Qualitative Interview Techniques for Human-Robot Interactions

Cindy L. Bethel, Jessie E. Cossitt, Zachary Henkel, Kenna Baugus

1 Introduction

A significant challenge associated with data collection for studies in human-robot interaction, especially with children, is determining the best methods for obtaining rich and meaningful data. In general, children under the age of 11 have a difficult time with being able to distinguish between different levels of information such as found in Likert scale-based surveys, understanding vague and abstract questions, and understanding partially labeled responses [3, 11]. As cognitive development increases after around age 11, then the use of survey data may be more accurate. However, the data from surveys and self-report questionnaires is limited in scope when compared with the richness of information obtained from structured and semi-structured interviews. This is typically true for different age groups including adolescents and adults. One approach that has been helpful is the use of the *forensic interview* approach to structured and semi-structured interviews [9]. The forensic interview approach is a structured protocol that has been established for obtaining critical information especially from children. It can also be useful for obtaining information from adolescents in addition to adults. The emphasis of the structured forensic interview is the use of open-ended prompts for information versus the use of focused recognition prompts that are frequently used in structured interviews [9]. The open-ended nature of the interview questions allows for more free responses that help to reduce the likelihood of the interviewer influencing the memory of the person being interviewed. It is important for accuracy of information gathering that leading questions or focused/closed-ended questions are not used as it could impact the person's memory of the event or experience.

The qualitative analysis of the data from using structured and semi-structured interviews can be tedious and challenging to perform, though the results can be

Cindy L. Bethel, Jessie E. Cossitt, Zachary Henkel, Kenna Baugus
Mississippi State University, Department of Computer Science and Engineering, P.O. Box 9637, Mississippi State, MS 39762-9637, e-mail: clb821, jec570, zmh68, kbb269@msstate.edu

beneficial to advancing our knowledge of interactions between humans and robots and other related technologies. This chapter begins with a brief overview of research related to the use of interviews for gathering information located in Section 2. In Section 3 there are details on how to perform studies using the forensic interview approach. The next section will cover the process of transcription and coding of the data, refer to Section 4. In Section 5, an example study is presented along with the transcription, coding, and analyses performed. Lastly, conclusions presented in Section 6.

2 Related work on qualitative interviews

This section includes some representative examples of studies that have used interviews for gathering information related to the topic of interest. It is not meant to be a comprehensive review of all possible studies, but rather to provide a foundation of examples related to the use of structured and semi-structured interviews in research.

A study conducted by da Silva et al. focused on doing a qualitative analysis of participants interactions with a robot as the robot conducted a motivational interview about physical activity [14]. This style of interview is a counseling approach that focused on interactive, two-way communication with the interviewees that leads them to make their own conclusions about the topic. A NAO robot was used to deliver the motivational interview using a script that was devised in such a way to provide a flow of conversation regardless of how questions were answered and was evaluated using Shingelton and Palfais criteria for assessing technology-driven adaptations of motivational interviewing. Participants who had expressed a desire to increase their level of activity engaged in the motivational interview with the robot where the robot asked them questions, and they responded before tapping the robots head to indicate that they had finished speaking. The interviews were not recorded due to concerns regarding participants levels of discomfort in the situation, and instead qualitative data was acquired in the form of typed answers to free response questions about the experience. The authors do express the desire to further their research in the future by recording the interviews after a short acclimation phase.

Participants responses were coded using *thematic content analysis* that defined a unit of the data set as all of the responses of a participant. The data set was analyzed to develop codes and create a list of themes and sub-themes. The data was coded by two coders working independently who then came together to discuss discrepancies and revise the coding scheme on agreement was achieved. Then two more coders with no knowledge of the study worked independently following the scheme to recode the data. Inter-rater reliability was then determined based on the percentage of agreement on coded data for each unit of the data set. The analysis showed that participants found the experience helpful and were especially fond of getting to hear themselves talk about their concerns while not having to fear being judged by a human interviewer [14].

Qualitative interviews were used to gather data about the effects of using a robotic seal as a therapeutic tool for the elderly in a study conducted by Birks et. al. [2]. The study took place at an eldercare facility in Australia where three recreational therapists, who were employees of the facility were trained to use Paro robotic seals as therapeutic tools and did so with residents of the facility over the duration of four months while keeping detailed journals of each session. After this time, the therapists were interviewed about their experiences using the seals as a therapeutic tool. Transcripts from the interviews with the therapists were created and provided verbatim responses for each interview. Following transcription, two researchers did a thematic analysis to code the collected data both from the interviews and the therapists session notes. A third researcher was used to check the finished coding. The thematic coding showed three themes in the responses that the authors list as "(1) a therapeutic tool that's not for everybody, (2) every interaction is powerful, and (3) keeping the momentum." The data was coded based on these three themes; however, some of the data was split into sub-themes. Results of the analysis showed that the robotic seals had high potential to be a useful therapeutic tool based on the experiences of these recreational therapists [2].

Jeong et. al. investigated the potential use of a social robot to offer therapeutic interactions with children in pediatric hospitals [7]. The project utilized a robotic teddy bear called Huggable and compared patients interactions with it to their interactions with a virtual teddy bear on a screen as well as with a normal plush teddy bear. Children who were patients at a pediatric hospital were given one of the three bears for a duration of four hours to interact with as they wished. During the sessions involving the Huggable robot and the virtual bear, the bears were operated remotely by a researcher outside the room who used the bears to communicate and play games with the children.

The interactions in each condition were video-recorded following an initial thirty minute introductory period. The video recordings were transcribed to include verbal data as well as data of physical movements. A transcription of the verbal data was created by a professional transcriber, and this transcription identified who was speaking for each utterance with possible speakers including "Patient, Huggable, Moderator, and Other." Physical movement was transcribed by the use of a number between 0 and 1 with 0 meaning no movement and 1 meaning fully active movement. The data that was coded from the transcriptions included average movement score, number of utterances, and length of interaction. Post study questionnaires were also given to the staff members who acted as moderators though it is not clearly stated what kind of questionnaires were given. Results showed that, of the three conditions, children were the most verbally and physically engaged with the Huggable robot. These results were partially corroborated by the moderators whose questionnaires reflected that they found the virtual bear and Huggable to be equal but both better than the standard plush bear [7].

In a study by A. M. Rosenthal-von der Pütten and N. C. Krämer, participants were interviewed about their opinions of pictures of robots in order to collect qualitative data about their attitudes towards robots that could be considered "uncanny" or very human-like without appearing fully human [12]. Both adults and children participated in the interviews where they were shown pictures and videos of robots with varying degrees of likeness to humans. Interviewers asked the participants questions based on an interview guide. The questions prompted participants to freely respond about their thoughts regarding robots in general as well as their emotions and anxiety levels when shown specific robots. Auditory recordings of the interviews were collected and then transcribed including pauses and excluding filler words. Responses were coded by two different individuals working independently using the software MAXQDA (`https://www.maxqda.com/`). Coding schemes for the data were determined based on responses to questions. For example, the response to a question regarding how a robot made a person feel would be coded as "positive," "negative," or "no response." The coded data was used to identify an extensive amount of information about perceptions of uncanny robots [12].

In an attempt to determine what causes breakdowns in interactions between children and technological tutoring systems, Serholt conducted a study where a robot tutor was placed in an elementary classroom to interact with students over the course of six months [13]. A NAO robot was used for the study as well as an interactive screen. The robot could carry out scripted interactions including talking and gesturing to the screen while instructing students on certain classroom tasks such as map reading and sustainability. Video recordings of the students' interactions with the robot were collected and later analyzed to determine when breakdowns occurred based on a list of breakdown indicators. Videos were always viewed at least twice for the sake of reliability and accuracy. Video segments that were determined to show breakdowns were coded based on the indicators that were shown as well as details of the interaction. Thematic analysis of the data showed that there were common themes causing breakdowns including "the robot's inability to evoke initial engagement and identify misunderstandings, confusing scaffolding, lack of consistency and fairness, and controller problems."

