electronics

MDPI

Article

Lessons in Developing a Behavioral Coding Protocol to Analyze In-the-Wild Child–Robot Interaction Events and Experiments

Xela Indurkhya [1,*] and Gentiane Venture [2]

[1] Department of Systems Mechanical Engineering, Tokyo University of Agriculture and Technology, Koganei Campus, Tokyo 184-0012, Japan

[2] Department of Mechanical Engineering, The University of Tokyo, Hongo Campus, Tokyo 113-0033, Japan; venture@g.ecc.u-tokyo.ac.jp

* Correspondence: xelaindurkhya@gmail.com

Abstract: Behavioral analyses of in-the-wild HRI studies generally rely on interviews or visual information from videos. This can be very limiting in settings where video recordings are not allowed or limited. We designed and tested a vocalization-based protocol to analyze in-the-wild child–robot interactions based upon a behavioral coding scheme utilized in wildlife biology, specifically in studies of wild dolphin populations. The audio of a video or audio recording is converted into a transcript, which is then analyzed using a behavioral coding protocol consisting of 5–6 categories (one indicating non-robot-related behavior, and 4–5 categories of robot-related behavior). Refining the code categories and training coders resulted in increased agreement between coders, but only to a level of moderate reliability, leading to our recommendation that it be used with three coders to assess where there is majority consensus, and thereby correct for subjectivity. We discuss lessons learned in the design and implementation of this protocol and the potential for future child–robot experiments analyzed through vocalization behavior. We also perform a few observational behavior analyses from vocalizations alone to demonstrate the potential of this field.

Keywords: child–robot interaction; in-the-wild HRI; behavioral coding; vocalization-based analysis

check for updates

Citation: Indurkhya, X.; Venture, G. Lessons in Developing a Behavioral Coding Protocol to Analyze In-the-Wild Child–Robot Interaction Events and Experiments. *Electronics* **2024**, *13*, 1175. https://doi.org/10.3390/electronics13071175

Academic Editors: Flavio Canavero, Eleuda Nuñez, Marie-Monique Schaper and Bipin Indurkhya

Received: 22 December 2023

Revised: 1 March 2024

Accepted: 18 March 2024

Published: 22 March 2024

1. Introduction

In the field of human–robot interaction, there are largely two ways to assess human responses to and attitudes toward robots. One is by observing behavior [1–3], the other is self-reporting, through means such as interviews and questionnaires [4–6]. It has become common to conduct visual behavioral analyses from video footage of in-the-wild human–robot interaction experiments [2,7,8]. Analytical measures can be elements such as how long children remain in a room with a dancing robot, or subjective measures such as the "goodness" of interaction between the robot and children as assessed by "judges" viewing the interaction on video [1].

However, these established methods can be somewhat limiting in the context of in-the-wild studies, particularly with child participants. Outside of a laboratory setting, it is common to run into situations in which the filming of children, even for experimental purposes, is highly scrutinized to the point where it might be forbidden or highly limited [4,9]. While in some contexts, there are behavioral analyses that can be performed according to established practices within these limitations, in many cases, performing a visual-based behavioral analysis of film data that is expressly designed to have poor visibility of the subjects is unlikely to yield ideal or even useful results. There have been studies conducted that have circumvented this potential problem by focusing on interviews with participants after an observed interaction in the wild [4].

Our lab has often experienced such restrictions in conducting in-the-wild child–robot interaction (CRI) experiments [9]. In obtaining consent to conduct events in collaboration with local preschools and kindergartens, the event was often required to be conducted separately from regular school, in order to make it something parents could opt-in to rather than opt-out of. From 2017 to 2019, our lab attempted to conduct several in-the-wild experiments with local preschools using the Pepper robot by SoftBank Robotics in Minato-ku, Tokyo, Japan. Due to constraints required by the school around the filming of children, these interactions of twenty to thirty children with one robot could only be filmed from the back of the room. This creates a difficulty in conducting traditional, visual behavioral analyses on these interactions as described above. With the video being taken from behind, it is impossible to view facial expressions unless the child turns away from the robot. Furthermore, with so many children involved, the children closest to the robot are frequently not visible at all. Interviews and questionnaires were frequently not possible or practical, or limited in their usefulness to analyze the interactions as a whole.

There are a handful of in-the-wild experiments that performed observation-based analyses, which relied heavily on a vocal element [3,6,10]. These papers continue to reference visual elements such as facial expression, and where visual data are available, it is valuable to incorporate into the analysis, just as there is value in incorporating verbal and vocal behavior into a primarily visual analysis. Due to the particular restrictions we have encountered around the acquisition of clear video data in an in-the-wild setting, we aimed to create a protocol which is able to analyze in a more systematic manner based on vocalization behavior without reliance on visual elements. While the audio is no less chaotic than the visuals on the existing experiments, they are less impeded by the constraints of the setting and camera placement than body language or facial expression analyses would be. Furthermore, as the acquisition of audio recordings in interactions with children are not viewed with the same suspicion as visual recordings, the ability to analyze interactions from audio data could provide more opportunities to conduct in-the-wild child–robot interaction studies in the future. Even in a setting where a camera is not welcome at all, audio recordings may still be permitted. The focus of this paper is to show the development of behavioral categories to begin a vocalization-based behavioral coding protocol, as well as to highlight some of the ways in which audio-based behavioral analysis may be useful, even where high-quality visual data are available. The behavioral coding scheme developed was based loosely upon a behavioral coding scheme used in behavioral studies of wild dolphin populations, which is designed to use what few visual cues are available to researchers above the water surface in the identification of five types of behaviors [11–13]. The transfer of the methodology from wildlife studies of dolphins to in-the-wild studies of human children was motivated by the fact that these are both methodologies designed to perform simple analyses of behavior with limited visibility and no ability to interact with the subjects being studied, though the environments and reasons and behaviors being studied are very different.

The process described in this paper is outlined in Figure 1. The paper is structured thusly: In Section 2, we describe the process through which children were recruited for the interactions used in this paper. In Section 3, we describe the literature background and the categories we outlined and defined for our vocalization-based behavioral coding protocol. In Section 4, we assess intercoder agreement through the first round of coding; in Section 5, we continue this process through another round, with redefined categories and one additional coder, and conclude that there is a need to use a multiple-coder analysis due to subjectivity of the data sets. In Section 6, we demonstrate some qualitative analyses that can be conducted with the same data based on information that can be gleaned from audio. In Section 7, we test our hypothesis about the subjectivity inherent in group CRI studies in the wild using a different data set and two new coders, using the final category definitions outlined in Section 5. We discuss our findings in Section 8, and lay out our conclusions concisely in Section 9.

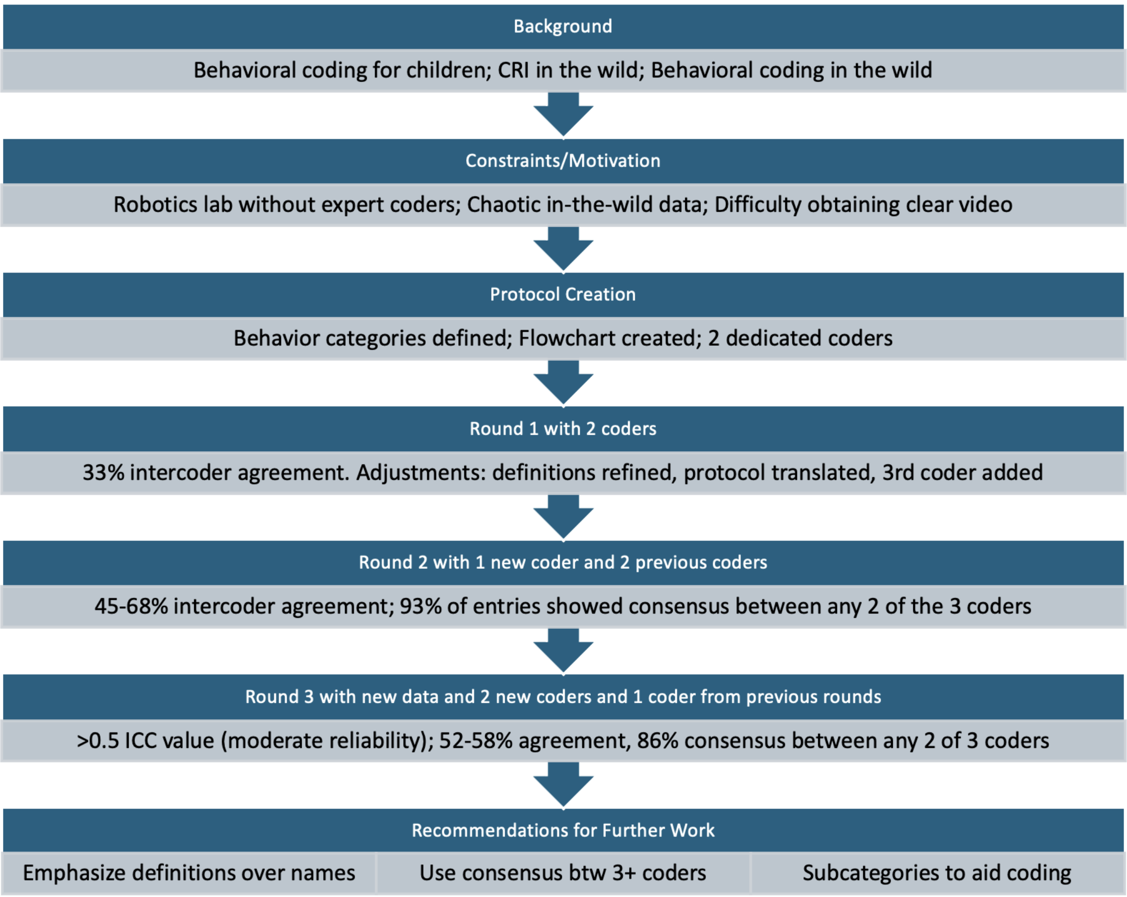
Background
Behavioral coding for children; CRI in the wild; Behavioral coding in the wild
Constraints/Motivation
Robotics lab without expert coders; Chaotic in-the-wild data; Difficulty obtaining clear video
Protocol Creation
Behavior categories defined; Flowchart created; 2 dedicated coders
Round 1 with 2 coders
33% intercoder agreement. Adjustments: definitions refined, protocol translated, 3rd coder added
Round 2 with 1 new coder and 2 previous coders
45-68% intercoder agreement; 93% of entries showed consensus between any 2 of the 3 coders
Round 3 with new data and 2 new coders and 1 coder from previous rounds
>0.5 ICC value (moderate reliability); 52-58% agreement, 86% consensus between any 2 of 3 coders
Recommendations for Further Work
Emphasize definitions over names
Use consensus btw 3+ coders
Subcategories to aid coding

