TODO

Anonymous CogSci submission

Abstract

TODO abstract Communication is important and stuff and kids need to learn it to. One task that illustrates the rich communicative capabilities is iterated reference games, where adults jointly converge on mutually understandable names for novel objects. Some prominent early experiments where children did not succeed led to the belief that children could not do this task. TODO FIX. However, more recent work with lessened task demands suggests that children may be more capable than previously thought. Here, we test children's communicative adaption to a partner with iterated reference games played by pairs of 4-5 year old preschoolers. Across a total of N pairs of children, we find children can accurately pick out referents for each other and that they are sensitive to the descriptions of their partners. TODO wrap up sentence.

Keywords: TODO keywords

Introduction

TODO introduction

Experiment 1 Methods

Our goal for experiment 1 was to test young children's ability to coordinate on descriptions to abstract shapes that their partner could understand. Young children can be very sensitive to task demands and cognitive load that can hide early abilities (TODO cite something for this), so we used a simple paradigm where an experimenter could scaffold the children's iteraction as needed. We adapted the experimental framework from (**leung?**), but further simplified it by reducing the total pool of targets and the number of trials children completed.

This experiment was pre-registered at `https://osf.io/kcv8j`.

Participants

4 and 5 year old children were recruited from a university nursery school during the school day. Children played with another child from the same class. Experiment 1 was conducted between June and August 2023. Pairs of children were included in analyses if they completed at least 8 of the 12 critical trials. We had 19 complete games and 1 incomplete, but included game. Of the 40 children, 21 were girls, and the median age was 57 months, with a range of 48-70 months.

Materials

For the target stimuli, we used four of the ten tangram images from (**leung?**), chosen based on visual dissimilarity. We coded the matching game using Empirica and hosted it on a lab server (CITATION). We then accessed the game on the web on tablets that were locked in a kiosk mode so children could not navigate to other websites or applications during the game.

Procedure

Once a pair of children agreed to play the game, a research assistant took them to a quiet testing room where the game was explained to them. Children were introduced to a stuffed animal "Smurfy" who wanted to play a matching game. Children sat across a table from each other, each with a tablet in front of them. On each trial, one child saw two images, one of them in a black box, and was asked to "say what they saw" in the black box so their partner (and Smurfy) could tap the corresponding image. The guesser saw the same two images (in a randomized order), but with neither boxed. Upon tapping an image, both children received feedback in form of a smiley or frowny face and an audible sound. After each trial, children's roles switched. Children passed Smurfy back and forth to help them keep track of whether they were the "guesser" or "teller" on a given trial.

Children completed two warm-up trials with black and white images of familiar shapes, followed by 3 blocks of the 4 targets. Targets were randomly paired with another of the critical images as the foil.

The experimenters running the game did not volunteer descriptions, but did scaffold the interaction, prompting children to describe the images, and sometimes repeating children's statements. The entire interaction was video-recorded.

TODO DIAGRAM OF EXPERIMENT + Picture of children playing the game as top bar

Data processing

Children's selections and the time to selection were recorded from the experiment software. Children's descriptions were transcribed from the video-recording, using Whisper (CITE) for the first pass and then hand-corrected by experimenters. Transcripts were hand-annotated for when each trial started, who said each line, and what referential descriptions were used.

We excluded trials where the "teller" did not produce a description, or where all description was unintelligible and im-

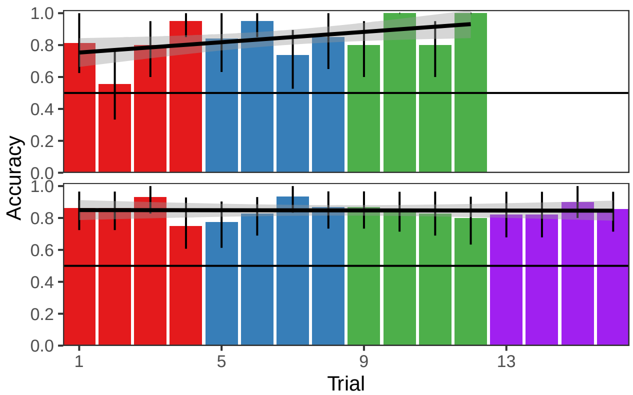

Figure 1: Children's accuracy at selecting the correct target over time. Experiment 1 had 3 blocks (12 total critical trials) and experiment 2 had 4 blocks (16 critical trials). Error bars are bootstrapped 95% CIs with a linear trend line overlaid.

possible to transcribe. After exclusions, we had 466 trials remaining.

