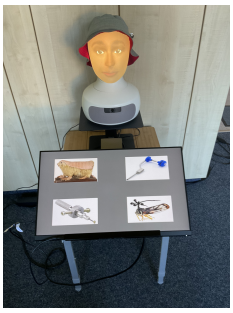
Children's Word Learning from Socially Contingent Robots under Active vs. Passive Learning Conditions

Fatih Sivridag
University of Göttingen
Göttingen, Germany
Leibniz Science Campus Primate Cognition
Göttingen, Germany
fatih.sivridag@uni-goettingen.de

Nivedita Mani
University of Göttingen
Göttingen, Germany
Leibniz Science Campus Primate Cognition
Göttingen, Germany
Nivedita.Mani@psych.uni-goettingen.de

(a) Furhat in the classroom for introduction

(b) Furhat with stimulus



(c) Furhat with question

Figure 1: Different stages of the experimental task.

ABSTRACT

Language is learned through social interactions, in which gaze has a special role because it can be used to guide the attention and reference objects easily. Children, starting from very early ages, are also very good at utilizing gaze to map labels to referenced objects. To achieve language teaching robots, we need to understand how these functions of gaze can be implemented most efficiently. To this aim, we allowed children to interact with a social robot to learn the labels of several objects in a naturalistic setting. In some trials the child guided the gaze and chose the object to be learned while the robot was following and in the others they changed the roles and robot guided the gaze and decided on the object to be learned. We measured how much children actually followed the robot's gaze and how many words they learned in these two conditions, referred to as active and passive learning conditions, respectively. The results indicate that although children followed the robot's gaze and learned words successfully, there were no meaningful differences in word learning between the two conditions. The rate of gaze following and time spent looking at the robot did not influence word learning, either. The implications of these results for use of robots in educational settings are further discussed.

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03.
https://doi.org/10.1145/3610977.3634931

CCS CONCEPTS

• **Social and professional topics** → **Children**; • **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → *Gestural input*.

KEYWORDS

Child-robot interaction, gaze, active learning

ACM Reference Format:
Fatih Sivridag and Nivedita Mani. 2024. Children's Word Learning from Socially Contingent Robots under Active vs. Passive Learning Conditions. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3610977.3634931

1 INTRODUCTION

Language is a social tool, which is primarily acquired through social interactions [15]. To achieve the promise of language teaching robots, these devices should be able to similarly sustain seamless social interactions. We, therefore, need further understanding of the factors that influence the quality of human-robot interactions, especially in early childhood. In particular, in order to ensure that robots attain social interactions that are qualitatively similar to human-human interactions, we must better understand the role of non-verbal communication in HRI. This is especially true for children for whom socio-communicative input in language learning is vital [44].

One of the non-verbal communication tools that enrich social interactions and has proved critical for language learning is gaze (c.f. [13]). Thus, adult gaze is considered an ostensive cue that signals to the child that the information being discussed is relevant

and meaningful [18], with children attending more to information accompanied by ostensive cues [38]. At the same time, learning outcomes are also improved when adults follow the gaze, and consequently, attention of the child, and label objects that the child is attending to [33, 42, 43]. Against this background, the current study examines the role of gaze in child-robot interactions, with regard to children's attention to and learning about objects that the robot is attending to versus objects where the robot follows the child's gaze and, consequently, attention.

2 BACKGROUND

2.1 Joint Attention in HRI

Considering the robustness of gaze following [39], humans are very good at attending to where the robot is looking at under various conditions [11]. For instance, a target cueing advantage, i.e., faster shift of gaze to a target indicated by another human's gaze, has been replicated in HRI [14, 30, 46]. Human-robot collaborative tasks are also completed more efficiently when the robot gaze is contingent with the task requirements [23].

Given the success of gaze following in HRI, subsequent research has focused on the factors that influence how efficiently humans follow robots' gaze. For example, although virtual agents' gaze information has been shown to improve task success in a card matching game, it provides low precision relative to embodied agents [8]. With embodied robots, objects can be referenced in 3D space with eye and head movements. Indeed, humans are better at following gaze in HRI when the robot moves its eyes and head in coordination instead of signaling gaze location with only head movements [6].

Nevertheless, studies comparing gaze following in human-human interactions and HRI report some critical differences. For instance, participants look at the face of the agent more than the target when interacting with a robot compared to a human [47]. This distracting effect might have significant implications especially for the use of social robots in children's language learning, where attending to the referent at the correct time has been shown to affect learning outcomes [48]. Participants also appear to be able to suppress the reflexive cueing effect in HRI [3]. Thus, in tasks requiring participants to direct their attention in a specific direction (cued by an arrow), participants' gaze reflexively orients to social cues such as a human partner looking in the opposite direction. In HRI, however, while participants were able to infer the direction of the robots' gaze, they did not reflexively orient to the location of the robot's gaze. The authors concluded that, while robot gaze may be as informative as human gaze at higher levels of processing, human gaze may have a special place at lower levels of sensory processing, e.g., reflexive attention, which they attribute to the different processing pathways for faces reported in neuroscience studies [16]. Furthermore, studies suggest that the anthropomorphism of the robot may influence attentional strategies, with participants inferring the intentions of more anthropomorphic robots better than less anthropomorphic robots [26, 27].

2.2 Joint attention in child-adult interactions

There has been much work on the development and influence of joint attention between children and their adult social partners on the quality of caregiver-child interactions and learning outcomes. These studies find that learning is boosted when the mature partner follows the attention of the child and labels an object that the child is directing attention to, relative to when the child's attention is recruited to an object before the partner labels this object [33, 42, 43, 48].

Thus, in one of the first studies on the subject, the authors observed mother-child interactions over a period of five months and examined the vocabulary development of the children [43]. The results suggest that children in the dyads where children were leading episodes of joint attention showed increased vocabulary gains relative to children whose mothers guided the child's attention more often. In another study, an experimenter introduced novel objects to children either while their attention was on the object or when they were attending to something else [42]. As in the observational study, children learned the label of an item better when the naming event occurred while they were already attending to the target item compared to the situations where children's attention shifted to the target item after the naming event occurred.