Figure 1. Workflow of the coding process described in this paper.

2. Selection and Participation of Children

The experiment discussed for the majority of this paper (with the sole exception of Section 7) was conducted with a preschool in suburban Tokyo in 2018 with Pepper (SoftBank Robotics, Minato-ku, Tokyo, Japan). The three sequential events were planned with teachers, and conducted as an optional after-school event. Parents and children were informed that an experiment would be conducted in which children would play with the robot, and would be filmed from the back of the room. They were informed that the images containing children and any part of the data containing identifying information would be used internally for research purposes only, and not made available through any social media or public forum.

For the data used in Section 7, these were two events conducted in India in 2019. Interactions used in this paper consisted of segments of a Wizard of Oz-style free interaction with Anki's Vector (Anki has since been acquired by Digital Dream Labs in Pittsburgh, PA, USA). One event was conducted at a small business in Madhya Pradesh, and the other at a private residence in Telangana. Children were recruited by word of mouth, with parents fully informed that the children would be taking place in a child–robot interaction experiments. Parents were present at all times, and assured that no identifying information would be released.

3. Creating a Behavioral Coding Protocol

The nature of behavioral coding is to take complicated real-world values and sort them into a defined code according to what is to be analyzed, thereby making it possible to analyze patterns and changes in behavior. The definitions chosen to sort behaviors necessarily both biases and limits the possibilities of interpretations. Behavioral coding is often applied in a lab scenario as a basis for comparing self-reported measures against observed measures [14]. Its ability to streamline and quantify qualitative information is of particular interest to us. In addition to providing a means of analysis of in-the-wild interactions, behavioral coding could provide an avenue through which various in-the-wild CRI experiments might be compared and contrasted to one another.

The closest child-directed behavioral coding scheme to the sort we wished to use was one in which caregiver reports were used to score the frequency of a given set of behaviors of a child toward the family's pet dog on a scale of 1 (never) to 6 (very often) [15]. However, this system relies on having an adult who can report on the child's behavior over sustained periods of private interaction with the dog (or robot, as the case would be for us), which is not the case with our in-the-wild CRI experiments. It is particularly tricky to assess the behavior of children based on an experimental interaction, and even studies that rely heavily on behavioral coding are often highly structured and use video data to assess the degree of specific emotions, such as distress [16,17].

Many behavioral coding schemes, especially those involving children, are designed to assess intensity of a particular emotion based on a preexisting hypothesis [16], which was not the case with our experiments. As is common with in-the-wild CRI experiments, the experiments were designed primarily with the goal of engaging and entertaining the children [18], many of whom would likely not participate in a laboratory experiment, with analysis being a secondary concern. Furthermore, many of these behavioral studies rely on having several people willing and able to train on a behavioral coding scheme [19], and at the time of this experiment, our lab only had access to two persons with the time and willingness to dedicate hours to training in a behavioral coding protocol. Questionnaires are a common way to assess the thoughts and feelings of a children after an experiment [1,20,21], as are interviews [4,21]. However, this is not ideal for group experiments where many children of age 5 and younger interact with the robot at once, and afterwards parents are eager to go home. We have tried approaching this by handing the parents questionnaires that they could optionally return to the school, but as these questionnaires must be prepared in advance of the experiment, there was no opportunity to discuss anything which happened over the course of the experiment. Additionally, our goal in analyzing these experiments was not to create a behavioral coding protocol that looked at the nuances of emotional states or gain insights into the personality of each particular child, but simply to assess general engagement with the robot on a group level. Therefore, we drew on methods used in field studies of wild dolphin populations, outlined in [11,12] and used to this day [13,22]. The similarity of dolphin behavioral studies to our in-the-wild scenarios lies in that while the majority of dolphins' lives take place underwater, the researchers studying them can usually only do so based on what is visible at the surface, and group behaviors were classified based on inferences that can be made from elements visible at the surface. Similarly, we sought to identify behavioral categories that could be inferred from audio even if visuals were obscured, which do not require the tracking of specific individuals. In developing our coding scheme, we consulted the guidelines laid out by Chorney et al. [23].

The guideline outlined in [23] consists of four major sections, and subsets of questions for each section. The four sections are: refining the research question, developing and refining the coding manual, piloting and refining the coding manual, and implementing the coding scheme. With regard to the research question, we are interested in the behavior of children (and, to a lesser extent, the adults around them). We are interested in their behavior as relates to the robot, and to what extent they appear to be engaging. The rest of Chorney's outline about refining the research question consists of defining observation periods and recording methods. As at present, we are using video files of in-the-wild experiments conducted in the past, we use the full files available. Our main focus is on developing and piloting the coding manual. The key elements of this process can be broken into three parts: converting the recording data into a transcript, developing and refining behavioral categories, and assessing inter-coder variation.

Because the research is being conducted in Japan by a multicultural, multilingual lab, the behavioral coding was designed bilingually, mindful of the cultural and linguistic differences both in general and specifically in regard to robots [6,24]. Our lab did not have access to more than one student trained in psychology who could understand and code in Japanese. Furthermore, many of our in-the-wild interactions were recorded with the understanding that the videos would not be viewed by anyone outside of the lab. Therefore, one of our aims was to design a behavioral analysis protocol for non-specialist coders with no experience in behavioral coding.

3.1. Creating a Transcript for Analysis

To ensure the coders could identify what they were coding, and to aid in checking inter-coder agreement, we first created a transcript. This process was conducted manually, as at the time of transcription, we could not find a software that could reliably pick up children's speech in Japanese from the quality of audio we had (sample audio is available at http://gvlab.jp/Material/BehaviorCodingProtocolv4/Experiment%20Audio%20Sample.mp3, (accessed on 17 March 2024) with the corresponding transcript at http://gvlab.jp/Material/BehaviorCodingProtocolv4/Experiment%20Audio%20Sample.pdf, (accessed on 17 March 2024); coders did not, at any point, code from audio alone, it was always accompanied by video. Video is redacted in this sample audio to protect the privacy of the children). In this study, we analyzed three interactions conducted at the same preschool over 3 months in early 2018. Children were aged 4–5, and there were 20–30 children in each interaction. Each experiment consisted of roughly four stages: (1) an introduction stage where the children enter the room, while the robot is immobile; (2) a stage where the robot conducts a roll call as a teacher would at the beginning of the school day, and the children respond as they would to a teacher; (3) an activity familiar to the children suggested by the teachers; and (4) a period after the experiment when the robot is once again immobile, as some children leave and some linger. The activity in the first experiment was a dance, and in the second experiment it was a picture book, which the robot narrated while the pictures were displayed on its screen, with additional questions designed to prompt contribution from the children. In the third experiment, both activities were performed, with a transition period in between.

