The Challenges of Evaluating Child-Robot Interaction with Questionnaires

Mike Ligthart  
Delft University of Technology  
m.e.u.ligthart@tudelft.nl

Rifca Peters  
Delft University of Technology  
r.m.peters@tudelft.nl

ABSTRACT
In this paper we reflect on the use of questionnaires as an evaluation tool in child-robot interaction research. We provide a case study containing eight user studies. While doing these user studies we ran into two major challenges: violations of the constructs used in questionnaires and a ceiling effect in the responses of the children. These issues are caused by a combination of factors such as, but not limited to, misinterpretations of questions, response biases, and the novelty effect. A first lesson learned is that a proper design of a questionnaire, and how questions are asked and answered, is essential. In this paper we discuss two questionnaire methods we have been developing that potentially could circumvent some of the issues. A second lesson learned is that user studies could benefit if they reflect the long-term nature of the child-robot interaction.

KEYWORDS
Child-Robot Interaction, Questionnaire Based Research, User Studies, Social Robotics

1 INTRODUCTION
Off-the-shelf social robots are used more and more in real-world applications for children. Healthcare in particular is a domain with many developments [1].

We are currently involved in two Child-Robot Interaction (CRI) healthcare projects for which we have done several user studies with children. In CRI research the use of questionnaires is a common tool for evaluation [2]. We use it in all our user studies. However, as we experienced, questionnaires are not always reliable. In this case study we will discuss two challenges we experienced while using questionnaires to evaluate a child-robot interaction. The first challenge, discussed in section 4, is a violation of the validity of the constructs used in questionnaires. The second challenge, a ceiling effect in the questionnaire responses of children, is discussed in section 5. For both challenges we will discuss the issues we observed and analyze the root-cause. Finally, we will discuss the lessons learned in section 6.

Rather than describing one study we chose to combine eight different user studies in one case study. The similarities and differences between the studies help us to better analyze the issues.

2 CASE STUDY CONTEXT
In this paper we focus on user studies where social robots were evaluated that provide affective and/or cognitive support. In Table 1 an overview is given of every included study. In this paper we limit ourselves to information relevant for discussing the challenges of using questionnaires to evaluate child-robot interactions.

The authors contributed equally and are listed alphabetically.

2.1 Application domains
The first application domain is diabetes care. Children diagnosed with type 1 diabetes mellitus need to develop new habits to improve their quality of life and avoid severe consequences. In the PAL project mHealth technology is developed to facilitate gradual increase self-management abilities and responsibilities. Intended robot functionalities are collaborating or competing in games, reviewing daily activities and self-management progress.

The second domain is pediatric oncology care (KWF-STW project). The focus lies on developing a social robot that can contribute to reducing stress and anxiety for pediatric oncology patients during their stay in the hospital. Intended robot functionalities are interactive affective storytelling, talking about the day with the robot, playing games and sleep coaching.

2.2 Aim of the studies
As most social CRI applications, both PAL and the KWF-STW project aim to affect the child’s mental state (e.g., mood) and/or behavior (e.g., develop habits) by exhibiting certain robot behaviors. The robot behaviors are designed to provide personalized and adaptive support.

The aim of all our described studies (see Table 1) is to evaluate whether the designed robot behaviors indeed have the desired effect on the child’s mental state or behavior.

From the PAL project the following evaluations are included:

A: repeated evaluation of child’s experiences during child-robot game play.
B: a comparison between a robot that gives direct instructions and asks knowledge questions afterwards and a robot that asks open-questions to engage the child in critical thinking.
C: usability and user experience evaluation of a digital diabetes diary where children interacted with a social robot and its virtual avatar.
D, E: measuring the children’s perception of robots that show various degrees of warmth, competence, and/or dominance during a presentation.

Furthermore from the KWF-STW project three pilot studies, that have been performed with school children to evaluate early prototypes, are included:

F: a comparison between a robot that likes explicitly what you like during a first contact and a robot that has a more nuanced response style.
G: a comparison between a plain and an interactive storytelling robot.