These studies provide some examples of different ways in which qualitative data may be collected and analyzed. It also provides some information regarding the richness of qualitative data and the possible benefits for using this approach. The use of interviews and open-ended questioning either through verbal interviews or from written responses, provides the ability to learn trends in data and responses that may not be evident from using surveys and self-report assessments. This approach allows users to make statements without being constrained by a set of survey questions for identifying their thoughts, feelings, and interactions. Overall, if performed correctly, there is less suggestibility and influence by the researchers and they may be able to obtain information that otherwise may not be possible. One caution, the Hawthorne Effect may still be an influence with this approach, in which participants may say what they feel the researcher wants to hear instead of what they are truly feeling [10]. Therefore, it is important when conducting a study that the researchers do not express any feedback on the responses given by participants or their performance during the study.

3 Approach

There are several approaches that could be used to perform a structured interview; however for the purpose of this chapter, the focus will be on the use of the forensic interview process and approach. This process was developed for use with children who had experienced maltreatment or were eyewitnesses to crimes to gather information for a legal case [9]. The forensic interview approach has been very effective especially when working with children. The use of open-ended questions allows the person being interviewed to rely on their memory of an experience rather than the recognition of options presented by the interviewer, which can be misleading. The forensic interview approach is beneficial in obtaining a person's feelings about an experience and less likely to introduce confounds into that process. This process encourages the interviewer to avoid yes/no questions and questions that provide limited responses, such as multiple choice. If a multiple choice question type is used then it is important to follow that up again with an open-ended question to gain additional knowledge. As part of this approach there is a protocol in place for the interviewer to follow. The protocol will be outlined in the following sections.

3.1 Introductory phase and guidelines

The first part of the forensic interview process is for the interviewer to introduce him or herself to the person who is being interviewed. Next it is important to share what the purpose of them participating in the study is and what tasks they are expected to do. Also the interviewer should tell the interviewee what his or her role is in the study and a bit about their background. Once that has been completed it is important to provide guidelines for the interview process. The following is an example of the items to include as part of the guidelines used in a forensic interview. It is important to provide these guidelines regardless of the age of the participant, but it is even more important when working with children.

Guidelines:

1. *If you do not know or remember the answer to a question, then just let me know.* With children it is recommended that you then practice that by giving an example, such as *"Tell me the name of my dog."* It is expected that the child would say that they did not know the name of your dog.
2. *If you do not understand the question, then ask me to provide the question in a different way to help you better understand.* This would be followed by an example such as *"Tell me your ocular color."* The child would likely say they did not understand. The interviewer would then change the question to be, *"Tell me your eye color."* The child should then be able to answer that question.
3. *If you need me to repeat a question or information, please ask me to repeat it.* This typically does not require an example, even when working with children.

4. *If I say something that is not correct, please let me know and give me the correct information.* As an example, you can tell a male child that is 10 years old, *"What if I said you are a 5 year old girl?"* The male child should correct you and tell you that he is not a 5 year old girl. Then provide the correct response.
5. *I ask that you please tell the truth as you recall it for any questions you are answering. Do you promise to tell the truth today during our time together?* It is important for whomever is being interviewed to agree to tell the truth throughout the process.

NOTE: it is important to audio/video record the interactions so that transcriptions can be accurately performed at a later date including the guideline process using proper informed consent as to how the data may be handled.

3.2 Rapport building

The next phase of the forensic interview process is considered *rapport building* with the person being interviewed. Most people will not be comfortable sharing information with someone they do not know and this is especially true with children. It is important to establish some level of rapport to overcome this. Rapport can be established by asking some general questions that are not necessarily critical information. Some examples of questions that can be asked are:

- *Tell me about who you live with.*
- *Tell me about the place in which you live.*
- *Tell me about what you like to do for fun.*

When asking questions using this approach it is important to actually make them statements instead of questions. Commonly, people will ask questions in the format of "Can you tell me about who you live with?" and that leads to a yes/no response typically. By making a statement, it requires the person to give more than just a one word response. This is one of the most challenging aspects to performing this type of interview for the interviewer.

3.3 Narrative Practice

The next phase of the forensic interview approach is called *narrative practice*. This involves asking the person being interviewed to discuss a topic in as much detail as they can remember. For example, it is common to ask them to tell you about everything they did that day from the moment they woke up until they arrived to meet with you. As part of this process, you may pick one aspect of their detailed account and request that they provide more details, such as *"tell me more about what you had for breakfast"* or *"tell me more about playing basketball with your* *friends today."* This gets the person used to discussing in detail about a topic area. Some additional possible follow on statements may be:

- *Tell me about how you were feeling when ... happened.*
- *Tell me about how that made you feel physically in your body.*
- *Tell me more about that.* - this is commonly used in the open-ended interview questions.
- *Tell me about what happened next.*
- *Tell me about the first time, the last time, or the most recent time that this happened.*
- *Tell me about how you learned to do ...*
- *Tell me about anything else you would like to share that happened.*

The important aspect of this phase is to get the person being interviewed used to sharing information in as much detail as possible on a topic that is not related to the exact topic of interest you are trying to investigate.

3.4 Sustantive disclosure interview phase

The substantive interview phase of the forensic interview protocol is the part in which questions of interest or investigation are discussed. The substantive interview is the key component of this qualitative interview approach. This is where you ask the person being interviewed about what is being studied. In the case of our examplar study presented in Section 5, the primary topic being investigated was children's experiences with bullying. These are sensitive topics that most people are not comfortable right away discussing with someone they do not know well. That is why following the forensic interview protocol is important to have a more comfortable interaction.

In the case of the exemplar study, participants were interviewed using the forensic interview protocol by either a human interviewer or a robot interviewer. In the case of the robot interviewer, following the forensic interview protocol allowed time for the person engaged with the robot to overcome the novelty effect [1]. The novelty effect is the initial excitement and positive interactions as a result of new technology. It may impact how a person views the technology and this may change over time when the newness of the interaction "wears off."

It is during the *substantive interview phase* that the interviewer asks the participant the key questions or statements of interest. Some recommended statements for eliciting information may be:

- *Tell me about how you felt emotionally ...*
- *Tell me about how you felt physically in your body when ...*
- *Tell me what symptoms you felt when you ...*
- *Tell me about what happened next.*

These are just some examples to guide the interview process. It is also possible to use the prompts and statements from the *narrative practice* phase of the forensic interview (refer to Section 3.3). The important aspect is to make sure that the statements or questions being asked are open-ended in nature to get the person talking in detail about the topic of interest. This may not seem natural to the interviewer at first, because so often in conversations there is a tendency to ask closed-ended questions or multiple choice questions. If that is needed to get people talking, then it is recommended to follow that type of question with an open-ended question or statement to obtain more details or information. This is where this approach holds the most value because it provides rich data and deeper insights than what can typically be obtained through a survey or self-assessment questionnaire. Also it is typically done at the time of interaction or immediately following an interaction so the information is fresh in the mind of the person being interviewed.

3.5 Cool-down/wrap-up phase

This is the last phase of the forensic interview protocol. In this phase it is important to return the person being interviewed to everyday life and normalcy. In this case, you may ask a statement such as *"Tell me about your plans for the rest of today."* This redirects the focus away from the topic of interest and gets them thinking about what is to come in their day and returns them to everyday life. At this point, for research purposes you can do a debrief of the study where you discuss the purpose of the study. You may also want to ask the participant to not share any information about the study with others so they can have a similar experience if they participate. This may also be a good time to have participants complete any relevant surveys on what is being investigated. It is always recommended that the interviewer thank the participant for his or her time and effort in being involved in the research. After this point, the researcher/interviewer can go through normal procedures for ending participation in the study as they would in any study such as paying out incentives or other closing activities.