The three experiments were transcribed manually into a spreadsheet. Each entry was a distinct vocalization made by a single speaker. Where there was a distinct vocalization cutting across another, lengthier vocalization by a single speaker, this would be represented as three separate entries: the beginning of the lengthy vocalization, the vocalization that cuts over it, and the subsequent part of the lengthy vocalization. There were no cases of a distinct single speaker speaking at length that would merit redacting the specifics from the transcript, except the robot when narrating a storybook. Where only a fragment of a sentence was audible, that fragment was entered as a single entry into the transcript.

Where there was laughter or indistinguishable voices or another vocalization difficult to transcribe, these were input as a single description entry in the transcript. As the three interactions analyzed were fully in Japanese, the description entries were input in English to avoid confusion. Children and adults were easily distinguished by speech pattern and voice, and there was no ambiguity.

The transcripts were designed not to be the source of the behavior analysis, but as a tool to ensure coders were responding to the same vocalizations, as well as a means to assess when different coders were hearing different things. Coders were instructed to code from the videos, and code for what they heard rather than what was written in the transcript.

3.2. Behavioral Categories

The interest of our study is in the vocal behavior of children and, to a lesser effect, the adults around them. The vocalizations of the robot(s) and experimenter(s), therefore, are excluded from the behavioral coding analysis. In defining the behavioral categories, we focused on defining a set of 5–6 behaviors which could be clearly distinguished from each other by voice alone.

Because our aim was to investigate engagement in the robot, broadly, we sought to isolate robot-related vocal behavior from non-robot-related vocal behavior, and subcategorize the latter. Table 1 lists the nominal categories used for behavior coding with their definitions, which were arranged for coders in the form of a flow chart, which first asked if the behavior was robot-related to separate robot-related vocal behavior (**To**, **About**, **Responsive**, **Reactive**, and **Dialogue**) from non-robot-related vocal behavior (**Other**), in addition to designating a category entries that could not be analyzed (**Unclear**) for any reason. The subcategories of robot-related behavior were then categorized by a series of questions in the flow chart, as shown in Figure 2. The expectation behind this categorization is that, based on the frequency of robot-related versus non-robot-related vocal behavior, we would be able to assess general levels of engagement in a way that could be quantified, and that through the subcategories, we would be able to see the broad type or types of engagement being exhibited.

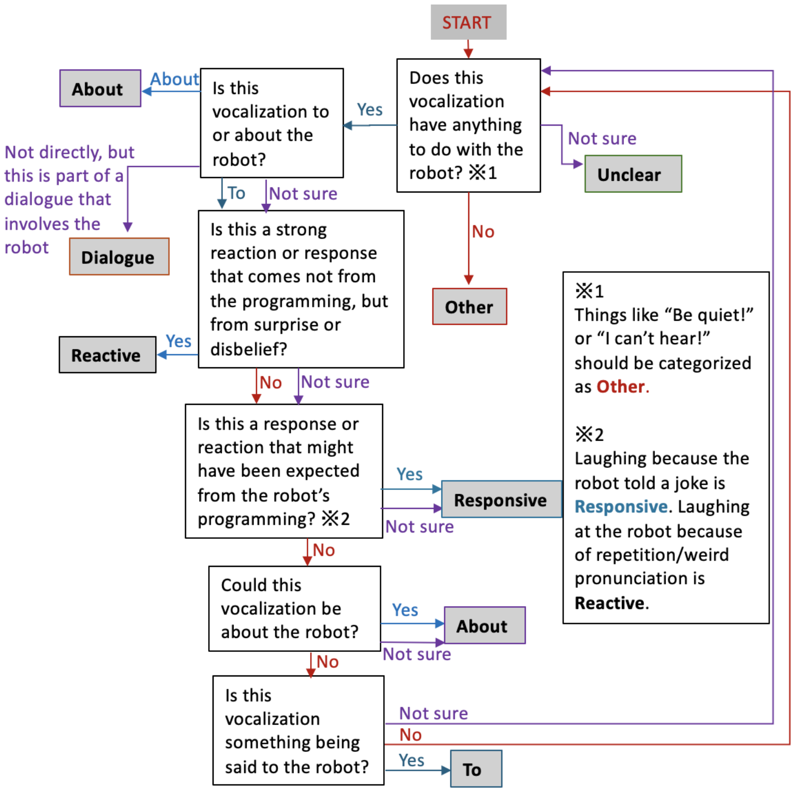
START
Does this vocalization have anything to do with the robot? ※1
Yes
Not sure
Unclear
No
Other
Is this vocalization to or about the robot?
About
About
Not directly, but this is part of a dialogue that involves the robot
Dialogue
To
Not sure
Is this a strong reaction or response that comes not from the programming, but from surprise or disbelief?
Yes
Reactive
No
Not sure
Is this a response or reaction that might have been expected from the robot's programming? ※2
Yes
Not sure
Responsive
No
Could this vocalization be about the robot?
Yes
Not sure
About
No
Is this vocalization something being said to the robot?
Not sure
No
Yes
To
※1
Things like "Be quiet!" or "I can't hear!" should be categorized as Other.
※2
Laughing because the robot told a joke is Responsive. Laughing at the robot because of repetition/weird pronunciation is Reactive.

Figure 2. Representation of the flowchart used in round 1 of behavioral coding.

Table 1. Definitions of nominal categories used for each round of coding.

Behavior Category	Round 1 Definition	Round 2 Definition	Final Definition
To	Directly addressing the robot	Directly addressing the robot	Directly addressing the robot, unprompted by the robot
About	Directly speaking about the robot	Directly speaking about the robot	Directly speaking about the robot
Responsive	Direct responses to the robot that could be predicted from its programming	Responses and reactions to the robot	Responses and reactions to words/actions of the robot
Reactive	Direct reactions to the robot indicating that the robot has done something unexpected	N/A	N/A
Dialogue	Part of a robot-related dialogue	Part of a robot-related dialogue; not directly to or about the robot	Robot-related, but none of the above
Other	Not robot-related	Not robot-related	Not robot-related
Unclear	Cannot be analyzed	Cannot be analyzed	Cannot be analyzed

We designed these behavioral definitions to be broad categories, applicable to a variety of in-the-wild child–robot interaction scenarios, primarily with the aim of identifying and differentiating a handful of types of interactivity with the robot. **To** suggests high interactivity, engaging the robot entirely unprompted; **Responsive** and **Reactive** suggest a more passive interactivity, in that they are interacting, but only because of some stimulus provided by the robot; **About** suggests interest without interactivity, as the robot is the subject of discussion but not being engaged with; and **Dialogue** is a catch-all label for other parts of robot-related activity that do not fit into any of the other labels. In order to counter the subjectivity inherent in these categories, the coders were instructed to code following a flow chart. As shown in Figure 2, categories were sorted through a series of questions, beginning from the broadest (**Other**, **Unclear**, and **Dialogue**), through the increasingly narrow categories (**About**, then **Responsive**, then **Reactive**, then **About** again to catch any points of ongoing ambiguity) and to the most exclusive (**To**). Each of the points on the flowchart were accompanied by prompting questions, as well as some clarifications (such as instructions for how to categorize certain points of ambiguity, such as laughter). **About** was on the chart twice because the prompting questions were different each time; in the first instance, the coder was asked, "Is this vocalization to or about the robot?" and in the second (after **Responsive** and **Reactive** behaviors have been categorized and therefore excluded), the coder was asked, "Could this vocalization be about the robot?" to minimize ambiguous behaviors being categorized as **To**.

In an attempt to see if it was possible to perform a similar analysis using behavioral categories that emerged from this specific experiment, a secondary, incidental behavioral category was also created, which was a category of potential interest for analysis that were identified on viewing the data in question. These categories were not given particular definitions beyond the word used to define the category. Coders could note instances of the behaviors in Table 2 in a separate column, if they seemed applicable. Coders were encouraged to apply their subjective interpretations to this category.

Table 2. Incidental behaviors of interest identified across the 3 experiments.

Behavior	Definition
Interactive	Addressing the robot as if seeking a response/reaction
Non-interactive	Behaviors that seem to presume the robot will not react/respond
Surprise	Indication that the robot has done something unexpected
Tsukkomi	Jokingly pointing out that the robot has done something odd
Insult	Calling the robot names or mocking its pronunciation
Fear	Use of words indicating fear (tone may not sound fearful)
Command	Instructing the robot to do something
Death	Robot is referred to as dying or dead
Impatience	Indication that they want to interact with the robot
Helpful	Behaviors that are meant to help the robot
Aggressive	Aggression seen toward the robot

Additionally, coders were asked to note where they felt the children were behaving toward the robot as though to a human or an object.

Finally, where the robot was referred to by name, it was noted whether the robot had been referred to by just the name with no suffix, or with the suffix -kun (a suffix used for boys by classmates). This last category was considered self-evident, and therefore notated only by the primary coder.