1Personal Assistant for a healthy Lifestyle: http://www.pal4u.eu/
2Improving Childhood Cancer Care when Parents Cannot be There – Reducing Medical Traumatic Stress in Childhood Cancer Patients by Bonding with a Robot Companion
Table 1: This table provides an overview of the studies that are included in this paper and lists the important characteristic of these studies. 'Duration' stand for minutes of interaction per encounter, 'frequency' stands for the number of interaction per day, and 'days' is the length of the study in number of days. In the last two columns we indicate whether the ceiling effect (CE) and/or the construct validity violations (CVV) hindered the research (where 'y' means yes, 'n' means no and '?' means we cannot know for sure).

<table>
<thead>
<tr>
<th>ID</th>
<th>Study Design</th>
<th>Duration</th>
<th>Frequency</th>
<th>Days</th>
<th>Instrument</th>
<th>Measure</th>
<th>CE</th>
<th>CVV</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>PAL Camp [9] explorative 5-15 min 1-4 4 questionnaire User experience</td>
<td>y</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Robot-tutor PAL camp 2 (role), within-subject 10 min 1 1 PuppyIr tool-kit [11] User experience</td>
<td>n</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>MyPAL [7] explorative 1-20 min 0.5 21 OBOBiDi Perception</td>
<td>y</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Perception study schools [10] 2 (warmth) * 2 (competence), between-subject 10 min 1 1 adjective stickers[10] Perception</td>
<td>y</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Perception study PAL camp [10] 2 (style), between-subject 10 min 1 1 adjective stickers[10] Perception</td>
<td>n</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>School pilot 1 - KWF-STW [in writing] 2 (response style), between-subject 10 min 1 1 OBOBiDi Perception</td>
<td>n</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>School pilot 2 - KWF-STW [in writing] 2 (storytelling style), between-subject 10 min 1 1 OBOBiDi UX - Enjoyment</td>
<td>y</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>School pilot 3 - KWF-STW [in writing] 2 (role), between-subject 10 min 1 1 OBOBiDi UX - Enjoyment</td>
<td>y</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

H: a comparison between a robot that has the role of a coach and an entertainer and two robots that have just one role.

### 2.3 Targeted population

Both the PAL and the KWF-STW project are aimed at children aged 7 to 14 with a severe illness that requires daily medical attention. However, it is important to note that to prevent overburdening the children [8] some evaluations were performed with healthy children.

### 2.4 Used off-the-shelf robot

The NAO Robot (SoftBank, France) was used in all included user studies.

### 3 METHODS

In this section we will discuss the relevant details of the used methods needed to analyses the issues we aim to tackle. Please consult the references listed in Table 1 for more details about the used methods.

#### 3.1 Experimental design

In all studies we aim to evaluate the effect of the implemented robot behavior on the children. For studies D, E, F, G, and H we created two versions of the robot behavior that were compared between-subject. For study B the comparison was within-subject. In studies A, and C all participants interacted with the same version of the robot.

#### 3.2 Participants

For the last three years several users studies took place during a holiday camp for children diagnosed with diabetes. Each year the camp was attended by approximately 20 children aged 7–11. The studies A, B, and E were part of these camps.

In study C, 13 participants, all patients at the diabetes care unit of the Gelderse Vallei hospital, completed the experiment. In study D, 101 school children participated and in the three KWF-STW pilot studies (studies F, G, and H) 79 school children between 8 and 10 years old participated.

#### 3.3 Experimental set-up and procedure

In the diabetes camp user experience studies (A) children engaged in various playful child-robot interactions. For example, playing a quiz with the robot, or team-up with the robot and another child in a sorting game.

At the start of A, the B study was executed. The children watched two robots giving, in turns, a 10-minute presentation about robotics and diabetes.