4 Transcription and coding of data

Once all the data is collected for a study, then the next phase of a qualitative study is to begin transcription of the interview data. This is a tedious process of listening to the audio portion of the audio-video recording of the interviews and typing out a transcript or record of everything that was said during the interview. There is some disagreement as to whether filler words should be included in the transcription, such as "um," "ahhh,", and similar. Depending on what is being studied, a person who is more nervous or uncertain may have higher numbers of these filler words and that may be important if the investigation is surrounding for example interactions between a human and a robot. It is up to the discretion of the researcher whether these types of utterances are included; however it is important to be consistent in what is included within a particular study. Once all of the data is transcribed, and there is a complete written record of the interview then this data is coded and categorized in a manner that can be interpreted and compared. This section presents details on the transcription and then the coding process.

4.1 Transcription process

Once video of an interview has been collected, it is necessary to create a written record or transcription of all the auditory data. The transcription process must reliably produce an accurate representation of everything said in the interview usually including filler words like "um" and "ah." It is helpful to have more than one person working on transcribing the same data in order to ensure a level of accuracy higher than just one person's understanding of what was said.

Typically, the transcription process involves watching a small bit of video and pausing to denote what is said before moving on to another small bit of video or re-watching, if necessary to capture more details. The process can be quite tedious and time consuming, but there are some tools to help it go more smoothly. An example of one such tool is the use of foot controls. When using foot controls for transcription, one has the ability to control the video playback with a foot pedal keeping the hands free for typing a written record of the data. One example that our research team have used is the Infinity in-USB-1 USB Computer Transcription Foot Pedal `https://smile.amazon.com/gp/product/B008EA1K66/ref=ppx_yo_dt_b_asin_title_o03_s00?ie=UTF8&psc=1`.

Another useful tool is software designed for use in video transcription. One example of transcription software is ELAN, which is produced by The Language Archive. ELAN is a program that allows the user to add in-time annotations to video or audio files and also supports searching and exporting the transcribed data [5]. This software is open source and available for download at `https://tla.mpi.nl/tools/tla-tools/elan/`. While there is no method to effortlessly create an audio-video transcription, taking advantage of available tools greatly eases the process. There is no easy way to perform this process, which is why many researchers prefer quantitative data collection methods such as surveys because it is much easier and straightforward to process the data and quickly obtain results. Often researchers will report in publications that qualitative data was collected as part of the study, and that the analyses will be presented at a later date. Typically, those results often do not appear in publications. It takes time and personnel to process this type of data, but it is often worthwhile because deeper knowledge and insights can be obtained from open-ended interview data.

4.2 Coding the transcribed data

After video data has been transcribed, it must be coded or translated into a format so that the data can be analyzed. The first step to coding transcribed data is to decide on a system to use to create quantitative data that makes sense with the kind of responses that are being coded. This could take the form of rating responses using a Likert scale, determining whether responses display certain qualities using a binary system, deciding which of a set of categories is the best fit for responses, sorting responses by recurring themes found through thematic analysis [13, 2, 14], or any other similar scientific system that makes sense for the data being analyzed. Adequate training of whatever system is chosen must be provided to the people that will be serving as data coders. Training usually involves providing the coders with sufficient documented information of how the data should be coded as well as testing the coders with a set of dummy data to ensure that they are capable of performing the task before they begin working with the actual data from the study.

It is essential to ensure that the coding of transcribed data produces consistent and reliable quantitative data. Such consistency can be reached by establishing strong inter-rater reliability. This is accomplished by having multiple people involved in coding the same data. The process requires at least three coders but can include more for an even higher level of reliability. In the case of three coders, two of these coders work independently coding the same transcribed data before coming together and finding any discrepancies in their coding. The third coder determines how the data should be coded in the case of the discrepancies [16]. Additional coders can be used in the initial coding process to provide more viewpoints as well as in the tie breaking process to form a committee to objectively determine coding in the case of discrepancies.

Data coding is a time intensive process, and because of the necessity of establishing reliability, it can be very labor intensive as well. While these factors are certainly drawbacks of the transcription and coding process, they are worthwhile because of the robust data that is produced when taking the time to properly analyze the qualitative data.

In addition to coding and analyzing transcribed data, the behaviors of the person during an interview may also be coded. This involves a similar process of having a minimum of two coders categorizing behaviors and/or facial expressions during an interaction. This may be important in human-robot interaction studies. They may code body positions, postures, and facial expressions or other behaviors of interest. This data may then be quantified and compared across participants of the study to determine trends in behaviors within the study and during the interactions.

A common measure for determining the accuracy of coding is a statistical measure known as *Cohen's Kappa*. Cohen's kappa is used to measure the level of agreement between two coders of qualitative data, whether it is transcription data or behavioral data. Many statistical analysis software packages can be used to perform this statistical test and there are also online calculators available such as `https://idostatistics.com/cohen-kappa-free-calculator/`. The scale to evaluate Cohen's Kappa and reliability is:

- 0.01 0.20 slight agreement
- 0.21 0.40 fair agreement
- 0.41 0.60 moderate agreement
- 0.61 0.80 substantial agreement
- 0.81 1.00 almost perfect or perfect agreement

It is commonly agreed that a Cohen's Kappa score of 0.60 or greater indicates satisfactory reliability within the coding of a set of data [10]. It is important to perform this evaluation and report it as part of any study that involves qualitative data.

4.3 Coding written responses

Qualitative interview data can sometimes take the form of written responses to open-ended questions rather than transcribed audio or video data. In such situations, coding the data works essentially the same way that it would if it did come from a transcribed video, the main difference being that transcription is not necessary because the data is already in a text format. This allows studies that use written responses to collect qualitative data (see [4, 14]) to do so while skipping the transcription step. This can save time and equipment, but does not always provide data that is as robust as that gathered by transcribing audio or video. Participants may get tired of writing or may not want to write out a response often resulting in shorter responses.

5 Examplar Qualitative Interview Study

This section provides an exemplar study that used qualitative interviews for information gathering with children. This was part of a broader project investigating the use of social robots as intermediaries for collecting sensitive information from children. Our research team conducted two studies focused on conducting forensic interviews with children concerning their experiences with bullying at school.

Though a rich array of data was collected during these studies (e.g., behavioral measures, participant self-reports, parental surveys, etc.) this section focuses on the qualitative data captured in a verbal structured forensic interview with participants concerning their perceptions of the interviewer (human or robot) immediately following their session.

Detailed findings from this research effort are available in a journal article (see Henkel et al. [6]); however in order to provide an example for this chapter, the relevant details are presented of the research questions, study design, data collected, analysis approach, and findings in this section with an expanded focus on the qualitative data. Excerpts from that article are presented as part of this chapter [6].

5.1 Research questions

The exemplar research effort focused on the question of how a forensic interviewer's characteristics (i.e., robot or human, and male or female) would affect the likelihood of children disclosing sensitive information related to their experiences with bullying at school. In addition to assessing disclosure behavior across conditions, we were interested in gaining an understanding of how participants perceived the robot or human conducting the forensic interview.

Though our past experiences conducting studies with children using social robots in sensitive domains helped to inform this inquiry, the approach can be characterized as exploratory in nature. While it was hypothesized that differences would exist between interviewer conditions, no prior predictions about the areas or directions in which differences would be observed were made. Determining the trends and exact differences between interviewer types was an investigative process, that was well supported by the use of the open-ended forensic interview approach [9].