Explaining the Flowchart

The flowchart begins with the question *Does this vocalization have anything to do with the robot?*, which serves to differentiate non-robot-related behavior from robot-related behavior, while also discarding those which cannot be categorized into the **Unclear** category. Should the coder answer *Yes*, the next question asks, *Is the vocalization to or about the robot?*, which serves to sort out cases where participants are clearly speaking about the robot in third person into the **About** category; this is also the stage at which vocalizations which cannot be clearly categorized are sorted into **Dialogue**. If the coder answers *To* or *Not sure*, they are next asked, *Is this a strong reaction or response that comes not from the programming, but from surprise or disbelief?*, which is a question designed to isolate out **Reactive** behaviors, i.e., behaviors indicating that the robot has behaved contrary to the expectations of the participants. The next question asks, *Is this a response or reaction that might have been expected from the robot's programming?*, which is designed to isolate out **Responsive** behaviors, i.e., engaging with the robot as intended. At this stage, with the **Reactive** and **Responsive** behaviors now coded, the answer *Not sure* leads to the same code as *Yes*. Because at the second question, the response *Not sure* led coders down this path, at this point, vocalizations, which could be interpreted as speaking about the robot in third person but have been included up to this point due to ambiguity, are sorted into **About** if the coder answers *Yes* or *Not sure* to the question, *Could this vocalization be about the robot*. Finally, answering *Yes* to the question *Is this vocalization something being said to the robot?* leads coders to the **To** category. Should the coders respond *No* or *Not sure* at this point, they are led back to the beginning of the chart to reassess the applicability of the **Unclear** and **Dialogue** categories. Should coders find themselves looping back to the beginning more than once, they were instructed to categorize the behavior as **Unclear**.

4. Coding Process and Inter-Coder Agreement Analysis

Initially, the coding process was conducted with two coders. Both coders were native Japanese speakers, and coded the full three experiments described in Table 3. For the purposes of testing the protocol, all instances of a robot or adult speaking were excluded from analysis, including only children's vocalizations. Coders were instructed to watch and listen to the experiment video recording, rather than coding from the transcript alone.

Table 3. Data from 3 experiments used for coding.

Experiment #	Duration	Total Entries	For Analysis
1	27 min	488	394
2	34 min	758	534
3	36 min	553	453

The intercoder agreement was then assessed. Results were analyzed by calculating percentage agreement and interclass correlation coefficients (ICC). ICC was calculated in Excel, using a two-factor ANOVA without replication and then applying the formula for two-way random effects and absolute agreement by a single rater [25]. While there are other methods to analyze intercoder agreement, they often rely on having a preset percentage of expected agreement [26], which we did not have. For all categories except the general behavioral coding section, where all applicable cells were filled in, percentages were calculated, excluding all cases where both coders left the field in question blank. To calculate the ICC, the nominal categories were converted to numerals from 1 to 9, as shown in Table 4, and blank cells were converted to 0. Numbers of the categories were chosen to roughly denote some sort of scale. The smaller numbers were assigned to the more robot-centric behaviors, the number 8 assigned to the least robot-centric behavior, and 9 assigned to behaviors that could not be categorized. It should be noted that, while we converted the categories into numerical values in order to be able to conduct a statistical analysis other than percentage of agreement between coders, the categories are not designed to be numerical, and thus any ICC value should be taken as more of a guidance in conjunction with percentage of agreement rather than as a definitive statistical test.

The results can be seen in Table 5. All ICC values are below 0.5, indicating low reliability.

Table 4. Conversion of nominal categories to numerical for the purposes of calculating ICC values. For the nominal coding categories, this order also reflected a hierarchy of categorization, wherein where there was ambiguity between two categories, the one lower (of larger numeric value) would be selected.

Nominal Coding	Incidental	Hmn/Obj	Numerical Value
To	Interactive	Human	1
	Helpful		1.5
About	Command		2
	Tsukkomi		2.5
Responsive			3
Reactive			3.5
Dialogue	Insult		4
	Aggressive		5
Other	Fear		6
	Surprise		6.5
	Death		7
	Non-Interactive		8
Unclear		Object	9

Table 5. Inter-coder agreement analysis using percentages and interclass correlation coefficients (ICC) after first round of coding.

Category	Percentage Agreement	ICC
Behavioral Coding	33.82 %	0.4316
Incidental Behaviors	12.30 %	0.1116
Human/Object	56.20%	0.4236

Discussion between coders revealed that several points were being interpreted differently. The definitions of the coding categories were not sufficiently clear, leading to such differences as coder 1 consistently coding **Responsive** where coder 2 coded **Dialogue**, and coder 1 often coding **To** where coder 2 coded **Responsive**.

These differences in understanding of definitions were much starker in the incidental behavior category, partly due to the absence of clear definitions for the behavior categories specified. For example, coder 1 interpreted “Aggressive” more violently than coder 2, who interpreted it to mean socially pushy. Consequently, coder 1 notated “Aggressive” 4 times, while coder 2 notated it 419 times. “Aggressive” consequently made up more than half of the incidental behaviors coded by coder 2. Upon discussion between coders 1 and 2, it became clear that while coder 1 had interpreted the word to mean “borderline violent”, coder 2 had interpreted the word to mean “socially pushy”. These differences in understanding can be attributed to the fact that while the interaction was in Japanese, these categories were named in English, in which the coders had a disparate understanding of keywords and their definitions. While inter-coder variation was to be expected, having encouraged coders to apply their subjective interpretation to the categories, this amount of variation means that this category cannot be assessed as a behavioral coding scheme; we considered, however, that these categories could still be used to note certain patterns and perform a qualitative analysis. Due to the particularly high subjectivity observed around the “Aggressive” category, however, that category was omitted going forward.

5. Refining the Coding Scheme

In order to refine the protocol, the behavior categories were redefined for round 2, as shown in Table 1. The incidental behavior category and the human-object category were not applied this time. The flowchart was also altered as shown in Figure 3, with both English and Japanese variations made available to coders.

Another attempt was made at refining the code, by asking both coder 1 and coder 2 to recode a segment of the data. Due to the low intercoder agreement observed during round 1, we also added a third coder, who was fluent in Japanese. The data used was once again from the dataset described in Table 3, but this time only a subsection of the data were coded, as described in Table 6.

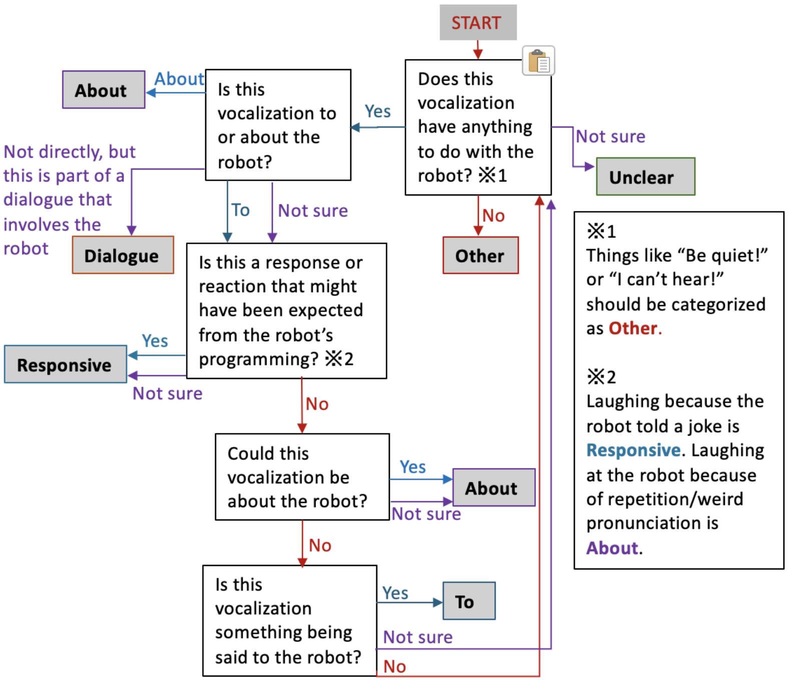
START
Does this vocalization have anything to do with the robot? ※1
Yes
Not sure
Unclear
No
Other
Is this vocalization to or about the robot?
About
About
Not directly, but this is part of a dialogue that involves the robot
Dialogue
To
Not sure
Is this a response or reaction that might have been expected from the robot's programming? ※2
Yes
Responsive
Not sure
No
Could this vocalization be about the robot?
Yes
About
Not sure
No
Is this vocalization something being said to the robot?
Yes
To
Not sure
No
※1
Things like "Be quiet!" or "I can't hear!" should be categorized as Other.
※2
Laughing because the robot told a joke is Responsive. Laughing at the robot because of repetition/weird pronunciation is About.

Figure 3. Representation of the flowchart used in round 2 of behavioral coding.

Table 6. Data used for round 2 of coding.