Statistical analysis

Statistical analyses were run in brms with weakly informative priors. We present the estimate and 95% credible intervals.

Experiment 1 Results

Accuracy

Our primary measure of interest was whether children could accurately communicate the intended target, as prior work is often interpreted as indicating the children at kindergarten age cannot communicate about abstract shapes successfully (CITATION). As shown in Figure 1, children were above chance in their selections. To confirm this and test for any changes in accuracy over time, we fit a mixed effects model of accuracy (correct.num∼ trial.num + (trial.num|game) + (1|target)). Children's accuracy was above chance (Odds Ratio: 3 [1.14, 8.09]) and accuracy slightly increased over the game (OR of one trial later: 1.17 [1.03, 1.39]).

Speed

As another measure of children's performance, we looked at how long each trial took to see if children were getting faster over time. We ran a Bayesian mixed effects model of how long each trial took over time: time.sec∼ trial.num + (trial.num|game) + (1|target). The first trial critical trial averaged 27.48 [20.48, 34.53] seconds, and children got faster over time (-1.22 [-1.9, -0.53]).

Description length

In iterated reference games in adults, a canonical finding is that the length of descriptions goes from long to short over repeated references to the initially hard to describe shapes. We looked at how long the descriptions the children used were to see if the same trend occurred. We ran a Bayesian mixed effects model of how long of a description the "teller" produced:

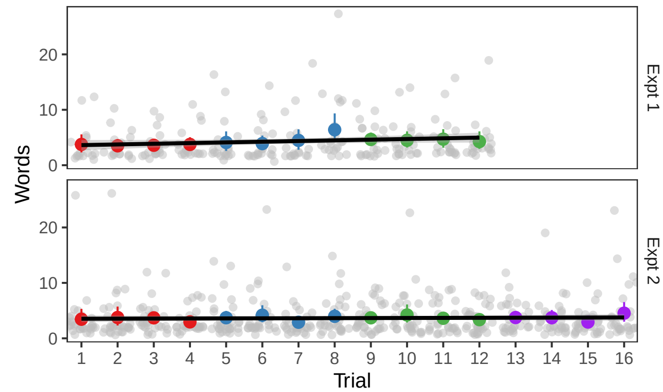

Figure 2: Length of referential expression (in words) produced by the speaker each trial. Grey dots are individual data points, colored dots are per trial means with bootstrapped 95% CIs.

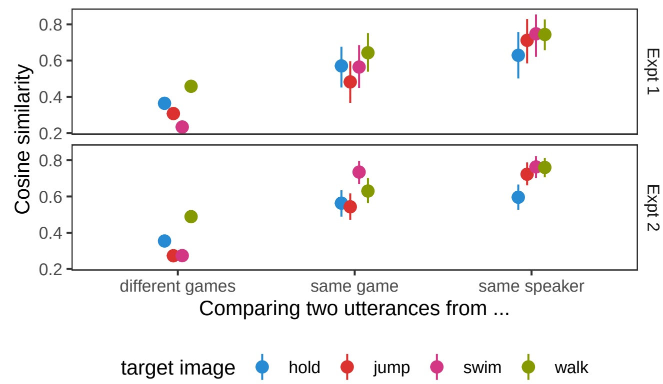

Figure 3: Semantic similarity between referential descriptions from the speaker to the same target under different circumstances. Different games refers to similarities across two speakers from different games, same game to similarities across the two participants in one game, and same speaker to descriptions from the same person in different blocks. Dots are means and lines are bootstrapped 95% CIs.

words∼ trial.num + (trial.num|game) + (1|target). The initial length was 3.66 [2.56, 4.84] and description length was relatively stable over time (0.12 [-0.05, 0.27], shown in Figure 2). How long of a description is initially warranted depends on how iconic or easily describable the shape is, as well as how large and close the contrast set is. In this case, the low number of options to distinguish and relatively easy to describe shapes may mean long initial descriptions are less necessary (todo mention how long the similar things are in leung). However, young children may also choose to provide shorter descriptions than adults would.