5.2 Study design

As the context surrounding data is critical to interpreting and making use of the data, this section describes the design of the two larger studies from which data regarding participant perceptions of the forensic interviewer were obtained. In *Study A* participants were between ages 8-12, while in *Study B* participants were between ages 12-17.

Both studies focused on using a structured forensic interview technique (as developed and investigated by Lamb et al. [9]) to obtain information about a participant's personal experiences with bullying at school (refer to Section 3. An interdisciplinary approach was taken to developing a dynamic script to guide the interview, with the research team's sociologist (experienced in this area of inquiry) ensuring full coverage of data typically collected during an investigation of bullying. This research was approved by and conducted with guidance from the Mississippi State University Institutional Review Board.

Both *Study A* and *Study B* followed the same base script addressing the areas of physical, relational, and verbal aggression, but *Study B* (older children) also included additional questions specifically addressing cyberbullying. Though following a pre-scripted structure during the interview, in all conditions a participant's responses determined the specific follow-up questions delivered by the interviewer. Additionally, information provided by the participant during the interview was incorporated into the follow-up prompts and interviewers responded to requests for clarifications or other spontaneous requests if needed.

Both studies used a between-participants design, employing random assignment to pair a participant and interviewer. In *Study A* possible interviewer assignments were: female human, male human, female humanoid RoboKind robot, or male humanoid RoboKind robot. In *Study B* the possible interviewers were: male human, male humanoid RoboKind robot, or male humanoid Nao robot.

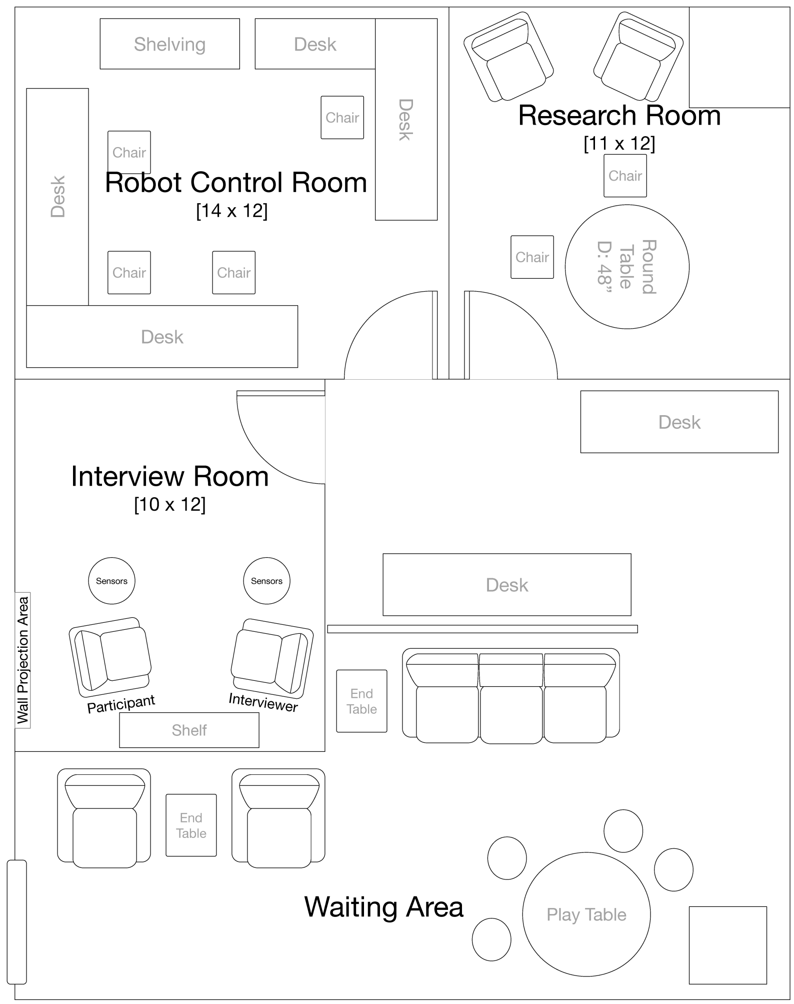

Fig. 1 Layout of the study rooms and space at the MSU Social Science Research Center

Study sessions were conducted in a dedicated lab space at the Social Science Research Center on the campus of Mississippi State University (refer to Figure 1). Two separate rooms were used for participant interactions, one for the researcher and participant to interact and the other for the participant and forensic interviewer to interact. All sessions (human and robot) were conducted using a Wizard-of-Oz [8] approach in which two hidden researchers worked collaboratively using camera feeds and a custom software system to remotely direct the flow of the interaction. These researchers entered spoken information into the software system, cued the questions an interviewer would deliver next, and ensured the interviewer remained engaged and responsive during the session. In the case of robot interviewers, the output of the software system directly activated the robot's speech and movement behaviors. Human interviewers delivered the same prompts, but were guided by a concealed teleprompter that projected prompts and cues on the wall directly behind the participant or onto a tablet the human interviewer held in his lap.

In total three robotic platforms were used between the two studies. In *Study A* male and female RoboKind R25 robots were used, while in *Study B* a male RoboKind R25 robot and blue colored humanoid Nao V5 robot with the same synthetic male voice were used (refer to Figure 2).

Fig. 2 Male Robokind R25 robot (left), Nao robot (center), and Female Robokind R25 robot (right)

5.2.1 Forensic interview about bullying experiences

The main portion of each study session was an interview conducted by either a robot or human. At the beginning of the interview the interviewer communicated the guidelines consistent with forensic interviews. The introduction script was as follows with excerpts from one participant:

INTERVIEWER: Hi <participant name>, it's nice to meet you. I'm the interviewer for the study today and would like to talk to you for a little while about your experiences at home and school. Does that sound okay to you?
PARTICIPANT: Yes [*nods*]
INTERVIEWER: Okay, great. If you can take a seat over there, we can get started. Once we are finished <researcher name> will come back and take you to the next part of the study.
INTERVIEWER: Like <researcher name> said before, my name is <interviewer name> and I spend a lot of time talking with kids about the different things they see and do at home and at school. Before we get started, do you have any questions for me?
PARTICIPANT: No.
INTERVIEWER: Before we begin, there are a few important things you should know. The first is if at anytime during the interview you are not sure about the answer to a question, just tell me that you don't know. Let's practice this. So if I asked you, what's my dog's name, what would you say?
PARTICIPANT: I dont know.
INTERVIEWER: If I ask a question during the interview that is confusing or that you dont understand, just tell me and I will explain what I mean. Let's practice this. So if I asked you, what is your ocular color, what would you say?
PARTICIPANT: I don't know what that means.
INTERVIEWER: It's okay to tell me you don't know what that means. So I'll ask it in another way. What color are your eyes?
PARTICIPANT: Blue.
INTERVIEWER: It's also important that if I say something wrong during the interview, you should tell me. For example, if I said you were a 4 year old boy, what would you say?
PARTICIPANT: I am not a 4 year old boy.
INTERVIEWER: And the last important thing is that you tell the truth during the interview. Do you promise you will tell the truth today?
PARTICIPANT: Yes, I promise.

Next, rapport with the participant was built via questions about the participant's hobbies, family, and events that occurred recently. This sequence also allowed the participant to practice providing a narrative description. An example exchange is as follows:

INTERVIEWER: So first, I'd like to get to know you better. What's something you like to do for fun?
PARTICIPANT: I like speech and debate. I'm in the speech and debate team at school.
INTERVIEWER: How did you learn to do that?

PARTICIPANT: Well this is my first year doing it. My best friend is on the team and she recruited me to do it because she thought I'd be good at it because I'm into politics. She asked me if I would join the team and then I did. My coaches have taught me a lot.