Experiment #	Duration	# Data Entries
1	6 min	90
2	8 min	123
3	3 min	99

There were six instances where coder 3 coded two possible behaviors simultaneously. Before conducting agreement analysis, one of the two options was selected by assessing whether either of the other coders had used one of these options. If yes, that option was selected. If no, or if coders 1 and 2 had each chosen 1 of the 2 possible options selected, then the option was selected which was of higher numerical value, according to Table 4.

The results are shown in Table 7. As before, ICC values were calculated by converting the nominal values to numerical values according to Table 4, excluding **Reactive**. Once again, all ICC values were below 0.5, indicating low reliability. Total agreement between all three coders was 45.85% with an ICC of 0.3913. Agreement between coders 1 and 2 was 60.70% with an ICC of 0.4751, indicating low reliability, but improved reliability from round 1. However, if majority consensus was taken, identifying cases where any two coders were in agreement, agreement was 93.45%.

Table 7. Inter-coder agreement analysis for round 2 of coding using percentages and interclass correlation coefficients (ICC).

Coders	Percentage Agreement	ICC
Total agreement	45.85 %	0.3913
Coder 1 vs. 2	60.70 %	0.4751
Coder 2 vs. 3	56.33%	0.3728
Coder 1 vs. 3	68.12%	0.3513
At least 2 agree	93.45 %	N/A

Discussions between coders revealed that there continued to be confusion over the definition of certain categories, especially **Responsive** and **Dialogue**. Each coder adhered to a slightly different definition of these words, resulting in internal consistency for that coder, but lack of consistency between coders. Despite the questions in the flowchart and the definitions provided, the nominal category names led to each coder having their own sense of what was meant by that category which, at times, conflicted with the definitions and flowchart provided. While the increased agreement between coders 1 and 2 was evidence that training, discussion and further refinement of the definitions had been effective, that it was still at 61% after the discussions reflected a combination of the coders' subjective interpretation of the category names overriding the flowchart and definitions, as well as a certain amount of ambiguity in the vocalizations being coded. The matter of the definitions is a problem that could be averted by shifting to use coders with a higher expertise in psychology studies and behavioral coding, to do so would have defeated the purpose of creating a protocol that could be used by a robotics lab without multiple psychologists available. Therefore, going forward, we saw the need for two further shifts: (1) better definitions for this set of behavior categories (the final definitions as shown in Table 1), which can be used without the flowchart, which we hypothesized was potentially introducing some confusion to the coding process by providing coders with two possible means by which to arrive at a category; (2) an emphasis placed on the provided definitions of category names, perhaps by using letters instead of full words to which each coder brings their subjective definition; and (3) the creation of an alternate protocol where behavioral categories are broken down into many specific subcategories. Additionally, we considered that, due to the coders being instructed to categorize every vocalization as well as the chaotic nature of the data, there would necessarily be disagreement in cases of high ambiguity. This is to be expected at some points in any language-based interaction [27], but is likely amplified in in-the-wild scenarios where there is limited context guiding the coders' interpretations regarding the vocalizations of particular individuals. This subjectivity prohibiting high levels of intercoder agreement is not unheard of in behavioral coding, though it is considered rare, and can be mitigated through the use of multiple coders [23]. Therefore, it may be that, rather than using a primary coder supplemented by secondary coders to assess consistency, it would be better to look at the cases where several coders agree and take those as the behaviors which are unambiguous and useful for analysis within this protocol.

6. Qualitative Analyses

While behavioral coding is one way to describe a qualitative data set, this streamlining of data also flattens it [23]. It is therefore beneficial to describe the data through additional means to gain a clearer picture of the interactions [28]. While the incidental behaviors identified as possibly relevant were not well-defined enough to be part of the coding protocol, we nevertheless saw value in using those categories to perform a qualitative analysis more specific to these particular interactions. We also sought to attempt an analysis of the children's use of suffixes with regard to the robot as another possible avenue for vocalization-based behavior analysis.

6.1. Incidental Behaviors

Coder 1 went through the interactions and noted where they observed instances of the behaviors in Table 2 as an observational analysis [29]. The "Aggressive" category was excluded due to the unusually high inter-coder variation seen between coders 1 and 2, though for the purpose of this type of analysis, it could have instead been split up into two separate categories, one for socially pushy behavior and one for violent behavior. "Interactive" and "Non-interactive" were intended to only apply to instances where the intent seemed to be clear; where the coder felt there was ambiguity, neither category was applied.

The experiments each had four distinct stages (pre-interaction, roll call, activity and post-interaction), and it was notable that, in many cases, these behaviors were identified to be concentrated in similar stages across all experiments. Table 8 shows the number of occurrences of each behavior by stage, and Table 9 shows the percentage of total analyzable behaviors each of these incidental behaviors represented within each experiment. The results in Table 9 were calculated from the total number of data points analyzed for each experiment: 393 for experiment 1, 534 for experiment 2, and 453 for experiment 3.

Table 8. Number of occurrences of each behavior of interest through all 3 experiments, sorted by stage.

Behavior	Pre-Interaction	Roll Call	Activity	Post-Interaction
Interactive	40	26	11	7
Non-interactive	11	6	6	25
Surprise	3	12	12	0
Tsukkomi	0	13	2	0
Insult	6	10	8	2
Fear	3	3	0	0
Command	3	6	5	1
Death	4	4	1	8
Impatience	4	5	0	0
Helpful	3	2	2	0

Table 9. Percentage representation of each incidental behavior within each experiment.

Behavior	Experiment 1	Experiment 2	Experiment 3
Interactive	13.20	3.75	2.65
Non-interactive	3.3	5.43	1.32
Surprise	4.06	1.87	0.22
Tsukkomi	1.52	1.5	0.22
Insult	2.79	1.31	1.77
Fear	0.51	0.19	0.66
Command	1.27	1.69	0.22
Death	1.78	1.5	0.44
Impatience	1.02	0.37	0.66
Helpful	1.52	0.19	0

All behaviors except helpfulness, which was not observed in experiment 3, were observed across all three experiments. Interactive behaviors (clearly seeking reactions from the robot) were most commonly seen during the pre-experiment and roll call stages, while non-interactive behaviors (clearly treating the robot as something that will not react or respond) were most often seen in the pre- and post-experiment stages. References to the robot dying or being dead were also most common in the pre- and post-experiment stages, as well as the roll calls stage, with the highest occurrence being in the post-experiment stage. Impatience and talk of fearing the robot were observed only in the pre-experiment and roll call stages. Surprise, tsukkomi (a joking way of pointing out that the robot has done something wrong) and insults were primarily observed during the roll call and activity stages. Commands and helpfulness were observed across the pre-experiment, roll call and activity stages.

These results paint a picture of the children acclimating to the presence and actions (or inaction) of the robot over the course of the experiment, with fear, impatience and attempts to interact with the robot observed principally in early stages, surprise and condescension (interpreted through "insult" and "tsukkomi" behaviors) at the robot's abilities and limitations observed in the middle stages when the robot is active, and commenting on the robot as an object after the experiment, though there are still some children who attempt to initiate interactions, give commands or comment on the robot's immobility as "death" at this stage.

However, this acclimatization did not entirely last from one experiment to the next. Fear behaviors were observed across all experiments. Surprise and tsukkomi were seen with less frequency from each experiment to the next, but were nonetheless present in every experiment. This may reflect the presence of children in each experiment who had not been present in the previous experiments, or simply an effect of the time that had passed between each experiment (approximately a month between each).

6.2. Naming the Robot

Whenever the robot was referred to by name, we noted whether it was named "kundzuke", referring to following the name with the suffix -kun (as children would refer to a male classmate) or "yobisute", without a suffix (as an object, pet, particularly close friend, or indicating derision). No other suffixes were used to refer to the robot.

The number of instances of children referring to the robot by name and whether or not they used a suffix are indicated in Table 10.

Table 10. Instances of the Pepper robot being called by name with a note of whether it is with the -kun suffix (kundzuke) or without suffix (yobisute), sorted by experiment.

Experiment	Kundzuke	Yobisute
1	46	32
2	13	32
3	6	30

At times, a pattern can be discerned, which is more indicative of group dynamics than the robot itself. Within the first 3 min of the first video, while the robot is still turned off before the first experiment, the children start out referring to it as "Pepper-kun", until one child speaks up nicknaming the robot "Toiletpepper", without suffix. There is a chorus of children repeating this nickname, and when the children once again refer to the robot by name shortly after, they drop the suffix and call it "Pepper". Not long after this, the children begin to wonder why the robot is not moving, and once again begin referring to it as "Pepper-kun". After that, for the remaining 4 min before the robot begins to move and commences roll call, and then through the roll call itself, there are instances of children referring to Pepper with and without the -kun suffix with no discernible pattern. During the dance activity and after the experiment however, Pepper is only referred to in the yobisute form without a suffix.