Such findings are typically explained with recourse to the demands on the child in ambiguous naming situations. Thus, in situations where many objects are simultaneously present in the visual field of the child, if caregivers label the object that the child is already attending to, the child does not need to infer which of the many objects in its visual field is the referent of the label [33]. Furthermore, labelling events where the child is already attending to the object are cognitively less demanding for the young learner, with fewer demands on their motor skills, cognitive control and selective attention.

While the studies reported above were mostly conducted with infants and very young children (up to 24-months of age), studies with older samples report mixed results. Such studies have typically focused on differences in learning outcomes when children actively choose the information to-be learned relatively to when they passively receive such information. Ackermann et al. [2020] report a passive boost in learning with 3-year-olds showing reduced learning of words in an active context, i.e., when they actively chose the objects whose label they wanted to hear, relative to a passive context, i.e., where they were presented with the labels of objects another child in a previous session had chosen to hear [2] (but see Partridge et al., [2015] for opposing results with 3- to 5-year-olds[31]). In keeping with this, Foushee et al. [2021] find that children's ability to learn words and facts actively improves across early childhood, matching learning from the passive condition (where they were explicitly told the label of an object) between 4.5- to 6-years of age [17]. Ruggeri et al. [2019] report that the memory boost associated with providing children active control over their study environment is small around 5-years, increases with age and is adult-like around 8-years of age [37]. In contrast, other studies with 7-year-olds report improved learning performance when children choose the item to-be learned compared to when the item was chosen by their interaction partner [40]. Taken together, the research reviewed in this section suggests that a) children show improved learning in child-adult interactions when adults follow the attention of the child and label objects that they are already attending to and b) there are, nevertheless, subtle nuances to children's active learning outcomes especially across early development that

require greater consideration. In what follows, we examine the extent to which such insights from research on early development have been capitalised on in HRI.

2.3 Joint attention in cHRI

Despite the abundance of research on the effects of different gaze behaviour in HRI conducted with adult participants, there are not many studies investigating how children respond to robot's gaze. The research in this area focuses mostly on children with Autistic Spectrum Disorder (ASD) and how these children respond to human gaze vs. robot gaze and the use of social robots as tools in therapy [21, 41, 45, 49]. Although typically developing children are recruited as control groups to whom the performance of children with ASD is compared, some informative observations about typically developing children's behaviour have also been made. For example, in a study investigating differences in responses to robot's joint attention elicitation attempts between typically developing children and children with ASD, typically developing children established joint attention with the robot 75% of the times [7]. In another study with typically developing and ASD samples, both groups looked at the robot's face more and cued targets less in their interactions with a robot compared to a human [12].

Nevertheless, the few studies conducted with typically developing children suggest that, similar to adults, children follow a robot's gaze to establish joint attention [4]. In one of the first studies investigating whether children successfully follow a robot's gaze, Movellan and Watson [2002], showed that even infants establish joint attention with a socially contingent robot similar to a human, concluding that humans generalize contingency to non-human social partners [24].

In another study done by Okumura et al. [2013], similarly, infants followed both the human and robot partners' gaze [29]. However, they learned significantly more when the partner was human compared to the robot. Finally, in another study with an older sample (6- to 11-year-olds) the authors used a NAO robot which provided hints about the target card in a memory card [28]. Children who noticed that the robot was trying to provide hints established mutual gaze (i.e., looked at each other's face) and followed the robot's gaze significantly more than the children who did not notice that the robot was trying to help. Furthermore, children required fewer attempts to find the correct card in the trials where the robot provided hints, showing that children used robot's gaze information in their decision making.

2.4 Current study

Most of the studies reviewed above have examined the extent to which children follow the robot's gaze and how this influences subsequent learning and decision making. Also, the robots employed in these studies had limited gazing capabilities; due to the fewer degrees of freedom in head and/or eye movements. In early development, however, studies suggest that following the child's gaze can boost learning outcomes in child-adult interactions and that children learn better when they actively control their learning environment. Against this background, the current study examines joint attention and learning outcomes when children follow the gaze of the robot, i.e., attend to and learn about objects that the robot is attending to, relative to when the robot follows the child's gaze and attends to and labels an object the child is attending to.

In particular, we allowed children to interact with Furhat, a back projected talking head, which assisted the child through a learning task, where the child was given the opportunity to learn the labels of a number of different objects. In half of the trials, Furhat chose one of the objects displayed on a screen, fixated this object (with head and eye movements) and then provided the child with more information about this object (passive learning). In the other half of the trials, Furhat followed the child's gaze (ascertained via an external eye-tracker) and provided the child with more information about one of the objects on the screen when the child fixated this object for longer than a pre-set threshold (active learning). Although active and passive learning are operationalized distinctly across disciplines, in keeping with the literature in developmental psychology and cognitive science, we refer to active learning in terms of situations where children elicit further information about an object actively – here via their initiation of joint attention – while passive learning refers to situations where children are provided with information about an object chosen by someone else, here those trials where Furhat initiates joint attention to an object. To increase the ecological validity of our investigation, we collected data in schools, as part of the afternoon activities of children while some level of typical background school noise was present. Our research questions (RQs) and corresponding hypotheses (Hs) are:

2.5 Research Questions

RQ1. Do children follow a robot's gaze in an interactive task in naturalistic settings?

H1. Considering the robustness of gaze following in human-human interactions and the findings of gaze studies in HRI, we expect the children to follow the robot's gaze on the screen (as operationalized by recurrence analysis [35]).

RQ2. Are there differences in learning outcomes when the robot provides children with information about an object the child is already attending to (active learning), relative to when the robot chooses the object that children are provided with more information about (passive learning)?