INTERVIEWER: How often do you get to do that?

PARTICIPANT: We have tournaments almost every weekend, and we have practice almost every day after school.

INTERVIEWER: Are there other people that you do that with?

PARTICIPANT: Yeah, we have a pretty big team.

INTERVIEWER: Thank you for telling me about that. I'd like to know about things that happen to you. Can you tell me everything that happened yesterday in as much detail as possible, from the time you woke up until you went to bed?

PARTICIPANT: I went to school, ate breakfast, went through my classes, and went home. I took a nap, then played a video game for a little while. I ate supper with my mom then went back to bed.

INTERVIEWER: Tell me more about playing a video game.

PARTICIPANT: I wasn't really doing much.

INTERVIEWER: Earlier you mentioned eating supper with your mom, tell me more about that.

PARTICIPANT: She had bought something for us to eat.

INTERVIEWER: Is there anything else you remember about yesterday?

PARTICIPANT: No.

Following the rapport-building and narrative practice phases, the interviewer began to explore the substantive issue of bullying experiences at school. After discussing how kids get along with each other at school the interviewer inquired about verbal aggression, relational aggression, cyber aggression (*Study B* only), and physical aggression. If the participant disclosed relevant aggression experiences at any stage, follow-up questions were asked to fully characterize their experience. A representative interaction is as follows:

INTERVIEWER: Now I'd like to learn a little bit more about you and your family. [*pause*] There are a lot of different types of families today. Tell me about your family and who lives with you.

PARTICIPANT: My parents are still married. They have been married for a really long time, and I have three other siblings. I have an older brother, older sister, and younger brother.

INTERVIEWER: Who is the person that spends the most time with you when you are at home?

PARTICIPANT: Probably my sister. Me and sister are best friends.

INTERVIEWER: How do you feel about that person?

PARTICIPANT: I love her. We act the exact same. I got really lucky that she was my sister.

INTERVIEWER: Now let's talk about your friends. Can you tell me about your closest friends at school?

PARTICIPANT: One of my closest friends, the one who recruited me to the debate team, we spend a lot of time together. Especially now that I got my license. We hang out all the time.

INTERVIEWER: If you were going to tell someone a secret, who would it be and why?

PARTICIPANT: Probably her because I know I can trust her with all of it, and she won't tell anyone else.

INTERVIEWER: I'd like to talk some about how the kids at your school get along with each other. Let's start with the way kids talk to each other. Please tell me about the different ways kids talk to each other at school.

PARTICIPANT: Well, it's really really clique-y at my school so there is the cheerleading clique and then there is like the baseball clique and just like different things like that but most people, they're pretty nice to you. A lot of people are pretty nice to you, to your face, but a lot of people talk when they're not with you. There are people who are a little bit rude to people who are different than them, which isn't fun to see but for the most part people are pretty nice face-to-face. The real problem is whenever they're not together.

INTERVIEWER: Sometimes kids say mean or unkind things to other kids to make them feel bad. Can you tell me about this happening at your school?

PARTICIPANT: Yeah, I've witnessed that a few times, especially with people who are different, like people with disabilities and stuff. I've witnessed that a lot, which is terrible to see. I have a friend who has Asperger's Syndrome, so obviously he functions a little bit differently, but he's still really really nice, but people just treat him so differently and kinda talk down on him and make fun of him when he does stuff that they're just not really used to, which isn't fair. There are some just rude things that–especially whenever people don't agree with each other about things, you shouldn't be mean about but a lot of people, if someone is different they automatically criticize.

INTERVIEWER: Do the kids who say mean things to other kids do this a lot?

PARTICIPANT: Yeah

INTERVIEWER: How often do they do this?

PARTICIPANT: I have witnessed it. Not so much about the disabilities thing but whenever people disagree about pretty much about anything; it happens almost everyday. Especially on social media, it happens so much, within the school. Then making fun of people that are different like people with disabilities and stuff, I witness that maybe like twice a week or three times a week.

INTERVIEWER: Which kids do they normally say mean things to?

PARTICIPANT: Recently, surprisingly, you wouldn't think this would be going on today in 2017, but recently it's been a lot of like political issues, which is crazy. Our volleyball team knelt during the national anthem so it made a big deal in the school, so there were people who agreed with it and people who disagreed with it, and all day people were just criticizing each other about it literally all day. High school students today have a really hard time accepting that someone else can believe something different than they do, and if someone thinks differently it is so quick and easy to criticize than it is to listen, so that happens a lot.

INTERVIEWER: How often do these kids say mean things to you?

PARTICIPANT: I am pretty confident in my opinions, if I think someone is doing something that I don't think is okay, I'll be like, "hey that's not okay," and I get criticized about that a good deal.

INTERVIEWER: What happens when they say mean things to <u>other kids</u>?

PARTICIPANT: A lot of times people don't stand up and say anything, and sometimes people will agree and say the same thing. But sometimes students will say, "hey that's not okay." But it's a lot easier to just go with the crowd and just laugh about it.

INTERVIEWER: How do you feel when they say mean things to <u>others</u>?

PARTICIPANT: It really sickens me to see it, because it's just so unfair I think.

INTERVIEWER: Sometimes people talk about other people and say mean or untrue things about them. For example, kids might spread rumors, gossip, or tell lies about someone to hurt them. Can you think of examples of this happening at your school?

PARTICIPANT: Oh yes. That happens all the time. I think that's the primary issue at my high school: gossiping and rumors being spread all the time.

INTERVIEWER: Let's think about the different groups of kids at your school. Sometimes kids will leave others out of the group or ignore them on purpose. Can you tell me about this happening at your school?

PARTICIPANT: Yeah, I think that does happen a lot.

INTERVIEWER: Are there kids who do this kind of thing a lot?

PARTICIPANT: Probably everyday. You see it everyday at lunch.

INTERVIEWER: Which kids do they normally leave out?

PARTICIPANT: The people who are not so popular, not dressed with the trends or whatever.

INTERVIEWER: Are there kids at your school who leave you out?

PARTICIPANT: Yeah, there has been friend groups that I was part of that I necessarily wouldn't always like get invited to things.

INTERVIEWER: What happens when they do this to <u>you</u>?

PARTICIPANT: It does kinda hurt cause it's like, "Why wasn't I good enough?" but most of the time I get over it pretty quickly and go talk to someone else.

INTERVIEWER: How do you feel when kids leave <u>others</u> out?

PARTICIPANT: It makes me feel bad, and sometimes I'll be like, "Do you want to come sit with us?" but sometimes I don't, because [*pause*] it's – it's a lot easier to not say anything. It is sad to admit, but sometimes it's just easier to say, "Well that really sucks," and kind of pity them but not do anything about it.

INTERVIEWER: Is there anything else you would like to share about that?

PARTICIPANT: I don't think so.

At the end of each aggression section, participants responding with relevant experiences were prompted to characterize the aggressors to identify the power dynamic between aggressor and victim. The interview's final phases focused on the participant's definition of bullying and closing the interview by thanking the participant and discussing any fun things they were planning to do in the near future.

5.2.2 Perceptions of the interviewer interview (PII)

After completing the forensic interview, a researcher led the participant to a separate room and verbally administered a semi-structured interview about their perceptions of the human or robot performing the forensic interview. The *Perceptions of the Interviewer Interview (PII)* was a set of open-ended questions developed and refined over the course of six different studies involving robots and humans interviewing children about sensitive topics. The interview was structured as follows:

General Perception Questions

- **Q1:** What did you think about *interviewer name* during the study?
- **Q2 - Robot only:** Do you think *interviewer name* was aware of what was going on around *her/him*? Why or why not?