In study C, participants interacted with the physical Nao in the hospital and at home they interacted with the virtual Nao that was present in myPAL. Both robot instantiations were treated as one entity. Participants could interact freely with the virtual robot at home.

The perception studies (D, E) consisted of a 10-minute presentation, including three multiple choice questions, given by a single robot. In all three experiments, the non-verbal behavior of the robot varied between groups.
In studies F, G and H the respective robots were set-up in a separate room at a school. Participants interacted one-by-one with the robot in one of the conditions for 10 minutes.

### 3.4 Measures and instruments

A wide range of measures and questionnaires were used in an attempt to answer all the respective research questions. In this section we will introduce the different style of questionnaires we used. In Table 1 the related measures are listed.

For A, a questionnaire was developed asking for children’s attitude towards the robot, focusing on likability and perceived educational value. This included nine 7-point Likert scale items and four open-questions.

For B the Fun Semantic Differential Scale, Smileyometer, and Children IMI Interest/Enjoyment Scale, from the validated PuppyIr User Evaluation Toolkit [11] were used. Additionally, one forced-choice question was added to measure perceived role.

For D and E, a new questionnaire was developed in collaboration with experts from cognitive psychology and human-computer interaction. This resulted in the adjective stickers-method, where 20 child friendly adjectives were selected that related to one of the four measurement constructs. Children were given these words as stickers, along with a response leaflet where they could indicate to what extend the word described their perception of the robot (see figure 1)[10].

For study C a different questionnaire method was developed. In an attempt to prevent asking leading questions [13] all items were formulated bi-directionally. For example, instead of the usual “Robin the robot was important for me to add activities. I strongly disagree [1][2][3][4][5] I strongly agree” the item was formulated as “Robin the robot wasn’t [1][2][3][4][5] was important for me to add activities”.

Furthermore, in a previous study3 a primacy-recency bias was encountered, where children blindly answered each question the same repetitively. To prevent this primacy-recency bias each question was printed on a separate piece of paper. The interviewer placed the paper in front of the participant, read the question slowly out loud and asked the participant to place a Playmobile figurine on the appropriate numbered square (see Figure 2). This resulted in one-by-one bi-directional items or OBOBiDi items for short. The OBOBiDi method was used in studies F, G, and H as well.

---

3The user study is not included in this case study because the authors were not directly involved in that study. The authors were warned for a primacy-recency bias.

---

### 4 CONSTRUCT VALIDITY VIOLATIONS

One of the challenges in child-robot evaluation is being sure that you measure what you aim to measure (i.e., construct validity). A question is not always understood properly or interpreted the same by every child.

#### 4.1 Observable behavior

A clear case where the construct validity was violated can be found in study B. Children assigned different categorical values to the robot’s role, but gave comparable rationals explaining their choice. The robot was perceived as a parent by a child, and as a teacher by another. When asking the children why they had selected this role, one mentioned that the robot was ’smart just like my father’ and the other commented that the robot was ’smart just like my teacher’. While another child said the robot to be ’like a father because he was so nice’. This observation indicates that children’s responses are influenced by their mental models.

#### 4.2 Root cause analysis

Children do not always interpret a concept or question the same. This may be explained by individual differences such as past experiences that lead children to develop different mental models [4]. Concepts are represented and categorized in the brain as (social) schemas. These schemas are (often unconsciously) activated and influence judgments. For example, one might associate authority with a teacher. Moreover, children might be unfamiliar with specific vocabulary or phrasing of the question.

### 5 CEILING EFFECT

With more than one user study we encountered a ceiling effect of the responses of children on one or more measures. A ceiling effect happens when participants give extreme answers to rating scales. This makes it impossible to conclude whether an actual effect was present [3].

#### 5.1 Observable behavior

We have observed children giving the maximum score in favor of the robot for most of the items in a questionnaire in multiple of our user studies (see table 1). This occurred typically for items related to user experience and more specifically to items focusing on the enjoyment of the interaction or the robot. For example, “I enjoyed my conversation with the robot”.