H2. Given the literature suggesting an active learning benefit when children actively control their learning environment reviewed above, we expect to find better word learning outcomes in active learning trials compared to passive learning trials.

RQ3. Do children attend more and longer to items that they actively choose to hear more about (active learning) relative to when the robot chooses the object to provide more information on (passive learning)?

H3. We expect children to look longer at the objects in active learning trials compared to passive learning trials, which may explain the previously reported active learning boost.

RQ4. Do children who follow the robot's gaze more also show improved learning performance in passive learning condition?

H4. As robot's gaze signals the item-to-be-learned and allows the child to allocate attentional resources accordingly, we expect to see an increase in learning performance with increased gaze following.

3 METHODS

3.1 Participants

A total of 96 children participated in the study. The children were recruited from local primary schools and were in first to fourth grades. Thus, all children received primary schooling in German and were regarded as being proficient in German. Demographic information pertaining to children's background and development could not be collected due to data being collected at the schools. 9 children were excluded from all the analyses due to technical problems (n = 8) or fussiness (n = 1) during the experiment leaving 87 children (47 females) with usable data. Age of the children ranged from 72 to 130 months ($Median_{age}$ = 97 months, IQR_{age} = 23 months) The study design was approved by the institute's ethics committee and necessary permits were obtained from the school administrations before we started testing. In addition, legal guardians of the children tested provided informed signed consent before their child participated in the study. Each child received a certificate of attendance and a printed picture of themselves with the robot, if the parent had given us consent for photos of their child to be taken.

3.2 Apparata and Stimuli

Furhat (Figure 1) is a robot head which back projects its face to a semi-translucent mask [5]. It can move its head along three axes. As its face is an image generated by a computer, it can exhibit human like eye-movements, synchronize lip movements to speech, and produce facial expressions and gestures such as blinking, winking, nodding, etc. It uses third-party text-to-speech and speech-to-text services to generate speech and transcribe user's speech to text. We programmed the behaviour of the robot using furhat-remote-desktop-API python package [36].

The robot and the child mutually interacted with a 24-inch, 1920x1080 pixel resolution touch screen which lay between the robot and the child, at approximately 45° angle; facing the child (Figure 1a). The robot and the screen were placed so as to give the impression that the robot could see what was on the screen from above. A Tobii pro X3-120 portable eye tracker was placed at the bottom of the screen to determine where the child was looking at during the experiment. The eye tracker was connected to a computer which sent head movement and speech commands to the robot so that the robot could follow the gaze of the child. We used the Tobii-research python package [1] to collect gaze data from the eye-tracker and the PsychoPy package (version 2022.1.2) [32] to control stimulus presentation and data collection from the children.

Stimuli were images of 12 pairs of objects (768x432 pixels) against a white background (see Figure 1b). Except for two pairs of objects used in practice trials, all objects were either exotic animals or hand tools. These objects were chosen after piloting with 5 children and discussion with their parents before data collection started. Also, potential confounding effects of prior knowledge was minimized by our randomisation measures. During training, each object was named 3 times embedded in simple carrier phrases beginning with the phrase "This is a/an .....". Subsequent sentences then provided brief information about the object such as where it lives (animals) or what it could be used for (tools). We used Amazon Polly's "Vicki-Neural" as the robot's voice.

3.3 Procedures

Children participated in the study in a separate, but not sound proof room at their schools, either during afternoon care or holiday care. Prior to the experimental task, children attended a short introductory session in which the robot introduced itself and explained its capabilities. Depending on the availability of the children, the introduction session was either carried out individually or in groups of up to five children.

After the introduction, each child did the experimental task individually. They sat on a chair in front of the touchscreen and the robot. After greeting the child and calibrating the eye-tracker, the robot instructed the child that they would see some pictures on the screen. The robot also told the child that, in some cases, the robot would choose what to talk about (passive learning) while, in others, the child could choose (active learning) just by looking at the image they wanted to know more about. This was followed by two practice trials; one where the robot chose the target item and one where the child chose the target object. Next, children completed ten training trials where they could choose to hear more about specific objects on the screen. Five active and five passive learning trials were randomly distributed across ten trials, with the robot telling the child at the beginning of each trial, whose turn it was to choose.

Passive trials, where the robot chose the target, began with two images appearing on the screen side by side and the robot looking at the child's face. Then, the robot attended to either one of the pictures or the middle of the screen randomly, with weights for the pictures being 3 and middle of the screen 0.5. Each fixation of the robot lasted either 1, 1.4, or 1.8 seconds, randomly assigned to fixations. In each trial, the robot had between 2 to 3 such fixations, randomly distributed across trials. Finally, the robot randomly chose one of the images, fixated on it for 1 second, and started talking about the object. The set-up, therefore, gave the impression that the robot was considering which of the images to talk about. Children's eye movements were recorded from the beginning of the trial until the robot started speaking.

Active trials, where the child chose the target, began the same way. Eye-tracking data were down sampled from 120 Hz to 36 Hz, with the average of last 10 fixation points on the horizontal axis used to determine whether the child was looking to the left or right of the screen. This was then fed to the robot such that the robot then attended to the same side of the screen as the child. Down sampling, averaging, and having only two areas of interest (left and right of the screen) prevented jerky head movements of the robot and simulated smooth gaze following. After the child had looked at the screen continuously for 5 seconds, we evaluated whether the child looked to the right or left side of the screen during the last 500ms. The robot then started talking about the object to the side of the screen the child was fixating. If gaze shifted between left and right during the last 500 ms, gaze data were collected for an additional second with the last 500 ms of gaze evaluated the same way. When the eye tracker did not register child's gaze due to calibration issues, an experimenter manually entered where the child was fixating using the keyboard and the robot started talking about chosen item.

After all trials were completed, the child had a chance to take a break, duration of which was determined by the participant. Next, the vocabulary task was presented in which four items that had been shown during the training trials appeared on the screen (Figure 1c). The robot then asked the child to tap on the item it named. The items in each question were two pairs from two trials. The robot did not give any feedback once the child answered. Finally, the robot thanked the child and concluded the experiment.