Understanding, Feelings, and Advice Questions

- **Q3:** How well do you think *interviewer name* understood what you said?
- **Q4:** How well do you think *interviewer name* understood how you felt?
- **Q5:** Do you think *interviewer name* could give you helpful advice if you had a problem? Why or why not?
- **Q6:** Sometimes people hide how they feel from others. Do you think you could hide how you feel from *interviewer name*? Why or why not?
- **Q7:** Are there things you could talk to *interviewer name* about that you could not talk to other people about? What kind of things could you talk about with *interviewer name*?
- **Q8 - Robot only:** How is *interviewer name* different from a human?

Helpfulness Questions

- **Q9:** Was *interviewer name* helpful to you?
- **Q10:** In what ways did you feel like *interviewer name* was helpful to you?
- **Q11:** How could *interviewer name* be more helpful to you?

Likability and Social Norms Questions

- **Q12:** Did you like *interviewer name*? Why or why not?
- **Q13:** Do you think *interviewer name* liked you? Why or why not?
- **Q14:** What would you do if *interviewer name* did not listen to you while you were trying to talk to *her/him*?

- **Q15:** What would happen if you did not listen to *interviewer name* while he was trying to talk to you?

Other Comments

- **Q16:** Do you have any other thoughts about *interviewer name* from the study that you'd like to share with us?

5.3 Study protocol

In both studies each session was an hour long and divided into four segments: (1) pre-interview tasks, (2) forensic interview about bullying experiences, (3) post-interview tasks, and (4) character guessing game with a robot. Prior to the interview, a researcher explained the study, obtained informed consent and participant assent, and administered a paper-based demographics survey in the "research room." After completing the demographics survey, the participant was guided to a separate "interview room" and spent about 30 minutes engaged in the forensic interview about bullying experiences. For children especially it is important to keep interviews short in length and should not go beyond 30 minutes. Once the forensic interview was complete, the researcher administered the PII in the "research room" and then the participants also completed a paper-based interviewer perception survey. Once that was completed participants were returned to the "interview room" and played a character guessing game with the robot, so that participants who did not get to experience the interview condition with the robot would have the opportunity to interact with the robot. After the character guessing game, participants were compensated for their involvement in the study, in this case they received a payment of $10 and a small gift.

5.4 Participants

Participants were recruited from a database of local children and parents that had expressed an interest in participating in research studies. The database is maintained by university researchers and is advertised through fliers, newspaper advertisements, and targeted advertisements on popular social media networks. Researchers used the database to contact parents with children that were eligible for each study. Participants who took part in the first study were ineligible for participation in the second study. Between the two studies, 70 female participants and 71 male participants were interviewed; 75 interviews were conducted by robots, and 67 interviews were conducted by humans.

5.5 Data and Analysis

A total of 147 one-hour sessions were conducted during the summer and fall of 2017, yielding 142 usable cases. Participants in *Study A* (younger children) were distributed between conditions to balance the interviewer-participant gender pairings. In *Study B* (older children), participants were randomly assigned to one of three interviewers [6].

The same male and female human interviewers were used across all human condition sessions in *Study A* and were social science majors in their early twenties. In *Study B* human interviews were conducted by a male undergraduate in the last year of his social science program and in his early twenties.

In this section, the analysis and results are presented from the *Perceptions of the Interviewer Interview (PII)* conducted by the researcher after the participant completed their forensic interview interaction concerning their experiences with bullying. Verbal and behavioral data captured during the main interview is currently undergoing analysis and will be reported in the future. The analysis in this section of the chapter examines effects present when responses from *Study A* and *Study B* were pooled as well as when they were analyzed independently.

5.5.1 Transcription and data coding approach

Verbal answers provided by participants during the *Perceptions of the Interviewer Interview (PII)* were converted to text by two independent research assistants using audio recordings of each session and the ELAN software package [5]. A third research assistant examined and resolved any discrepancies between the transcriptions, yielding a final text transcript for each participant. If responses were unable to be determined from audio recordings alone, video of the session was consulted to clarify verbal responses and capture non-verbal responses.

Two researchers independently evaluated text transcriptions of participant responses to items Q2 - Q13 from Section 5.2.2, coding responses first for agreement or appraisal (depending upon the question) and then for any social factors present in the response.

A five-point coding scheme for indicating agreement or disagreement was developed for items Q2, Q5, Q6, Q7, Q9, Q12, Q13 using the following coding guidelines, see Section 5.2.2 [6]:

- **No** - A verbal or non-verbal response indicating complete disagreement.
- **Indecisive negative** - A verbal response that primarily indicated disagreement but also included reservations, conditions, minor uncertainty, or hypothetical alternatives.
- **Indecisive** - A non-verbal or verbal response that ultimately indicated uncertainty.

- **Indecisive positive** - A verbal response that primarily indicated agreement but also included reservations, conditions, minor uncertainty, or hypothetical alternatives.
- **Yes** - A verbal or non-verbal response indicating complete agreement.

Similarly, a five point coding scheme for appraising performance was developed for items Q3 and Q4 with the following coding guidelines, refer to Section 5.2.2 [6]:

- **Very poor** - A verbal response that indicated exceptionally poor performance.
- **Poor** - A verbal response that indicated performance that was slightly problematic or did not fully meet expectations.
- **Indecisive** - A verbal or non-verbal response that ultimately indicated uncertainty.
- **Well** - A verbal response that primarily indicated performance was acceptable or met expectations.
- **Very well** - A verbal response that indicated superb performance or exceeding expectations.

In addition to the five established agreement and appraisal codes, a not applicable (NA) category was created for cases when a participant was not asked or did not provide an answer to a specific item. For the analysis presented in this chapter the agreement and appraisal scales were collapsed from five points to three points by combining the first two and last two categories on each scale [6].

Each response was also examined for supporting social factors that participants used to justify their answers. Items Q2, Q5-Q7 and Q9-Q13 included explicit follow-up prompts, which often elicited social factors, while responses to Q8 were primarily composed of social factors. Two researchers collaboratively generated a list of six base social factors from their observations of study sessions and by examining a small sample of transcribed responses to each item. During the coding process researchers discussed and created additional sub-categories within these six main factors when doing so assisted in more precisely characterized responses. Responses were coded for the following main social factors [6]:

- **Appearance:** Responses that referenced the interviewer's physical appearance but did not incorporate elements of the interviewer's behavior. (Positive Example: *"yeah because I liked her blue hair"*, Negative Example: *"He doesn't have life in his, in his eyes so you don't feel like you're talking to somebody alive"*.)
- **Demeanor:** Responses that referenced the interviewer's behavior or personality, often as it related to being understanding, interested, concerned, or helpful. (Positive Example: *"she was really nice and, uh she would listen a lot more than some other people would to me"*, Negative Example: *"I don't think she could see me uh really feel"*.)

- **Interviewer behavior:** Responses that highlighted a specific action the interviewer took during the interaction with the participant. (Example: *"I could tell he was listening at least most of the time because he like when I would say something he would ask like questions like specifically about that and stuff like that"*.)
- **Knowledge:** Responses that referenced the interviewer's knowledge. (Positive Example: *"he seemed pretty smart"*, Negative Example: *"I don t think she understood it at all because like I said before she was most likely programmed to do that"*.)
- **Social confidence:** Responses that discussed social judgment, privacy, or trust as it related to the interviewer.(Positive Example: *"I feel probably better talking to him than a person because I mean like when you're saying it to the person it's basically just like a guilt trip right there"*, Negative Example: *"it felt good to talk about it and at the same time I had to stop myself because it's a robot and I don't know where it was going to, so I just kinda like stop myself"*.)
- **Non-specific:** Responses that expressed only uncertainty, were not specific, referenced intuition, or emphasized the participant's own traits rather than evaluating the interviewer's traits or behaviors.