Throughout the second and third experiment, the children sometimes use a -kun suffix, but in every stage of both of these experiments, there are more instances of the children referring to Pepper without a suffix than with -kun. This is shown in Table 11. The results show that, upon becoming acclimatized to Pepper's presence, the children increasingly referred to it without a suffix, though there was still the occasional child who kept the suffix.

Table 11. Instances of the Pepper robot being called by name with a note of whether it is with the -kun suffix (kundzuke) or without suffix (yobisute), separated by each stage of each experiment.

Experiment	Stage	Kundzuke	Yobisute
1	pre-interaction	25	19
1	roll call	21	10
1	dance	0	2
1	post-interaction	0	1
2	pre-interaction	3	6
2	roll call	6	9
2	storybook	0	6
2	post-interaction	4	11
3	pre-interaction	1	3
3	roll call	0	9
3	dance	0	1
3	transition	0	8
3	storybook	3	6
3	post-interaction	2	3

This could reveal that with repeated exposure to Pepper, many children began to view the robot more as an object than a classmate. It could also signal increased familiarity from some children, and derision from others. “Toiletpepper” continued to get thrown around throughout the three experiments, sometimes with the -kun suffix. At one point, one child cuts across someone shouting “Toiletpepper” by saying “Toumorokoshipepper”, “toumorokoshi” being the Japanese word for corn.

As a group, it is noteworthy that all through the three experiments, there remain children who use the kundzuke form of address, which could signal that in some way they feel the robot is akin to another child, deserving of a certain degree of polite address.

7. Coding with New Data and Coders

While our results of the coding protocol suggested that there was a need for better training of coders and clearer definitions, we also hypothesized that, to some degree, the ambiguity of in-the-wild vocalization data results in variable interpretations between coders. In order to test this hypothesis, we conducted a third round of coding using a different data set. In keeping with our aim that this coding process be accessible to non-expert coder, only one coder from rounds 1 and 2 was consulted, while the other two coders were new to the process. A training process was designed to familiarize the new coders with the process, and the flowchart was omitted in keeping with the conclusion outlined in Section 5 that the presence of both definitions and a flowchart may have been introducing ambiguity to the protocol.

7.1. Methods

These data were taken from two in-the-wild interactions conducted with 4–9 children aged 3–8 in India in the English language in 2019 and recorded on video with audio. Due to the lower number of children and the lack of video restrictions in recording these interactions, these data had a higher quality of both video and audio, and it was possible to make out which individual was speaking in most cases. Three coders all native or fully fluent in English (though none native speakers of Indian English) were recruited: coder 1 from the first two rounds as well as two new coders, coders 4 and 5, who were provided with instructions and the final category definitions as shown in Table 1, with no flowchart, and instructions to code each category as the first letter of the category name (i.e., To = T, About = A, Dialogue = D, etc.). As before, only coder 1 had experience in behavioral coding; coders 4 and 5 did not. Sections of each interaction were selected and manually transcribed for coding and training. Two data sets were created for training. The first training data set consisted of a fake transcript with no corresponding video or audio, 31 entries long; the second training set was taken from a segment of an interaction that would be excluded from analysis: a video-and-audio segment 2 min long consisting of 41 transcript entries. After each round of training, the coders would be consulted about each point of disagreement; if it was a mistake, then it would be corrected, and if it was a difference of interpretation of the audio and video, then it would be left as it was. Following the training, the coders were instructed to code one segment from each of the in-the-wild experiments. Experimental set 1 was a 2 min segment consisting of 62 transcript entries; experimental set 2 was a 1.5 min segment consisting of 59 transcript entries. Segments selected from the experiments for experiments and training sets were chosen from sections with little to no use of Indian-English-specific terms and expressions to minimize any artifacts due to the coders’ lack of experience with Indian English. Both experimental data sets were included in intercoder agreement analysis. Entries where all coders agreed a data point was Unclear were removed before assessing intercoder agreement.

7.2. Results

The results are shown in Table 12. This time, with only the definitions provided and the coders using letters to code the categories, the names of the categories were not mentioned by any coder as a reason for any of their choices. Intercoder agreement percentage was found to be roughly equivalent or a little lower than that seen between three coders in round 2, while ICC value was found to be higher, with the total ICC between all three coders being in the range of 0.5 to 0.75, indicating moderate reliability. The same percentage of agreement was observed between coders 1 and 5 as well as coders 4 and 5, but points of agreement and disagreement between coders continued to vary, as can be seen from the differing ICC values. Taking majority consensus where any two coders agreed led to a rate of 86.27% agreement, lower than the 93.45% agreement observed between three coders in round 2, but with a much higher ICC value that indicates moderate reliability. While the flowchart had been omitted this time to force coders to focus on the definitions, it is possible that some of the specific instructions contained within the flowchart (such as how to code different types of laughter and answers to specific questions) do lead to higher intercoder agreement in terms of direct proportions. However, in conjunction with the interviews conducted during the training sets, we believe that while the definition of subcategories may aid in the process of raising intercoder agreement, to a large degree, the lack of agreement seen in rounds 2 and 3 is the result of lack of clarity or context in in-the-wild data, emphasizing the need for multiple coders in order to establish a consensus on which vocalizations' meanings are highly subjective.

Table 12. Inter-coder agreement analysis of round 3 of coding with 2/3 new coders and new data using percentages and interclass correlation coefficients (ICC).

Coders	Percentage Agreement	ICC
Total agreement	41.18%	0.5110
Coder 1 vs. 4	52.94%	0.4880
Coder 4 vs. 5	57.84%	0.5397
Coder 1 vs. 5	57.84%	0.5114
At least 2 agree	86.27%	N/A

8. Discussion and Lessons

While some in-the-wild studies view the noisy effects of the environment and bystanders as a detraction from the human–robot interaction that is the focal point of the experiment [30], our analysis attempted to incorporate these environmental vocalizations into the interpretation, with moderate success. The definitions used in the first and even second rounds of coding were insufficient, and contributed to the lack of agreement between coders. In order for the behavioral coding protocol to be applied reliably, improvement is needed. While we did not perform a round of coding with the final definitions, we ended with more specific and useful definitions than in round 1.

While some changes between the rounds of coding did bring the ICC value closer to the necessary bare minimum of 0.5, it was only the third round of coding, conducted with new data and new coders, that brought it above that threshold. However, using the ICC also required conversion of the categories to numerical values. While an effort was made to assign numerical values in a way that had some internal logic, they are not designed to be numerical values. Though it is standard to have a single primary coder and a single secondary coder to check agreement, rounds 2 and 3 of coding saw a roughly equivalent rate at which at least 2 of the 3 agreed, which was in the range of 86–93% majority agreement. We attribute this to the chaotic nature of in-the-wild group CRI experiments with little to no structure, and therefore see the need for consensus building between coders as a feature of the subjectivity inherent to the interpretations of some of the vocalizations [23].

After the first round of coding, we cut out the **Reactive** category designed to identify instances where the children's reactions indicate that the robot has done something unexpected. However, other studies have found this useful in vocalization-centric analyses that incorporate only some visual elements [3], indicating that this might be a category worth revisiting, perhaps as more of a focal point than we tried in round 1. The investigation of reactions to unexpected experiences with the robot is an aspect of several in-the-wild studies [31,32], and while that can be performed without reliance on this particular behavioral coding scheme, if the categories were to be redefined for a future study, the reintroduction of the **Reactive** category may be worth consideration.

Another aspect to consider is that this coding protocol was conducted without any training set or training process for the coders. The coders were simply introduced to the protocol and asked to code the experiments. After the initial 33.82% agreement between coders 1 and 2 in the first round, they exhibited 60.70% agreement in the second round. While some of this may be attributed to the redefinition of the categories, the training they experienced in the first round and the discussion after likely also had an effect. With more training and clearer definitions applied as shown in Table 1, we believe the 93% agreement rate between any two of the three coders in round 2, as well as the 86% agreement rate for the same conditions in round 3, is evidence that there can be more agreement and consistency between coders. Furthermore, the results of the round of coding attempted with new data and coders detailed in Section 7 may demonstrate that with training and clearer definitions, ICC values could be raised to the range of moderate reliability, despite the lower overall agreement percentage. However, the nature of in-the-wild interactions is that there is often a lack of clarity on the context of a given vocalization, and this introduces an unavoidable element of subjectivity.

Use of multiple coders in cases where there is unavoidable subjectivity is an established method in behavioral coding [23]. While there is some inherent subjectivity in language use [27], the in-the-wild group interaction setting adds further ambiguity to the type of data we wish to analyze, as with so many children present, the context of many vocalizations are not clear even with video. While subjective measures can be assessed by taking the average of a numerical value given by multiple coders [33], as we are not using numerical values, we recommend majority consensus. Going forward, even if agreement is assessed by majority out of at least three non-expert coders, it is still necessary to consider each coder's experience level, and train the coders as needed before the experiment. In round 3, we found two short rounds of training with feedback after each round to be sufficient in curbing cases where coders misunderstand the process or the definitions of each category. With majority agreement analysis, even if there is not 100% agreement, analysis could discount data points where all three coders disagree in the same manner as those which are coded as **Unclear**. Additionally, the number of coders could be increased in order to enhance the reliability of results.

