paperwork required for human studies and for determining risks and maintaining privacy of participants in the United States and in many countries. The process may be different depending on the country, but regardless of whether it is required in your country or not, the preparation of a study protocol is good practice to ensure a high quality study.

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 when data is actually collected. 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 many of the studies conducted in HRI do not have large or appropriate sample sizes. It is important to recruit participants that will appropriately represent the population being studied. If your research involves questions associated with children, then your study population should be using children as participants. If you are performing research with first responders then it is important to include them as your participants in the study. In the case of special populations, it may be more difficult to get larger or appropriate sample sizes, but every effort should be made to use the target population for the research questions. Too often, the population selected is a convenience sample, that typically involves college-age students because that is what is most available to university researchers. It may be challenging, but it is important to the validity of the research to use a population that represents the end user and evaluates the technology being investigated. 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 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. You must obtain written permission to recruit participants from these different populations. If you work with first responders, you need to contact the agencies and request written permission to use their employees as participants. For schools, it may be necessary to present your study before the school board to obtain permission to use the students. There are special considerations if you are working with military personnel. If you are developing research for these populations, you need to use them to determine the viability of the research.

4.10 Preparing Institutional Review Board or Ethics Committee Documents

The next step in planning, designing, and executing a human study is the preparation of the Institutional Review Board or Ethics Committee 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) or the Ethics Board are committees at each university. The process is similar at most universities and in other countries if they have this requirement. In some cases it may be more extensive than what is described in this section. The committee members are from diverse backgrounds and they review proposals for research conducted with human subjects. The IRB or Ethics Boards were established to review the procedures of proposed research projects, to determine if there are any known or anticipated risks to participants in the study, methods of recruiting participants, and how the participants' confidentiality is 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 [47]. 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 or Ethics Board 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. Start this pro-

cess early because in some cases this process can take considerable time, especially if your research involves vulnerable populations such as children, elderly, prisoners, etc. If you are using a vulnerable population or your study has risks involved, then the study will often have to be approved by the entire board and this can take a month or more to complete that process, especially if changes are requested by the committee. It is important to keep this in mind when planning your data collection.

4.11 Recruiting Participants

Once IRB or Ethics Board 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 many 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 at least 24 hours in advance, but it may also be helpful to send a reminder an hour before his or her appointment. 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 and backup data.

A helpful scheduling tool is software to make timeslots available. This is often included in the software for participant pools, such as SONA or PRP that many university Psychology departments use in the United States. There are also software products like Calendly (<https://calendly.com/>) and YouCanBook.me (<https://youcanbook.me/>). These allow the researcher to link to their personal calendar and participants can sign up through the software that will place them on the researcher's calendar. This can be quite useful and the software can be setup to send reminders. In the case of the calendly.com software, you can also include questions related to inclusion/exclusion criteria that allows researchers to screen participants prior to them signing up for the study.

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 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 were 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. To date, this study remains one of the most complex studies of this scale in the HRI community.

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 viewed first or Packbot viewed first), and operating mode assignments were balanced for age and gender.



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

5.2 Determining Sample Size

An *a priori* power analysis was conducted for this study and was based on using two groups, power of 0.80, 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 [50]. The same sample size was calculated using the G*Power3.1 software [22]. 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 $\hat{f} = 0.25$ (results prior data collections were used to determine the effect size), $\alpha = 0.05$, 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 prior 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 [8]. 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 Section 3) and used as input in the *a priori* power analysis [14].