3.4 Data analysis

RQ1. To examine whether children follow the robot's gaze, we looked at the recurrence scores of the child and robot. That is, we calculated the proportion of time in each passive learning trial where the child and the robot were looking at the same picture. To achieve this, we first calculated the number of frames required for the robot to complete a gaze shift from one side of the screen to the other and from the middle to the sides using a video of a mock session of the experiment. We also estimated the delay between the robot receiving the attend command and actuating the gaze shift. Although the robot's eyes move before the head, we took the beginning of head movement as the start of gaze shift. Using the time stamps of attend commands and data log of attention durations, we generated gaze data of the robot with a sampling rate of 1000 Hz in the same coordinate system as the eye-tracking data obtained from the child. Then, we synchronized the robot's and child's gaze relative to the start of each trial and determined where they were looking at each time point.

To examine whether child and robot gazes remained on the same picture significantly longer in our data compared to a scenario where the child and the robot were looking at the screen randomly, we conducted a permutation analysis. We excluded trials where children looked at the screen less than 10% of the whole trial duration (n = 4; 2 females) with the assumption that this was due to calibration issues. We randomly shuffled the robot's fixation within each trial for each participant 999 times and calculated recurrence scores on the randomly shuffled data. For each run of the permutation, we first averaged the recurrence scores for each participant and then across all participants to obtain 999 recurrence scores. Actual recurrence scores obtained from the original data were also represented in the permuted data; yielding 1000 recurrence scores in total. To retain the auto-correlated nature of the eye-movements, we shuffled blocks of fixations rather than shuffling individual gaze points. We then calculated the probability that the actual recurrence scores came from a random distribution by using the equation below:

$$p = \frac{\sum I(R_i \geq A)}{N}$$

Where:

- $I(R_i \geq A)$ is an indicator function that equals to 1 if the recurrence score (R_i) is greater than or equal to the actual recurrence score (A), and 0 otherwise.
- N is the total number of recurrence scores obtained from permutations

RQ2. To examine whether children learned more from active learning trials relative to passive learning trials, we ran a Generalized Linear Model. Children's responses at test were binary coded (1: correct, 0: incorrect) and entered as the response variable in the model. Learning condition (active vs. passive), age (in months) and their interaction were entered as predictors. For this and subsequent models, the random effects structure and transformations are detailed below.

RQ3. To examine whether children looked at the objects on the screen more during active learning trials compared to the passive learning trials, we fitted a model with the proportion of looking to the target item on screen as the response variable and learning condition (active, passive), age and their interaction as predictors.

RQ4. To examine the relationship between recurrence of gaze and learning performance, we fitted another Generalized Linear Model with children's responses at test (binary coded as above) as the response variable and recurrence proportion, age and their interaction as predictors. Since the robot only led eye gaze in passive trials, this analysis was conducted on the data from only the passive trials.

All analyses were done in R, version 4.2.2 [34], using "glmer" function of lme4 package [9]. We dummy coded learning condition (0: passive, 1: active). Age was centred and z-transformed to a mean of zero and a standard deviation of one to ease model convergence and interpretation wherever possible. All models detailed above used binomial error distribution and logit link function. We also included the random intercepts of participants and item, as well as random slopes of condition, age, and their interaction within item and condition within participant whenever these random effects were identifiable. Correlations between random intercepts and slopes were initially included in the models, but were removed due to convergence issues. Correlation parameters were close to -1 or 1, or their exclusion led to only a minor decrease in model fit. The study was pre-registered on Open Science Framework [1]. Python script for the experiment and R script for the data analysis can also be reached on the same platform [2].

4 RESULTS

The mean and standard deviation of proportion of correct answers separated by condition and age as well as mean and standard deviation for recurrence scores are presented in Table 1. Number of participants are different for recurrence scores due to calibration related exclusions. Children did not take breaks longer than 60 seconds between learning trials and vocabulary task, therefore the vocabulary scores should be interpreted as the results of an immediate vocabulary task.

RQ1.: Permutation analysis with shuffled recurrence scores returned a random distribution of recurrence scores with a range between 0.191 and 0.214. Comparison of recurrence scores obtained from the experimental data ($Mean_{exp\_rec}$ = 0.215, SD_{exp_rec} = 0.117) with randomly permuted data ($Mean_{perm\_rec}$ = 0.199, SD_{perm_rec} = 0.003) revealed that in the experiment children followed the robot's gaze significantly more compared to a random distribution (p = 0.001), confirming our first hypothesis. Figure 2 plots the actual recurrence score against the random distribution obtained from 1000 permutations. The narrow range of recurrence scores obtained from the permutations is an expected outcome of keeping the auto-correlated nature of gaze data. Thus, within a trial the robot does

[1] osf.io/q7x2h
[2] osf.io/f9m8k/?view_only=f36ad217db854b9cbdb94e689bb4219a

Table 1: Mean and standard deviation of proportion of correct answers in the vocabulary task per age and condition as well as recurrence score.

		Active		Passive		Recurrence		
Age	n	mean	sd	mean	sd	n	mean	sd
72-84	10	0.78	0.42	0.72	0.45	9	0.16	0.19
85-96	31	0.78	0.42	0.74	0.44	29	0.22	0.20
97-108	19	0.77	0.42	0.77	0.43	18	0.22	0.16
109-120	18	0.87	0.34	0.86	0.35	18	0.25	0.18
121-132	9	0.82	0.39	0.88	0.32	9	0.26	0.16

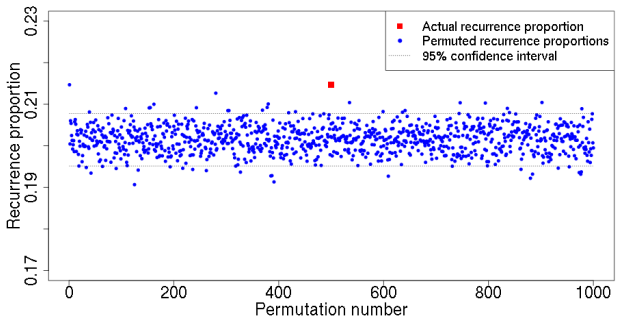


Figure 2: Actual and permuted recurrence scores. Dottet lines show 95 % confidence intervals.

not change gaze often and then only between two images and the rest of the screen (in some trials it fixates only on one location due to the randomizations in its gaze behaviour design). Since we treat each fixation as a block while shuffling the data, randomly rearranging the robot's fixations does not lead to a large change in the structure of the fixations. The significant results obtained despite these limitations strongly suggest that the children were responsive to robot's gaze.