Each response was coded for social factors independently and any conflicts were resolved through discussion. A majority of conflicts in coding involved one coder selecting a less specific area of the same main category or one coder applying a single code when multiple were merited. After all coding conflicts were resolved, a final review of the responses associated with each social factor was conducted to ensure consistency.

5.5.2 Agreement and appraisal

After the initial data coding process, responses to items, which included an agreement or appraisal prompt were sorted into three categories (positive, negative, or indecisive) for analysis. For agreement prompts *Yes* and *Indecisive Positive* were grouped into the *Positive* category and *No* and *Indecisive Negative* were placed in the *Negative*. For appraisal prompts *Very Poor* and *Poor* were grouped as *Negative* while *Very Well* and *Well* formed the *Positive* group. In a small number of cases a researcher inadvertently skipped an item, a participant offered no decipherable response, or a technical error prevented capturing the participant's response and the participant was excluded from the analysis for the affected items [6].

The responses from both studies were merged and the frequencies of coded responses for each item were compared when participants were grouped by study, interviewer type (human or robot), and participant gender. Further analyses were conducted within the context of each study to better understand the source of any significant differences. No statistically significant differences on any agreement or appraisal items were identified when splitting responses into two groups based on the study they participated in (younger children versus older children) [6].

Robot vs. human interviewers
When responses were split into those with a human interviewer (male or female) and those with a robot interviewer (female R25, male R25, or male Nao), statistically significant differences were observed for items related to how well the interviewer understood the participant and the perceived ability of the interviewer to give helpful advice (Q3-Q5).

Participants with a human interviewer were more likely to respond that they were uncertain how well the interviewer understood what they said (Q3) (14.93%) in comparison to those with a robot interviewer (4.23%, Fisher's Exact Test $p = 0.01$). Furthermore, 7.04% of participants with robot interviewers felt the interviewer did not understand what they said, while none of the participants in human interviewer conditions reported a lack of understanding (refer to Figure 3) [6].

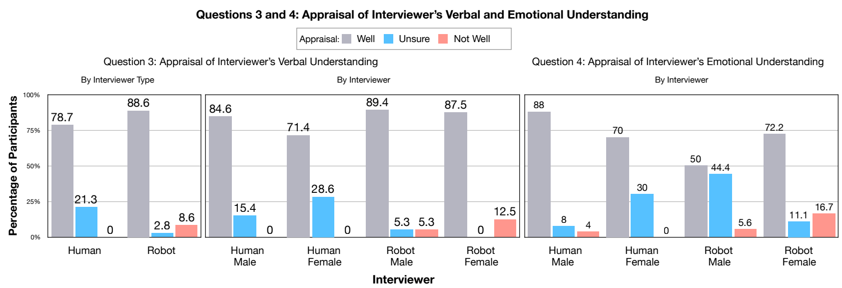

Fig. 3 Participant responses to questions about the interviewer's ability to understand what they said and how they felt.

When asked how well the interviewer understood the way they felt (Q4) (see Figure 3), participants in the robot conditions were more likely to indicate *Negative* (14.29%) or *Indecisive* (22.86%) responses compared with the responses in the human interviewer conditions with 1.54% indicating *Negative* and 13.85% indicating *Indecisive*, Fisher's Exact Test $p < 0.01$).

Participants in robot interviewer conditions were more likely (5.33%) than those with a human interviewer (1.54%) to indicate that they were uncertain if the interviewer could provide them helpful advice if they had a problem (Fisher's Exact Test $p = 0.01$). Furthermore, those in the robot conditions were more likely report that the interviewer would be unable to provide helpful advice (9.33%) in comparison to those in the human condition (0%).

An examination of *Study A* independently shows the only statistically significant difference (Fisher's Exact Test $p < 0.01$) that occurred between the human and robot interviewers was in responses to Q3 (see Figure 3). In the human conditions (male and female interviewers) 78.72% participants appraised the interviewer's ability to understand what they said as *Positive*, while 21.28% were *Indecisive*. In comparison, 88.58% of participants in robot interviewer conditions (male and female) appraised the interviewer's ability to understand what they said as *Positive*, 2.86% were *Indecisive* and 8.57% reported the interviewer did not understand well [6].

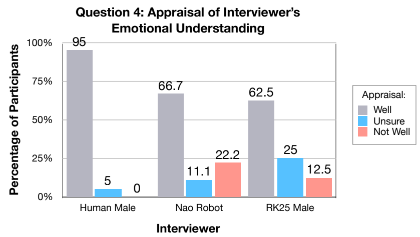

Fig. 4 Older participant's assessment of how well the interviewer understood their feelings, divided by interviewer.

Figure 4 illustrates that when analyzed independently, responses from *Study B* yield a statistically significant difference between participants in the different interviewer conditions (Human male, Nao male, R25 male) on Q4 (Fisher's Exact Test, $p = 0.05$). In the male Nao robot condition 66.67% of participants reported that the interviewer understood how they felt, while 22.22% indicated the robot did not have a good understanding of how they felt. Of the participants in the male R25 condition 62.5% indicated the robot understood how they felt, while 12.5% responded that the robot did not understand how they felt. In the human interviewer condition 95% of participants felt the interviewer understood how they felt [6].

Participant gender

When responses were split into groups based on participant gender (male or female) statistically significant differences were present for items concerning interviewer understanding of feelings and ability to provide advice (Q4,Q5), refer to Figure 5. As displayed in Figure 5, Participants identifying as female were more likely (26.47%) to report uncertainty when asked how well the interviewer understood how they felt in comparison to male participants (10.45%, Fishers Exact Test $p = 0.01$), who were more likely to endorse the *Negative* option (13.43%) in comparison to females (2.94%). Furthermore, as shown in Figure 5, participants identifying as female were more likely (97.18%) to indicate the interviewer could provide them helpful advice in comparison to participants identifying as male (85.51%, Fisher's Exact Test $p = 0.04$).

Figure 5 illustrates significant differences found within each study. When examined separately, responses from *Study A* show significant differences between participant reported genders and appraisals of the interviewer for Q6 and Q13. Compared to 35.71% of male participants, 66.67% of female participants reported that they could not hide how they felt from the interviewer (Fisher's Exact Test $p = 0.02$) [6].

Furthermore, within *Study A* female participants (64.29%) were more likely than male participants (50%) to perceive that the interviewer liked them (Fisher's Exact Test $p = 0.02$). While 16.67% of male participants responded that they felt the in-

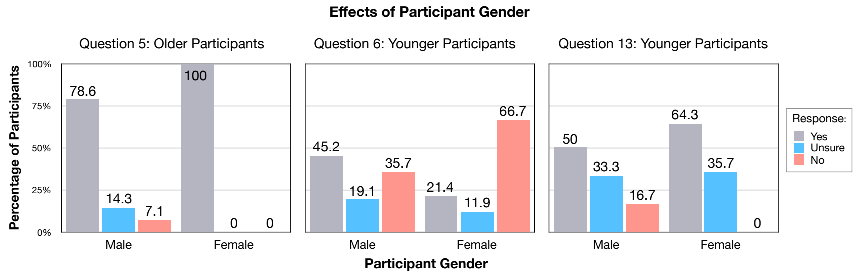

Fig. 5 Gender differences among participants in each study across all types of interviewers.

terviewer did not like them, none of the female participants reported the perception that the interviewer did not like them [6].

All female participants (100%) in *Study B* reported that the interviewer could provide helpful advice, while 78.57% of male participants indicated the interviewer could provide helpful advice (Fisher's Exact Test $p = 0.02$) [6].

5.5.3 Social factor mentions

Following the analysis of agreement and appraisal responses, an examination of the explanations that participants provided for their responses was conducted. Each response was tagged with all relevant social factors (described in Section 5.5.1). Data from *Study A* and *Study B* were combined and analyzed as a whole [6].