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

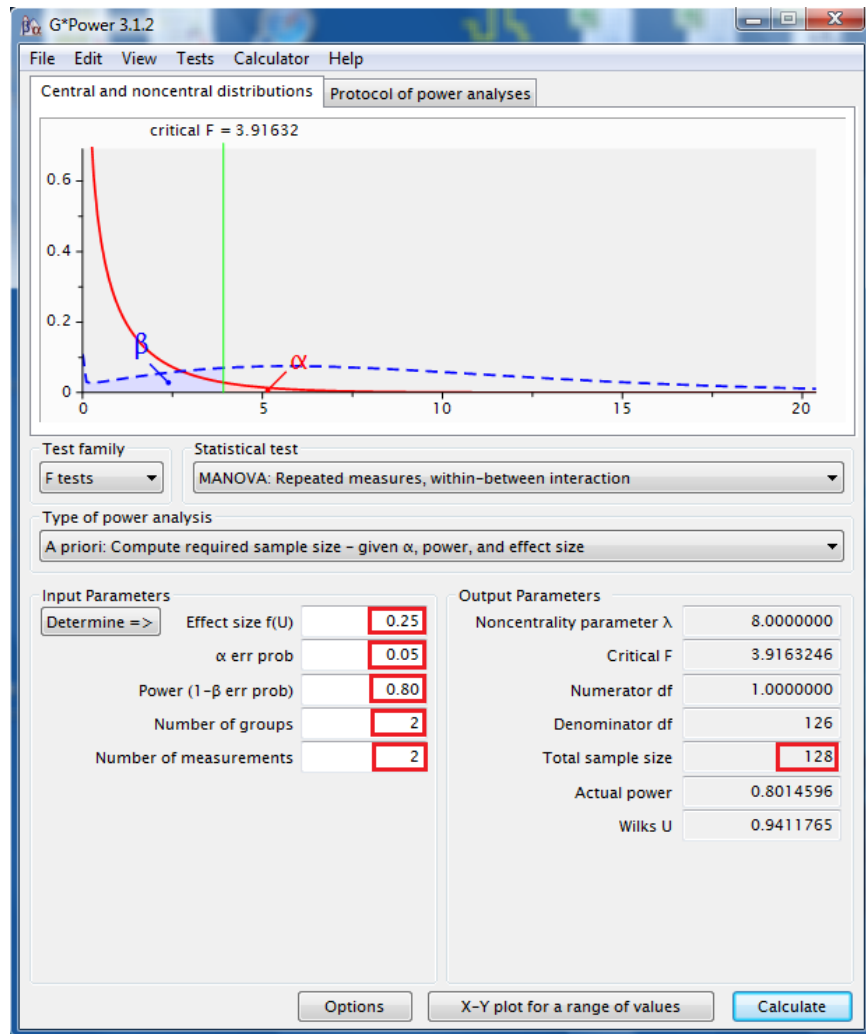


Fig. 3 G*Power3.1 example using the data for the exemplar study.

of the use of non-facial and non-verbal affective expression for naturalistic social interaction in a simulated disaster application.

Self-Assessments: 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 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 make note of the questions that participants found confusing and/or required further explanation. This can be done in a participant log or as part of their paperwork so that you are not relying on memory after the data collection.

In the case of one assessment, the Self-Assessment Manikin (SAM) [10] the valence and arousal questions were easily interpreted; however the questions related 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 [3].

As part of the validation process for self-assessments it is important to ask each question in at least two different but similar ways and then perform a statistical test known as Cronbach's alpha to determine tau-equivalent reliability or internal validity for the items on the assessment. A Cronbach's alpha value greater than .70 indicates acceptable validity for the items evaluated. For more information on Cronbach's alpha consult [24] or a statistical textbook.

Another issue associated with self-assessment data is in the analysis of this data. Likert scale and Semantic Differential scales are not continuous data. This type of data can be analyzed using t-tests or a more appropriate approach would be to perform a Chi-square or related test such as Kruskal-Wallis test for non-parametric data. Unfortunately, many researchers use an ANOVA test, which is not appropriate given that the scales are not continuous. The issue with the Chi-square test is that there needs to be at least five (5) items in each bin to run the test. This may be accomplished by combining categories such as strongly disagree and disagree in a Likert scale and agree and strongly agree categories. For the purpose of this chapter, the details of a Chi-square and/or the Kruskal-Wallis ranks tests will not be covered in detail, but most statistical software can be used to perform these types of analyses. It is important when analyzing data that the appropriate statistical test is used. It is also important that when reporting those statistical tests that the results from the actual tests are presented and for statistically significant results the p-value and the effect size for the statistical test used are also reported. You should include a scale for interpreting the effect size for the reader as a reference and state the level of effect (e.g., small, medium, large).

Psychophysiology Measures: 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 to obtain 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 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.

Video-Recorded Observations: 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 cameras and infrared illuminators. When recording video observation data, synchronizing multiple cameras can be a challenge. In the case of 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 the 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 [4].

Structured Interviews: 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 and to not accidentally use the materials. The interviews can be transcribed in detail, then coded and analyzed in a quantitative manner through categorizing the comments.

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 [38][46]. 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 popu-

lation 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.



Fig. 4 Confined space simulated disaster site with the lights on.

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 the experiences gained in completing a large-scale, complex, controlled human study in HRI recommendations are presented 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.
6. Include participants who would be the expected population associated with the research questions being investigated. For example, if the study is researching the responses of children to a robot, then use children as participants in the study.

These recommendations offer guidelines and practical tips to determine the best approach to design a comprehensive study in robotics and human-robot interactions. 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's 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 or Ethics Board 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 a volunteer hours requirement and are willing to volunteer.
- Recruit participants through the use of 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 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 [21][30][50] 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.