RQ2.: The model examining whether children learned more from active trials relative to passive trials yielded a significant main effect of age, while condition or its interaction with age were not significant ($\beta_{condition}$ = 0.127, p = 0.5; β_{age} = 0.320, p = 0.03; $\beta_{condition:age}$ = -0.271, p = 0.17). Figure 3 plots change in probability of giving a correct answer in both conditions as a factor of age according to the model results. A full-null model comparison between the model reported above and a null model excluding predictors of interest (i.e., condition and its interaction with age), yielded a non-significant result ($\chi^2(2)$ = 2.48, p = 0.29). Thus, adding condition and its interaction with age did not improve model fit significantly. Particularly, while children showed learning of the label-object associations, there was no evidence that children showed improved learning in active learning trials relative to passive trials.

RQ3. The model testing the relationship between children's proportion of looking to the target object and performance at test did not yield any significant effect of the critical predictors or their interaction (β_{target_look} = 0.630, p = 0.2; β_{age} = 0.295, p = 0.13; $\beta_{target\_look:age}$ = -0.176, p = 0.7). Comparison of the full model with the model lacking the predictors of interest also showed that

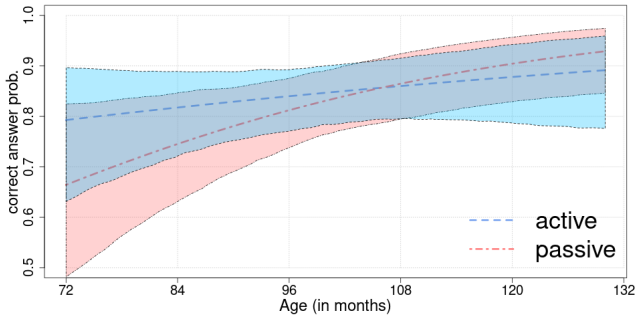


Figure 3: Change in probability of giving a correct answer in both conditions depending on age. The values were obtained based on the estimates of the model. Shaded areas show 95 % confidence intervals for fitted values.

inclusion of proportion of looking to the target and its interaction with age did not significantly improve the model fit ($\chi^2(2)$ = 1.98, p = 0.37). This suggests that children's performance at test was not predicted by their attention to the target objects during the learning phase in passive trials. Figure 4 shows the fitted values from this model.

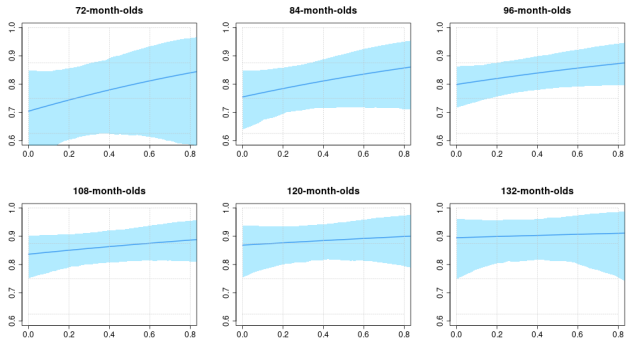


Figure 4: Children's probability of giving a correct answer (y-axis) as a function of target looking proportion (x-axis) and age. Shaded areas show 95 % confidence intervals for fitted values.

RQ4. Similarly, the model testing the relationship between children's gaze following, as indexed by their recurrence score, and performance at test did not yield significant effects of the critical predictors or their interaction ($\beta_{recurrence}$ = -0.099, p = 0.90; β_{age} = 0.558, p = 0.05; $\beta_{recurrence:age}$ = -0.782, p = 0.37). Again, a full-null comparison with a model excluding the recurrence score and its interaction with age did not change the model fit significantly ($\chi^2(2)$ = 0.79, p = 0.67). Thus, children's gaze following behaviour did not predict their performance at test. Figure 5 shows the fitted values from this model.

Lastly, we ran an exploratory model comparing children's attention to the robot in active and passive learning trials. This model included the proportion of time spent looking at the robot as the response variable and condition, age and their interaction as predictor

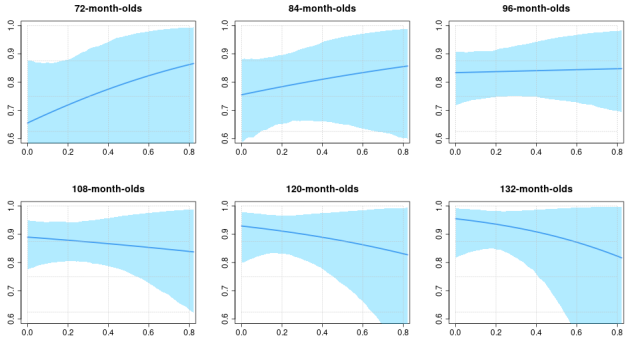


Figure 5: Change in probability of giving a correct answer (y-axis) with recurrence score (x-axis) plotted for each age in months. Shaded areas show 95 % confidence intervals.