While we did base the coding categories upon the same sorts of principles applied in wild dolphin studies, the resulting protocols are very different in a number of ways. In the protocol used for wild dolphins, there are only a few visual factors that researchers need to observe: speed and direction of movement, presence of fish and fowl, and the distribution of the dolphin pod [12]. There are cases of ambiguity, especially for a coder new to the protocol, but there is no point at which a category could not be applied to the observation of a visible dolphin pod, though multiple categories might be required. In contrast, the vocal behavior of humans has far more ambiguity, and cannot always fit into one of the designated categories. The biggest difference comes from the fact that one requires observing the behavior of the group as a whole, while the other isolates and codes individual behavior and attempts to code it in the context of the group. This difference is rooted in visual versus auditory behavior observation, as it is more difficult to assess the behavior of a group as a whole using an audio medium. An alternate audio-based behavioral coding approach might be to define categories of overall group behavior over fixed units of time, such as silence, non-robot related cacophony, robot-related cacophony, etc. While breaking interactions up into units of time rather than by individual vocalizations might introduce a higher degree of subjectivity, such an approach would likely also require more training, and therefore could yield more reliable results.

Beyond the behavior coding protocol, however, vocalization-based behavioral analysis offers the opportunity for further studies with children circumstances where filming is not possible or practical. Here, we have only presented a simple analysis of incidental behaviors identified through the three experiments, and an analysis of how the children referred to the robot with suffixes particular to Japanese language and culture. Our results consistently showed that, through the experiments, the children increasingly treated the robot less like a classmate and more like an object. This finding would be consistent with other studies, which have found that as the novelty effect wears off, the users' more science-fiction-derived expectations of the robot abate [31]. These observation-based analyses are qualitative in nature, and are more useful in describing individual events or series of events than they are at effectively contrasting different experiments or producing reproducible results. One of the strengths of in-the-wild studies is the variety of unexpected behaviors and reactions that might be encountered, and there is value to using qualitative methods to describe them; however, in both the psychological sciences and the tech industry, more quantitative approaches are generally considered. The qualitative analysis of an individual event or series of events may be more akin to a sociological approach [28], which may be an approach worth considering for more in-the-wild robot interaction studies, especially as robots become more common among humans in workplace settings [34].

This is not intended an exhaustive list of the possibilities in vocalization analysis. For instance, in English, one might look at the use of gendered pronouns related to the robot. Unlike Japanese, English is a language where pronouns are frequently necessary, which provides more opportunity to assess if the robot is being spoken of as an object, or if it is being gendered and how. There are many other approaches that could be explored, including the impacts of the robot's language on the children's use of language and imitation behaviors [35–38], as well as cases when the children cannot understand the robot's accent or vice versa [10]. While this approach to analysis is not the ideal for all group CRI in-the-wild analyses, such as those where there are better video data and clear one-to-one interactions with the robot, or in which individuals can be differentiated from one another [31,39], we believe it could be useful particularly in those cases where language is already a point of focus [3,6,10]. In labs with access to coders with a higher level of expertise in behavioral studies, perhaps these categories could provide the basis for the definition of categories in such a way as could lead to higher rates of intercoder agreement. In recent years, machine learning approaches have been utilized in behavioral coding as an alternative to the use of human coders [40,41], and with the progress in large language models, this could be an approach used to analyze vocalization data as well.

9. Conclusions

Our vocalization-centric behavioral coding protocol was designed to assess general engagement with the robot in in-the-wild interactions for robotics labs without experienced coders. We were not able to reach a high rate of agreement between two coders. Based on the rate of majority consensus observed between three coders, we believe that the combination of our use of non-expert coders and the inherent subjectivity of group in-the-wild data warrants an approach that relies on building consensus between multiple coders. In addition to the use of multiple coders, we recommend two alterations to the protocol to be applied in the future: emphasis of the provided definitions of the coding categories, and the use of specific subcategories to aid the coders. We also have shown several ways in which verbal behavior can be analyzed, alongside or in the absence of visual data of child–robot interactions, highlighting this approach as a potential direction for future CRI studies in the wild.

10. Ethics Statement

Both sets of experiments were performed as opt-in events with parents present. Parents were informed in advance that this would be an experiment in which we would observe the children's responses to the robots, and that the video recordings would be for internal use only, with no identifying screenshots or identifying information released to the public. Furthermore, events were conducted with full priority given to the wishes and requests of the school/hosting location, including the structure of the event, the types of activities, and location of the camera. Because the events were opt-in events designed in collaboration with the host location, with the parents informed and present, with no identifying information intended to be collected (participants were anonymous to the researchers), and what identifying information was incidentally collected not to be released, by Japanese law, as well as the ethics protocols of our institution in the years when these events were conducted (2018 and 2019), this study did not meet the threshold required for review by an ethics committee. These ethical practices are the same as seen in other publications, such as [2,9].

Author Contributions: Conceptualization, X.I. and G.V.; methodology, X.I.; data curation, X.I.; formal analysis, X.I.; investigation, X.I. and G.V.; resources, G.V.; writing—original draft preparation, X.I.; writing—review and editing, X.I. and G.V.; visualization, X.I.; supervision, G.V.; project administration, G.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from parents of all subjects involved in the study with the understanding that no identifying information of any participant would be released.

Data Availability Statement: The datasets presented in this article are not readily available to protect the privacy of the children involved in the experiments as per our ethical guidelines. Requests to access the datasets should be directed to GVLab (http://gvlab.jp/ accessed on 17 March 2024).

Acknowledgments: Thank you to Takamune Izui, for conducting the preschool event described and for participating in the coding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tanaka, F.; Movellan, J.R.; Fortenberry, B.; Aisaka, K. Daily HRI evaluation at a classroom environment: Reports from dance interaction experiments. In Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI), Salt Lake City, UT, USA, 2–3 March 2006; pp. 3–9. [CrossRef]
2. Venture, G.; Indurkhya, B.; Izui, T. Dance with Me! Child-Robot Interaction in the Wild. In Proceedings of the Ninth International Conference on Social Robotics (ICSR), Tsukuba, Japan, 22–24 November 2017; pp. 375–382. [CrossRef]
3. Wróbel, A.; Źróbek, K.; Schaper, M.M.; Zguda, P.; Indurkhya, B. Age-Appropriate Robot Design: In-The-Wild Child-Robot Interaction Studies of Perseverance Styles and Robot's Unexpected Behavior. In Proceedings of the 32nd IEEE International Conference on Robot & Human Interactive Communication (Ro-Man 2023), Busan, Republic of Korea, 28–31 August 2023; pp. 1451–1458. Available online: https://arxiv.org/abs/2310.12899 (accessed on 17 March 2024).
4. Nomura, T.; Uratani, T.; Kanda, T.; Matsumoto, K.; Kidokoro, H.; Suehiro, Y.; Yamada, S. Why Do Children Abuse Robots? In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Portland, OR, USA, 2–5 March 2015; pp. 63–64. [CrossRef]
5. Tobis, S.; Neumann-Podczaska, A.; Kropinska, S.; Suwalska, A. Unraq—A questionnaire for the use of a social robot in care for older persons. A multi-stakeholder study and psychometric properties. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6157. [CrossRef]
6. Sienkiewicz, B.; Sejnova, G.; Gajewski, P.; Vavrecka, M.; Indurkhya, B. How language of interaction affects the user perception of a robot. In Proceedings of the International Conference on Social Robotics (ICSR2023), Doha, Qatar, 3–7 December 2023; Ali, A.A., Ed.; pp. 308–321. Available online: https://arxiv.org/abs/2310.15321 (accessed on 17 March 2024).
7. Serholt, S.; Pareto, L.; Ekström, S.; Ljungblad, S. Trouble and Repair in Child–Robot Interaction: A Study of Complex Interactions With a Robot Tutee in a Primary School Classroom. *Front. Robot. AI* **2020**, *7*, 46. [CrossRef]
8. Diaz-Boladeras, M.; Paillacho, D.; Angulo, C.; Torres, O.; Gonzalez-Dieguez, J.; Albo-Canals, J. Evaluating Group-Robot Interaction in Crowded Public Spaces: A Week-Long Exploratory Study in the Wild with a Humanoid Robot Guiding Visitors Through a Science Museum. *Int. J. Humanoid Robot.* **2015**, *12*, 1550022. [CrossRef]
9. Coronado, E.; Indurkhya, X.; Venture, G. Robots Meet Children, Development of Semi-Autonomous Control Systems for Children-Robot Interaction in the Wild. In Proceedings of the 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM), Toyonaka, Japan, 3–5 July 2019; pp. 360–365. [CrossRef]
10. Singh, D.K.; Kumar, M.; Fosch-Villaronga, E.; Singh, D.; Shukla, J. Ethical Considerations from Child-Robot Interactions in Under-Resourced Communities. *Int. J. Soc. Robot.* **2023**, *15*, 2055–2071. [CrossRef] [PubMed]
11. Altmann, J. Observational study of behavior: Sampling methods. *Behaviour* **1973**, *49*, 227–267. [CrossRef]
12. Mann, J. Behavioral sampling methods for cetaceans: A review and a critque. *Mar. Mammal Sci.* **1999**, *15*, 102–122. [CrossRef]
13. Clarkson, J.; Christiansen, F.; Awbery, T.; Abbiss, L.; Nikpaljevic, N.; Akkaya, A. Non-targeted tourism affects the behavioural budgets of bottlenose dolphins Tursiops truncatus in the South Adriatic (Montenegro). *Mar. Ecol. Prog. Ser.* **2020**, *638*, 165–176. [CrossRef]