References

1. Bartneck, C., Kulic, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* **2009**(1), 71–81 (2008)
2. Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., Belpaeme, T.: From characterising three years of hri to methodology and reporting recommendations. In: *The Eleventh ACM/IEEE International Conference on Human-Robot Interaction*, pp. 391–398. IEEE Press (2016)
3. Bethel, C.L.: *Robots without faces: Non-verbal social human-robot interaction*. Dissertation, University of South Florida (2009)
4. Bethel, C.L., Bringes, C., Murphy, R.R.: 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, CA (2009)
5. Bethel, C.L., Murphy, R.R.: Use of large sample sizes and multiple evaluation methods in human-robot interaction experimentation. In: *2009 AAAI Spring Symposium Series, Experimental Design for Real-World Systems*. Palo Alto, CA (2009)
6. Bethel, C.L., Murphy, R.R.: Review of human studies methods in hri and recommendations. *International Journal of Social Robotics* (2010). DOI 10.1007/s12369-010-0064-9
7. Bethel, C.L., Salomon, K., Burke, J.L., Murphy, R.R.: Psychophysiological experimental design for use in human-robot interaction studies. In: *The 2007 International Symposium on Collaborative Technologies and Systems (CTS 2007)*. IEEE, Orlando, FL (2007)
8. Bethel, C.L., Salomon, K., Murphy, R.R.: Preliminary results: Humans find emotive non-anthropomorphic robots more calming. In: *4th ACM/IEEE International Conference on Human-Robot Interaction (HRI2009)*. San Diego, CA (2009)
9. Bethel, C.L., Salomon, K., Murphy, R.R., Burke, J.L.: Survey of psychophysiology measurements applied to human-robot interaction. In: *16th IEEE International Symposium on Robot and Human Interactive Communication*. Jeju Island, South Korea (2007)
10. Bradley, M.M., Lang, P.J.: Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry* **25**, 49–59 (1994)
11. Burke, J.L., Murphy, R.R., Riddle, D.R., Fincannon, T.: Task performance metrics in human-robot interaction: Taking a systems approach. In: *Performance Metrics for Intelligent Systems*. Gaithersburg, MD (2004)
12. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G.: *Handbook of Psychophysiology* Fourth Edition, fourth edn. Cambridge Handbooks in Psychology. Cambridge University Press, United Kingdom (2017)
13. Chien, S.E., Chu, L., Lee, H.H., Yang, C.C., Lin, F.H., Yang, P.L., Wang, T.M., Yeh, S.L.: Age difference in perceived ease of use, curiosity, and implicit negative attitude toward robots. *ACM Trans. Hum.-Robot Interact.* **8**(2), 9:1–9:19 (2019). DOI 10.1145/3311788. URL <http://doi.acm.org/10.1145/3311788>
14. Cohen, J.: *Statistical power analysis for the behavioral sciences*, 2nd edn. Lawrence Earlbaum Associates, Hillsdale, NJ (1988)
15. Compagna, D., Marquardt, M., Boblan, I.: Introducing a methodological approach to evaluate hri from a genuine sociological point of view. In: *International Workshop in Cultural Robotics*, pp. 55–64. Springer (2015)
16. Conti, D., Di Nuovo, S., Di Nuovo, A.: Kindergarten children attitude towards humanoid robots: what is the effect of the first experience? In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 630–631. IEEE (2019)
17. Dautenhahn, K.: Some brief thoughts on the past and future of human-robot interaction. *ACM Trans. Hum.-Robot Interact.* **7**(1), 1–3 (2018). DOI 10.1145/3209769
18. Dautenhahn, K., Walters, M., Woods, S., Koay, K.L., Nehaniv, C.L., Sisbot, A., Alami, R., Siméon, T.: 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)*, pp. 172–179. ACM Press, New York, NY, USA, Salt Lake City, UT (2006)

19. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* pp. 319–340 (1989)
20. Elara, M.R., Wijesoma, S., Acosta Calderon, C.A., Zhou, C.: Experimenting false alarm demand for human robot interactions in humanoid soccer robots. *International Journal of Social Robotics* **2009**(1), 171 – 180 (2009)
21. Elmes, D.G., Kantowitz, B.H., Roediger III, H.L.: *Research Methods in Psychology*, 8th edn. Thomson-Wadsworth, Belmont, CA (2006)
22. Faul, F., Erdfelder, E., Lang, A.G., Buchner, A.: G*power 3: A flexible statistical power analysis program for social, behavioral, and biomedical sciences. *Behavior Research Methods* **39**(2), 175–191 (2007)
23. Fisicaro, D., Pozzi, F., Gelsomini, M., Garzotto, F.: Engaging persons with neuro-developmental disorder with a plush social robot. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 610–611. IEEE (2019)
24. Gliem, J.A., Gliem, R.R.: Calculating, interpreting, and reporting cronbachs alpha reliability coefficient for likert-type scales. In: *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education* (2003)
25. Goodwin, C.J.: *Research in Psychology-Methods and Design*. John Wiley & Sons, Inc., Hoboken, NJ (2003)
26. Gravetter, F.J., Forzano, L.A.B.: *Research Methods for the Behavioral Sciences - Fifth Edition*, fifth edn. Cengage Learning, Stamford, CT, USA (2016)
27. Greenwald, A.G., McGhee, D.E., Schwartz, J.L.: Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* **74**(6), 1464 (1998)
28. Hatcher, L.: *Advanced Statistics in Research: Reading, Understanding, and Writing Up Data Analysis Results*. Shadow Finch Media, Saginaw, MI, USA (2013)
29. Itoh, K., Miwa, H., Nukariya, Y., Zecca, M., Takanobu, H., Roccella, S., Carrozza, M.C., Dario, P., Atsuo, T.: Development of a bioinstrumentation system in the interaction between a human and a robot. In: *International Conference of Intelligent Robots and Systems*, pp. 2620–2625. Beijing, China (2006)
30. Johnson, B., Christensen, L.: *Educational Research Quantitative, Qualitative, and Mixed Approaches*, 2nd edn. Pearson Education, Inc., Boston, MA (2004)
31. Kamin, S.T., Lang, F.R.: The subjective technology adaptivity inventory (stai): A motivational measure of technology usage in old age. *Gerontechnology* (2013)
32. Kidd, C.D., Breazeal, C.: Human-robot interaction experiments: Lessons learned. In: *Proceeding of AISB'05 Symposium Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction*, pp. 141–142. Hatfield, Hertfordshire, UK (2005)
33. Kiesler, S., Goodrich, M.A.: The science of human-robot interaction. *ACM Trans. Hum.-Robot Interact.* **7**(1), 1–3 (2018). DOI 10.1145/3209701
34. Kulić, D., Croft, E.: Physiological and subjective responses to articulated robot motion. *Robotica* **Forthcoming**, 15 (2006)
35. Lazar, J., Feng, J.H., Hochheiser, H.: *Research Methods in Human-Computer Interaction*. John Wiley & Sons, Ltd., West Sussex, United Kingdom (2010)
36. Liu, C., Rani, P., Sarkar, N.: Affective state recognition and adaptation in human-robot interaction: A design approach. In: *International Conference on Intelligent Robots and Systems (IROS 2006)*, pp. 3099–3106. Beijing, China (2006)
37. Moshkina, L., Arkin, R.C.: Human perspective on affective robotic behavior: A longitudinal study. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, pp. 2443–2450 (2005)
38. Murphy, R.R., Riddle, D., Rasmussen, E.: 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 (2004)
39. Mutlu, B., Hodgins, J.K., Forlizzi, J.: 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 (2006)

40. Mutlu, B., Osman, S., Forlizzi, J., Hodgins, J.K., Kiesler, S.: 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, UK (2006)
41. Nomura, T., Suzuki, T., Kanda, T., Kato, K.: Altered attitudes of people toward robots: Investigation through the negative attitudes toward robots scale. In: Proc. AAAI-06 workshop on human implications of human-robot interaction, vol. 2006, pp. 29–35 (2006)
42. Olsen, D.R., Goodrich, M.A.: Metrics for evaluating human-robot interactions. In: Performance Metrics for Intelligent Systems Workshop (2003)
43. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(10), 1175–1191 (2001)
44. Preece, J., Rogers, Y., Sharp, H.: *Interaction Design-Beyond Human-Computer Interaction*, 2nd edition edn. John Wiley and Sons, Ltd., West Sussex, England (2007)
45. Rani, P., Sarkar, N., Smith, C.A., Kirby, L.D.: Anxiety detecting robotic system - towards implicit human-robot collaboration. *Robotica* **22**(1), 85–95 (2004)
46. Riddle, D.R., Murphy, R.R., Burke, J.L.: Robot-assisted medical reachback: using shared visual information. In: IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2005), pp. 635–642. IEEE, Nashville, TN (2005)
47. Schweigert, W.A.: *Research Methods & Statistics for Psychology*. Brooks/Cole Publishing Company, Pacific Grove, CA (1994)
48. Shaughnessy, J.J., Zechmeister, E.B.: *Research Methods in Psychology*. McGraw-Hill, Inc., New York (1994)
49. Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., Goodrich, M.: Common metrics for human-robot interaction. In: 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction. ACM Press, Salt Lake City, Utah, USA (2006)
50. Stevens, J.P.: *Intermediate Statistics A Modern Approach*, 2nd edn. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ (1999)
51. Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., Montanier, J.M., Oranç, C., Oudgenoeg-Paz, O., García, D.H., et al.: Second language tutoring using social robots: A large-scale study. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 497–505. IEEE (2019)
52. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology* **54**(6), 1063–1070 (1988)