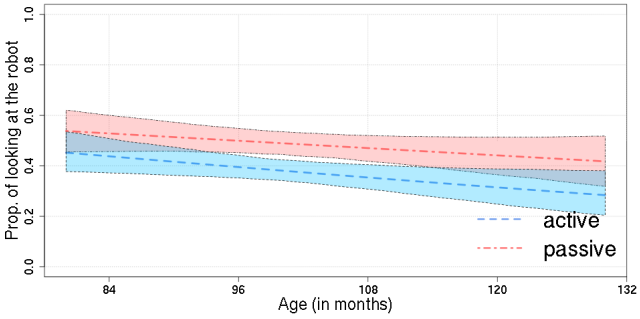


Figure 6: Children's looking proprtion to the robot as a function of condition and age. Shaded areas show 95 % confidence intervals for the fitted values.

variables. We excluded trials in which the child was looking outside the screen during the whole trial (n = 38), leaving 792 trials in total. Although we only have gaze data when children were looking at the screen, in the context of this exploratory analysis, we assume that children were looking at the robot while their gaze was away from the screen during the trials. Although random effect structure was similar to the previous models, we used "glmmTMB" function of glmmTMB package [22] as our response variable required a beta error distribution.

The analysis revealed significant effects of condition and age, but not for their interaction ($\beta_{condition}$ = -0.44, $p < 0.001$; β_{age} = -0.17, p = 0.04; $\beta_{condition:age}$ = -0.069, p = 0.39). The model output is plotted in Figure 6. As expected, children looked at the robot more during passive learning trials relative to active learning trials. Comparison of the model with a reduced model where the variables of interest (condition and age) were dropped showed that exclusion of these significantly deteriorated the model's ability to explain the variance in the data ($\chi^2(2) = 29.4$, $p < 0.001$). Bringing together the results from **RQ2.** and this exploratory analysis, our findings suggest that although children were looking more at the robot in passive trials relative to active trials, there was no evidence for differences in performance at test across conditions, i.e., they showed similar learning success in both conditions.

5 DISCUSSION

The current study examined children's gaze following behaviour in cHRI and the extent to which learning outcomes differed across child-led (active learning) and robot-led (passive learning) interactions, as indexed by either child or robot gaze. As expected, we found that children can follow a robot's gaze and learn words from interactions with robots, at least in the naturalistic settings in which this study was conducted. However, contrary to our expectations, we did not find any differences in learning outcomes across child and robot-led interactions. In particular, children showed similar learning outcomes regardless of whether the robot labelled an image that the child was looking at or the robot chose the image to be labelled by directing its gaze towards this object and subsequently labelled this image. Surprisingly, this similarity of learning outcomes persisted despite the robot distracting children's attention away from the object in robot-led interactions, i.e., passive learning trials, relative to child-led interactions, i.e., active learning trials. It should be noted that although our test of vocabulary learning was only tapping receptive vocabulary, which is only one facade of vocabulary learning, other one being expressive vocabulary, receptive proficiency is a first step towards expressive competence and our task finds that receptive competence is not affected by the different conditions.

Our finding that children followed the robot's gaze successfully in a noisy environment with little instruction and learned the label of the object that the robot was referring to using only gaze highlights that robot's gaze is sufficient to establish reference-referent connections in this context. This finding is in line with the previous studies [7, 12, 24, 29]. While the referenced items were spatially distant in most of the previous studies [23, 25, 30, 47], in our study the interaction took place on a relatively small space. Evidence of following the robot's gaze despite this limited space suggests that gaze cues are utilised efficiently in cHRI such that robot's gaze can distinguish items in close proximity to each other. Further research is needed to determine the range of resolutions at which this reference-referent connection can be successfully established.

Considering the age range of our sample and the mixed results in the literature regarding improved learning outcomes when children receive actively elicited information, our failure to find a difference in learning outcomes across robot-led and child-led interactions is not surprising. On the one hand, we note that despite the age range tested in the current study overlapping with many of the previous studies in the literature, a critical difference in our study is the use of a robot as the interaction partner. However, we do not consider this a contributing factor in our failure to find differences in learning outcomes across child-led and robot-led interactions for the following reason. On the one hand, we note that previous studies suggest that the active learning boost increases with age and is adult-like only around 8-years of age. Furthermore, our findings suggest that children adapted to attention leading roles easily and they learned the object-label pairs as evidenced by the high learning scores in Table 1. The failure to find a difference across conditions may be attributed, on the other hand, to our task design, which allowed children to learn words in a naturalistic setting. Data collections were carried out in children's schools, their typical learning

environment, and the task was introduced as a "learning game", which might have primed children towards better learning performance. Furthermore, contrary to most of previous studies where labels were either introduced in isolation or embedded in neutral carrier sentences [2, 31], our labels were embedded in informative meaningful contexts. Thus, children could leverage their knowledge of concepts associated with the novel words towards word learning performance. This could lead to children performing well regardless of conditions, eliminating previously reported effects of impoverished performance during passive learning, and highlights the importance of future research aiming for such naturalistic settings in testing paradigms.

We also found no evidence for a meaningful relationship between children's gaze following behaviour and word learning success in passive learning trials. This, too, is unsurprising given that the probability of learning the label of an item was high, even in passive trials ($mean_{vocabulary\_score}$ = 0.8). While not all children followed the robot's gaze as successfully as the others, the high levels of learning performance achieved suggest that children were, nevertheless, able to gather the necessary information with regards to the item being labelled, even from briefer gaze exchanges. This is further supported by our exploratory analysis which showed that, while children looked more at the robot during passive learning trials, they learned equally well in both conditions, which is not in line with the some other findings in the literature [20]. In other words, despite the distracting effect of the robot randomly looking at pictures prior to the labelling events, children could successfully form the reference-referent association. It remains a possibility that the contingency of the gaze exchanges during and after the labelling event, during which time the robot continued to fixate the labelled object might have proved adequate to support the reference-referent connection [19]. An alternative but related explanation of this finding might be children's attention to broader socio-pragmatic context. This explanation suggests that redundancy of multi model cues, rather than fixation time on the target improves learning [10]. It is possible that in our experiment these redundant cues such as robot's head movements, use of real objects and real information about them suppressed the potential small effect of looking time to the target on learning outcome. Future research, may, therefore, focus on the extent to which differences in participants' attention to the labelled object during and after the labelling event guide performance.