14. Hayes, A.T.; Hughes, C.E.; Bailenson, J. Identifying and Coding Behavioral Indicators of Social Presence With a Social Presence Behavioral Coding System. *Front. Virtual Real.* **2022**, *3*, 773448. [CrossRef]
15. Arhant, C.; Beetz, A.M.; Troxler, J. Caregiver reports of interactions between children up to 6 years and their family dog-implications for dog bite prevention. *Front. Vet. Sci.* **2017**, *4*, 130. [CrossRef] [PubMed]
16. Borelli, J.L.; Lai, J.; Smiley, P.A.; Kerr, M.L.; Buttitta, K.; Hecht, H.K.; Rasmussen, H.F. Higher maternal reflective functioning is associated with toddlers' adaptive emotion regulation. *Infant Ment. Health J.* **2021**, *42*, 473–487. [CrossRef]
17. Calkins, S.D.; Smith, C.L.; Gill, K.L.; Johnson, M.C.; Maternal, M.C. Maternal Interactive Styles across Contexts. *Soc. Dev.* **1998**, *7*, 350–369. [CrossRef]
18. Ros, R.; Nalin, M.; Wood, R.; Baxter, P.; Looije, R.; Demiris, Y.; Belpaeme, T.; Giusti, A.; Pozzi, C. Child-robot interaction in the wild: Advice to the aspiring experimenter. In Proceedings of the the 2011 ACM International Conference on Multimodal Interaction, Alicante, Spain, 14–18 November 2011; pp. 335–342. [CrossRef]
19. Pesch, M.H.; Lumeng, J.C. Methodological considerations for observational coding of eating and feeding behaviors in children and their families. *Int. J. Behav. Nutr. Phys. Act.* **2017**, *14*, 170. [CrossRef]
20. Schaper, M.M.; Márquez Segura, E.; Malinverni, L.; Pares, N. Think-4-EmCoDe framework: Highlighting key qualities in embodied co-design techniques for children. *Int. J. Hum. Comput. Stud.* **2023**, *177*, 103065. [CrossRef]
21. Syrdal, D.S.; Dautenhahn, K.; Robins, B.; Karakosta, E.; Jones, N.C. Kaspar in the wild: Experiences from deploying a small humanoid robot in a nursery school for children with autism. *Paladyn* **2020**, *11*, 301–326. [CrossRef]
22. Ribarič, D.; Clarkson, J. Nautical tourism affects common bottlenose dolphin (Tursiops truncatus M.) foraging success in a NATURA 2000 site, North-Eastern Adriatic Sea. *Mediterr. Mar. Sci.* **2021**, *22*, 285–296. [CrossRef]
23. Chorney, J.M.L.; McMurtry, C.M.; Chambers, C.T.; Bakeman, R. Developing and modifying behavioral coding schemes in pediatric psychology: A practical guide. *J. Pediatr. Psychol.* **2015**, *40*, 154–164. [CrossRef] [PubMed]
24. Bartneck, C.; Nomura, T.; Kanda, T.; Suzuki, T.; Kato, K. A cross-cultural study on attitudes towards robots. In Proceedings of the HCI International, Las Vegas, NV, USA, 22–27 July 2005; pp. 1981–1983.
25. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
26. Haidet, K.K.; Tate, J.; Divirgilio-Thomas, D.; Kolanowski, A.; Happ, M.B. Methods to Improve Reliability of Video Recorded Behavioral Data. *Res. Nurs. Health* **2009**, *32*, 465–474. [CrossRef] [PubMed]
27. Boyd, R.L.; Schwartz, H.A. Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *J. Lang. Soc. Psychol.* **2021**, *40*, 21–41. [CrossRef]
28. Babbie, E. *The Practice of Social Research*, 12th ed.; Cengage Learning: Wadsworth, OH, USA, 2010.
29. Michaelis, J.E.; Cagiltay, B.; Ibtasar, R.; Mutlu, B. "Off script:" Design opportunities emerging from long-term social robot interactions in-the-wild. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI '23), Stockholm, Sweden, 13–16 March 2023; pp. 378–387. [CrossRef]
30. Neerincx, A.; Edens, C.; Broz, F.; Li, Y.; Neerincx, M. Self-Disclosure to a Robot "In-the-Wild": Category, Human Personality and Robot Identity. In Proceedings of the RO-MAN 2022—31st IEEE International Conference on Robot and Human Interactive Communication: Social, Asocial, and Antisocial Robots, Naples, Italy, 29 August–2 September 2022; pp. 584–591. [CrossRef]
31. Ahtinen, A.; Beheshtian, N.; Väänänen, K. Robocamp at home: Exploring families' co-learning with a social robot: Findings from a one-month study in the wild. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI '23), Stockholm, Sweden, 13–16 March 2023; pp. 331–340. [CrossRef]
32. Wróbel, A.; Źróbek, K.; Indurkhya, B.; Schaper, M.M.; Gunia, A.; Zguda, P.M. Are robots vegan? Unexpected behaviours in child-robot interactions and their design implications. In Proceedings of the CHI EA '23: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–7. [CrossRef]
33. Blöte, A.W.; Miers, A.C.; Westenberg, P.M. The role of social performance and physical attractiveness in peer rejection of socially anxious adolescents. *J. Res. Adolesc.* **2015**, *25*, 189–200. [CrossRef]
34. Dobrosovestnova, A.; Hannibal, G.; Reinboth, T. Service robots for affective labor: A sociology of labor perspective. *AI Soc.* **2022**, *37*, 487–499. [CrossRef] [PubMed]
35. van den Berghe, R. Social robots in a translanguaging pedagogy: A review to identify opportunities for robot-assisted (language) learning. *Front. Robot. AI* **2022**, *9*, 958624. [CrossRef]
36. Rohlfing, K.J.; Altvater-Mackensen, N.; Caruana, N.; van den Berghe, R.; Bruno, B.; Tolksdorf, N.F.; Hanulíková, A. Social/dialogical roles of social robots in supporting children's learning of language and literacy—A review and analysis of innovative roles. *Front. Robot. AI* **2022**, *9*, 971749. [CrossRef]
37. Sommer, K.; Davidson, R.; Armitage, K.L.; Slaughter, V.; Wiles, J.; Nielsen, M. Preschool children overimitate robots, but do so less than they overimitate humans. *J. Exp. Child Psychol.* **2020**, *191*, 104702. [CrossRef]
38. Peter, J.; Kühne, R.; Barco, A. Can social robots affect children's prosocial behavior? An experimental study on prosocial robot models. *Comput. Hum. Behav.* **2021**, *120*, 106712. [CrossRef]
39. Kim, Y.; Chen, H.; Algohwinem, S.; Breazeal, C.; Park, H.W. Joint Engagement Classification using Video Augmentation Techniques for Multi-person HRI in the wild. In Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, London, UK, 29 May–2 June 2023; pp. 698–707. Available online: https://arxiv.org/abs/2212.14128 (accessed on 17 March 2024) .

40. Wu, C.; Liaqat, S.; Helvaci, H.; Chcung, S.C.S.; Chuah, C.N.; Ozonoff, S.; Young, G. Machine learning based autism spectrum disorder detection from videos. In Proceedings of the 2020 IEEE International Conference on E-Health Networking, Application and Services, HEALTHCOM 2020, Virtual, 1–2 March 2021. [CrossRef]
41. Bennett, C.C.; Stanojević, C.; Šabanović, S.; Piatt, J.A.; Kim, S. When no one is watching: Ecological momentary assessment to understand situated social robot use in healthcare. In Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21), Virtual, 9–11 November 2021; pp. 245–251. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.