For each participant it was computed whether or not the participant made mention of each social factor across their entire response to the *Perceptions of the Interviewer Interview (PII)*. This was done to limit the influence of participants who cited the same factors repeatedly for multiple items. As each factor could have a positive or negative valence, this resulted in a set of twelve binary variables for each participant indicating whether or not the participant discussed the factor. For the analysis associated with this chapter, the sub-categories for each factor were not examined, rather they were counted as representing their higher level category [6].

Figure 6 compares the percentage of participants in human and robot interviewer conditions citing each social factor. With the exception of the *Knowledge+* social factor, all other positive social factors were referenced significantly more by participants in a robot interviewer condition [6].

Since participants in robot interviewer conditions had the opportunity to respond to two additional questions (Q2 and Q8), an analysis was conducted in which all responses were removed from these questions. Removing these items resulted in a loss of statistical significance between human and robot interviewers for the area of *Social Confidence*, which includes components of general trust, social judgment, and maintaining privacy ($\chi^2(1) = 2.67, p = 0.10$, Cramer's $V = 0.19$, small effect). When incorporating all questions, 41.33% of participants in the robot conditions

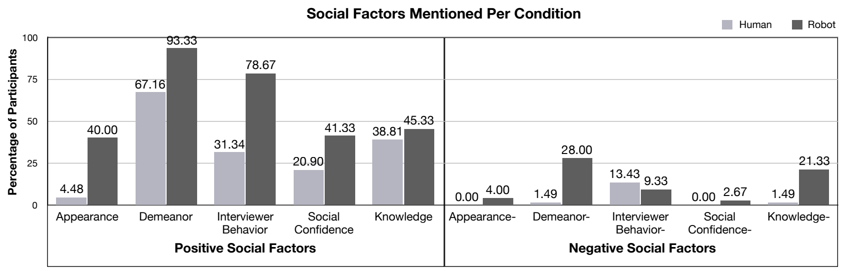

Fig. 6 Positive and negative mentions of each social factor for human and robot interviewers.

discussed factors related to *Social Confidence*, but when excluding responses to the question *How is the interviewer different from a human?* (Q8) this declined to 34.67% of participants in the robot interviewer conditions. Of the participants in the human interviewer conditions, 20.9% identified positive factors related *Social Confidence* [6].

5.6 *Reporting*

[6] The exemplar study presented in Section 5 provides different ways in which qualitative data can be analyzed, interpreted, and reported. There are different types of statistical tests required depending on the type of data being evaluated and how it is converted from textual transcriptions, coded, and evaluated as a quantitative value. In most cases the data was evaluated using non-parametric statistical measures such as a Chi-square test using Cramer's V for determining effect size (refer to `http://vassarstats.net/newcs.html`) or a Fisher's Exact Test (see `https://www.socscistatistics.com/tests/fisher/default2.aspx`), which is used when sample sizes are smaller than what is needed to calculate a Chi-square. It is recommended that a statistical methods book (e.g., [15]) be consulted or software packages used such as SPSS, SAS, or similar be used for a better understanding of these techniques. A detailed discussion of these statistical tests are beyond the scope of this chapter.

6 Conclusions

This chapter presents the basics associated with using a forensic structured interview approach for gathering richer data from human-robot interaction studies. Although survey and self-assessment data is a place to start and is useful in human-robot interaction studies, the content is limited. Study participants are forced to provide information associated only with what information is requested based on the items selected by the researcher. This type of data can only provide a certain amount of information about interactions between robots and humans.

In order to obtain additional information from participants of user studies, it is recommended that researchers enhance studies through the use of structured or semi-structured interview questions. This allows participants of the study to provide their own insights and feelings regarding the interactions and gives an opportunity for them to provide additional knowledge and insight. This type of data can be challenging, time intensive, and tedious to evaluate; however, it can provide additional insights that may not be discerned using other methods of evaluation.

As discussed in the Related Work Section 2, there are many different types of interview techniques that can be used. The focus of this chapter was on the use of an investigative interview technique, known as the forensic interview [9]. This approach has been successfully used in different types of investigations and it is especially effective for use in studies involving children. The approach for using the forensic interview protocol were presented in Section 3 and included the introduction, guidelines, rapport building, narrative practice, substantive disclosure, and cool-down/wrap-up techniques. A discussion on the process of transcription and coding of the qualitative data was presented in Section 4. An exemplar study was presented to help reinforce and demonstrate the process of how to conduct a forensic interview, and then examples of the how the transcription, coding, and analyses process was performed for this type of study. Although the process of using structured interviews with open-ended types of questions seems to incur significant effort, the results may transform human-robot interaction research and researchers are encouraged to consider using this approach in the design of their studies.

References

1. 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)
2. Birks, M., Bodak, M., Barlas, J., Harwood, J., Pether, M.: Robotic seals as therapeutic tools in an aged care facility. Journal of Aging Research **2016** (2016)
3. Borgers, N., Hox, J., Sikkel, D.: Response quality in survey research with children and adolescents: The effect of labeled response options and vague quantifiers. International Journal of Public Opinion Research **15**(1), 83–94 (2003)
4. Foster, R.K.: An investigation of training, schemas, and false recall of diagnostic features. Masters thesis, Mississippi State University (2015)
5. Hellwig, B., Van Uytvanck, D., Hulsbosch, M., Somasundaram, A., Tacchetti, M., Geerts, J.: ELAN - Linguistic Annotator. The Language Archive, MPI for Psycholinguistics, Nijmeg, The Netherlands, 5 edn. (2018)
6. Henkel, Z., Baugus, K., Bethel, C.L., May, D.C.: User expectations of privacy in robot assisted therapy. Paladyn, Journal of Behavioral Robotics **10**(1), 140–159 (2019)
7. Jeong, S., Breazeal, C., Logan, D., Weinstock, P.: Huggable: Impact of embodiment on promoting verbal and physical engagement for young pediatric inpatients. In: 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 121–126 (2017)

8. Kelley, J.F.: An iterative design methodology for user-friendly natural language office information applications. ACM Transactions on Information Systems (TOIS) **2**(1), 26–41 (1984)
9. Lamb, M.E., Orbach, Y., Hershkowitz, I., Esplin, P.W., Horowitz, D.: A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the nichd investigative interview protocol. Child abuse & neglect **31**(11-12), 1201–1231 (2007)
10. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human-Computer Interaction. John Wiley and Sons, Ltd., West Sussex, United Kingdom (2010)
11. de Leeuw, E.D.: Improving data quality when surveying children and adolescents: Cognitive and social development and its role in questionnaire construction and pretesting. In: Report prepared for the Annual Meeting of the Academy of Finland: Research programs public health challenges and health and welfare of children and young people, pp. 1–50 (2011)
12. Rosenthal-vonderPütten, A.M., Krämer, N.C.: Individuals evaluations of and attitudes towards potentially uncanny robots. International Journal of Social Robotics **7**(5), 799–824 (2016)
13. Serholt, S.: Breakdowns in childrens interactions with a robotic tutor: A longitudinal study. Computers in Human Behavior **81**, 250–264 (2018)
14. da Silva, J.G.G., Kavanagh, D.J., Belpaeme, T., Taylor, L., Bleeson, K., Andrade, J.: Experiences of a motivational interview delivered by a robot: qualitative study. Journal of Medical Internet Research **20**(5) (2018)
15. Stevens, J.P.: Intermediate Statistics A Modern Approach, 2nd edn. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ (1999)
16. Syed, M., Nelson, S.C.: Guidelines for establishing reliability when coding narrative data. Emerging Adulthood **3**(6), 375–387 (2015)