Taken together, an important merit of this study is the finding of robust learning performance using a social robot in a naturalistic setting at a school in a simple learning task, where children could individually interact with and learn from it. The robot was in a separate room, where children could visit without interrupting their daily flow and without requiring much human resources for supervision. Here, we found surprisingly high learning performance with children showing learning in conditions, e.g., the passive condition and despite decreased social contingency, that have typically showed impoverished learning in the past. Thus, this study shows that, when scaled for different tasks and subjects, a robot with similar capabilities might be implemented as a learning aid at schools efficiently. A key takeaway from the current findings is that our results reiterate the importance of gaze to infer reference, that can be reliably and smoothly used by children in cHRI.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions grant agreement No 857897. We would also like to thank Dr. Christina Keller and Charlotte von Iljin for their invaluable help in data collection, Dr. Roger Mundry for his guidance on data analysis, and Leibniz Science Campus Primate Cognition for additional funding.

REFERENCES

[1] Tobii AB. 2023. *Tobii pro SDK, version 1.11.0.* Retrieved September 15, 2008 from https://pypi.org/project/tobii-research
[2] Lena Ackermann, Chang Huan Lo, Nivedita Mani, and Julien Mayor. 2020. Word learning from a tablet app: Toddlers perform better in a passive context. *PloS one* 15, 12 (2020), e0240519. https://doi.org/10.1371/journal.pone.0240519
[3] Henny Admoni and Brian Scassellati. 2012. Robot Gaze Is Different From Human Gaze: Evidence that robot gaze does not cue reflexive attention. In *Proceedings of the "Gaze in Human-Robot Interaction" Workshop at HRI.*
[4] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63. https://doi.org/10.5898/JHRI.6.1.Admoni
[5] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers.* Springer, 114–130. https://doi.org/10.1007/978-3-642-34584-5_9
[6] Samer Al Moubayed and Gabriel Skantze. 2012. Perception of gaze direction for situated interaction. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction.* 1–6. https://doi.org/10.1145/2401836.2401839
[7] Salvatore Maria Anzalone, Jean Xavier, Sofiane Boucenna, Lucia Billeci, Antonio Narzisi, Filippo Muratori, David Cohen, and Mohamed Chetouani. 2019. Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment. *Pattern Recognition Letters* 118 (2019), 42–50. https://doi.org/10.1016/j.patrec.2018.03.007
[8] Gérard Bailly, Stephan Raidt, and Frédéric Elisei. 2010. Gaze, conversational agents and face-to-face communication. *Speech Communication* 52, 6 (2010), 598–612. https://doi.org/10.1016/j.specom.2010.02.015
[9] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01
[10] Amy E Booth, Karla K McGregor, and Katharina J Rohlfing. 2008. Socio-pragmatics and attention: Contributions to gesturally guided word learning in toddlers. *Language Learning and Development* 4, 3 (2008), 179–202.
[11] Jean-David Boucher, Ugo Pattacini, Amelie Lelong, Gerard Bailly, Frederic Elisei, Sascha Fagel, Peter Ford Dominey, and Jocelyne Ventre-Dominey. 2012. I reach faster when I see you look: gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in neurorobotics* 6 (2012), 3. https://doi.org/10.3389/fnint.2010.00005
[12] Wei Cao, Wenxu Song, Xinge Li, Sixiao Zheng, Ge Zhang, Yanting Wu, Sailing He, Huilin Zhu, and Jiajia Chen. 2019. Interaction with social robots: Improving gaze toward face but not necessarily joint attention in children with autism spectrum disorder. *Frontiers in Psychology* 10 (2019), 1503. https://doi.org/10.3389/fpsyg.2019.01503
[13] Melis Çetinçelik, Caroline F Rowland, and Tineke M Snijders. 2021. Do the eyes have it? A systematic review on the role of eye gaze in infant language development. *Frontiers in psychology* 11 (2021), 589096. https://doi.org/10.3389/fpsyg.2020.589096
[14] Thierry Chaminade and Maria M Okka. 2013. Comparing the effect of humanoid and human face for the spatial orientation of attention. *Frontiers in neurorobotics* 7 (2013), 12. https://doi.org/10.3389/fnbot.2013.00012
[15] E. V. Clark and M. Casillas. 2016. *First Language Acquisition.* Routledge.
[16] Nathan J Emery. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews* 24, 6 (2000), 581–604. https://doi.org/10.1016/S0149-7634(00)00025-7
[17] Ruthe Foushee, Mahesh Srinivasan, and Fei Xu. 2021. Self-directed learning by preschoolers in a naturalistic overhearing context. *Cognition* 206 (2021), 104415. https://doi.org/10.1016/j.cognition.2020.104415
[18] György Gergely and Gergely Csibra. 2013. Natural pedagogy. *Navigating the social world: What infants, children, and other species can teach us* (2013), 127–132.
[19] Simon Haykin and Zhe Chen. 2005. The cocktail party problem. *Neural computation* 17, 9 (2005), 1875–1902. https://doi.org/10.3389/fnhum.2019.00386
[20] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot*

interaction. 67–74. https://doi.org/10.1145/2696454.2696457

[21] Hideki Kozima and Cocoro Nakagawa. 2006. Interactive robots as facilitators of childrens social development. In *Mobile robots: Towards new applications*. IntechOpen.

[22] Arni Magnusson, Hans J. Skaug, Anders Nielsen, Casper W. Berg, Kasper Kristensen, Martin Maechler, Koen J. van Bentham, Benjamin M. Bolker, and Mollie E. Brooks. 2017. *glmmTMB: Generalized Linear Mixed Models using Template Model Builder*.