This was observed in studies F, G, and H where the mean scores on the 5-point Likert scale enjoyment questions were all higher than 4.5, in A where the mean score on a 7-point Likert scale was 6.5,
and in B where two different robots received a mean score of 4.1 and 4.3 on a 5-point scale measuring enjoyment. With a few exceptions, in these examples all the participants gave the maximum score regardless of their experience.

It is important to note that no ceiling effect was observed for questions related to user’s perception of the robot in any of our user studies. Furthermore, in study C also no ceiling effects were observed including the enjoyment questions.

5.2 Root cause analysis

The question is why do children give the robot mostly a maximum score? This could be attributed to one or more response biases. Children have been found to be susceptible for an extreme response bias where they give maximum scores for the sake of giving maximum scores. Furthermore, children can also give maximum scores because they expect that the researchers want those scores. This is called the social desirability response bias [12].

Another type of bias that is likely to play a role is a stimulus bias called the novelty effect [5]. Because interacting with a robot is a new and thrilling experience for the children they give the robot high ratings regardless of the behavior of the robot. This also explains why the ratings are less extreme after repeated interactions over a longer period of time, like in study C. That is because the novelty effect fades after multiple interactions [6].

However, the question of why the ceiling effect predominately manifests itself on user-experience questions such as enjoyment remains. Possibly when asked about a personal experience that was fun children inherently give a maximum score. More indirect questions about the robot (e.g. children’s perceptions) and non-robot related questions might be less affected by the novelty effect.

It is important to note that the primacy-recency effect, where children give the same answers to all items repetitively was not observed in any of our studies. The counter measures, namely using the OBOBiDi-method or the adjective stickers-method, seem to be successful in overcoming the primacy-recency effect.

6 LESSONS LEARNED

The first lesson we learned from the issues we encountered while using questionnaires to evaluate child-robot interaction is that proper questionnaire design is essential. When the language and concepts used in questionnaires do not match the level and experiences of the children there is room for misinterpretations. This affects the validity of the answers.

We have furthermore learned that while responding to questions children can be biased in different ways. The first step is to be aware of the possible biases. Note that the biases we discussed in the previous section are just a subset of all the possible biases that influences questionnaire responses [12].

Proper questionnaire design might be a (partial) solution to at least some of the biases. For example, we have seen that the OBOBiDi-method (e.g. in study C) and the adjective stickers-method (e.g. in study D & E) circumvent the primacy-recency bias. The adjective stickers-method could furthermore prove to be robust against the extreme and social desirability response biases, because its way of questioning is more indirect. More research is needed to validate both methods and to increase the robustness against different response biases.

A final lesson is that in child-robot interaction research we need to deal with the novelty effect. It not only affects questionnaire responses but also behavioral responses, e.g. during the interaction, that is often measured additionally, or as an alternative, to questionnaires [5]. Practically the only way to overcome the novelty effect is to conduct a more long-term study where the children interact multiple times with the robot. However, this exposes a more fundamental problem for child-robot applications and that is maintaining the interest of users [6]. In order to properly evaluate a child-robot application it is important that a user study reflects the long-term and repeated nature of the interaction. It strengthens research fidelity, helps to overcome the novelty effect, and it gives better insights in the actual performance of the robot.

ACKNOWLEDGEMENT

This work is partly funded by the EU Horizon 2020 PAL project (grant number 643783). This work is also part of the research programme Technology for Oncology (project number 15198) which is financed by the Netherlands Organization for Scientific Research (NWO), the Dutch Cancer Society (KWF), the TKI Life Sciences & Health, Assolutions, Brocaef, Cancer Health Coach, and Focal Meditech. We would like to thank all the healthcare professionals, teachers, involved students, and most importantly the children without whom the presented studies would not have been possible.

REFERENCES