[23] AJung Moon, Daniel M Troniak, Brian Gleeson, Matthew KXJ Pan, Minhua Zheng, Benjamin A Blumer, Karon MacLean, and Elizabeth A Croft. 2014. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 334–341. https://doi.org/10.1145/2559636.2559656

[24] Javier R Movellan and John S Watson. 2002. The development of gaze following as a Bayesian systems identification problem. In *Proceedings 2nd International Conference on Development and Learning. ICDL 2002.* IEEE, 34–40. https://doi.org/10.1109/DEVLRN.2002.1011728

[25] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 518–523. https://doi.org/10.1109/ICHR.2006.321322

[26] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 61–68. https://doi.org/10.1145/1514095.1514109

[27] Bilge Mutlu, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 69–76. https://doi.org/10.1145/1514095.1514110

[28] Eunice Mwangi, Emilia I Barakova, Marta Díaz, Andreu Català Mallofré, and Matthias Rauterberg. 2018. Dyadic gaze patterns during child-robot collaborative gameplay in a tutoring interaction. In *2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 856–861. https://doi.org/10.1109/ROMAN.2018.8525799

[29] Yuko Okumura, Yasuhiro Kanakogi, Takayuki Kanda, Hiroshi Ishiguro, and Shoji Itakura. 2013. Infants understand the referential nature of human gaze but not robot gaze. *Journal of experimental child psychology* 116, 1 (2013), 86–95. https://doi.org/10.1016/j.jecp.2013.02.007

[30] Linda Onnasch, Eleonora Kostadinova, and Paul Schweidler. 2022. Humans can't resist robot eyes–reflexive cueing with pseudo-social stimuli. *Frontiers in Robotics and AI* 9 (2022), 848295. https://doi.org/10.3389/frobt.2022.848295

[31] Eric Partridge, Matthew G McGovern, Amanda Yung, and Celeste Kidd. 2015. Young children's self-directed information gathering on touchscreens. In *Proceedings of the 37th annual conference of the cognitive science society, austin, tx. Cognitive science society*.

[32] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior research methods* 51 (2019), 195–203. https://doi.org/10.3758/s13428-018-01193-y

[33] Alfredo F Pereira, Linda B Smith, and Chen Yu. 2008. Social coordination in toddler's word learning: Interacting systems of perception and action. *Connection Science* 20, 2-3 (2008), 73–89. https://doi.org/10.1080/09540090802091891

[34] R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[35] Daniel C Richardson and Rick Dale. 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science* 29, 6 (2005), 1045–1060. https://doi.org/10.1207/s15516709cog0000_29

[36] Furhat Robotics. 2021. *Furhat Remote API, version 1.0.2*. Retrieved September 15, 2008 from https://pypi.org/project/furhat-remote-api

[37] Azzurra Ruggeri, Douglas B Markant, Todd M Gureckis, Maria Bretzke, and Fei Xu. 2019. Memory enhancements from active control of learning emerge across development. *Cognition* 186 (2019), 82–94. https://doi.org/10.1016/j.cognition.2019.01.010

[38] Atsushi Senju and Gergely Csibra. 2008. Gaze following in human infants depends on communicative signals. *Current biology* 18, 9 (2008), 668–671. https://doi.org/10.1016/j.cub.2008.03.059

[39] Stephen V Shepherd. 2010. Following gaze: gaze-following behavior as a window into social cognition. *Frontiers in integrative neuroscience* 4 (2010), 5. https://doi.org/10.3389/fnint.2010.00005

[40] Zi Lin Sim, Michelle M Tanner, Nina Y Alpert, and Fei Xu. 2015. Children Learn Better When They Select Their Own Data.. In *CogSci*.

[41] Adriana Tapus, Andreea Peca, Amir Aly, Cristina Pop, Lavinia Jisa, Sebastian Pintea, Alina S Rusu, and Daniel O David. 2012. Children with autism social engagement in interaction with Nao, an imitative robot: A series of single case experiments. *Interaction studies* 13, 3 (2012), 315–347. https://doi.org/10.1075/is.13.3.01tap

[42] Michael Tomasello and Michael Jeffrey Farrar. 1986. Joint attention and early language. *Child development* (1986), 1454–1463. https://doi.org/10.2307/1130423

[43] Michael Tomasello and Jody Todd. 1983. Joint attention and lexical acquisition style. *First language* 4, 12 (1983), 197–211. https://doi.org/10.1177/014272378300401202

[44] Lev S. Vygotsky. 2016. *Thought and language (Abridged from 1934; A. Kozulin, Trans.)*. MIT Press, Cambridge MA.

[45] Zachary E Warren, Zhi Zheng, Amy R Swanson, Esubalew Bekele, Lian Zhang, Julie A Crittendon, Amy F Weitlauf, and Nilanjan Sarkar. 2015. Can robotic interaction improve joint attention skills? *Journal of autism and developmental disorders* 45 (2015), 3726–3734.

[46] Eva Wiese, Patrick P Weis, and Daniel M Lofaro. 2018. Embodied social robots trigger gaze following in real-time HRI. In *2018 15th International Conference on Ubiquitous Robots (UR)*. IEEE, 477–482. https://doi.org/10.1109/URAI.2018.8441825

[47] Chen Yu, Paul Schermerhorn, and Matthias Scheutz. 2012. Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 2 (2012), 1–25. https://doi.org/10.1145/2070719.2070726

[48] Chen Yu and Linda B Smith. 2012. Embodied attention and word learning by toddlers. *Cognition* 125, 2 (2012), 244–262. https://doi.org/10.1016/j.cognition.2012.06.016

[49] Zhi Zheng, Lian Zhang, Esubalew Bekele, Amy Swanson, Julie A Crittendon, Zachary Warren, and Nilanjan Sarkar. 2013. Impact of robot-mediated interaction system on joint attention skills for children with autism. In *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 1–8. https://doi.org/10.1109/ICORR.2013.6650408