TODO add anecdotes here with some examples of children's descriptions.

Convergence

While description length is used as a proxy for measuring convention formation, better measurements for convergence look at the actual content of the utterances, which does not always track description length (cite boyce 2024). With only 3 repetitions for each target, we do not have enough data to look at convergence over time. However, we can still analalyze the semantics of the descriptions to see if children are influenced by their partner's descriptions or not.

Following cite Boyce et al 2024, we use Sentence-BERT to embed the descriptions in a semantic vector space, and then use the cosine between embeddings as a measure of semantic similarity between descriptions. We compare the semantic similarities between descriptions of the same target based on who produced the description. Our question is whether the two children within the same game produce descriptions that are more similar to each other than two children in different games. As shown in Figure 3, different children in the same game do produce more similar descriptions than children in different games, although the descriptions are less similar than descriptions from the same child in different rounds. We modeled this as sim∼ same_game + same_speaker + (1|target). Utterances were more similar if they came from the same partnership (0.222 [0.184, 0.26]) and were slightly more similar still if they came from the same person with the partnership (0.135 [0.077, 0.193]). The big differences in descriptions between games compared to within games is a measure of partner sensitivity – children are more likely to use descriptions semantically similar to their partner's than to another child's. This provides weak evidence for some coordination between chidren.

TODO could provide anecdotes about children's coordination approaches

Experiment 1 Discussion

Experiment 2 Methods

In experiment 1, children were above chance accuracy in their selections. In experiment 2, we aimed to repeat the same paradigm with a tighter experimental script to reduce possible influence of experimenters. Additionally, we aimed to fix sources of confusion and frustration in experiment 1. As most pairs in experiment 1 completed the game fairly quickly, we added a 4th experimental block to allow for more analyses of change over time.

As Experiment 2 was very similar to Experiment 1, here we note the differences. Experiment 2 was pre-registered at `https://osf.io/y2dax`.

Participants

Experiment 2 was run between March and August of 2024, at the same university preschool as experiment 1. No children participated in both experiments. 30 pairs of children completed all 16 critical trials, and 1 pair of children completed between 8 and 16 critical trials. Our target age range was 4 and 5 year olds, but one almost 4-year-old was unintentionally included. Of the 62 children, 30 were girls, and the children had a median age of 56 months, and a range of 45-69 months.

Materials

The same 4 critical images were used as in experiment 1, although this time, children saw these images 4 times. In response to some children struggling with the abrupt switch from nameable to non-nameable shapes, we introduced more practice trials for experiment 2. We used a total of 4 practice trials, designed to transition from easily recognizable shapes to slightly blockier black and wide shapes to smooth the transition to the critical trials. ## Procedure The procedure was much the same. We added an initial "bubble popping" exercise to give children practice with how to tap the tablet appropriately (this was an issue for some children in the first experiment). The smurfy puppet was swapped out for a more attractive smurfy stuffed animal. The experimental script was fully written out so children all recieved the same instructions. To prevent experimenter's influencing children's descriptions or understanding of descriptions, we wrote up contingency statements that the experimenter could use to prompt children who were not giving descriptions or making selections. TODO link to materials for where the script is.

Data processing

Data was processed in the same way as experiment 1. After excluding trials where children did not give a description or where the experimenter echoed a child's description, we had 466 trials total.

Experiment 2 Results

We report the same set of analyses and model results as in Experiment 1 as well as additional analyses of how the semantic content of children's descriptions changes over time.

Accuracy

As in Experiment 1, in Experiment 2, children's accuracy was above chance (Odds Ratio: 5.95 [3.07, 11.89]) and accuracy slightly increased over the course of the game (OR of one trial later: 1.01 [0.94, 1.09], Figure 1). This confirms that children are able to communicate with each other about these abstract shapes.

Speed

In Experiment 2, the first critical trial averaged 22.06 [15.86, 28.58] seconds, and children got faster over time (-0.7 [-0.99, -0.41]). Children were initially faster in Experiment 2 than Experiment 1, possibly due to the increased number of practice trials and pre-training on how to press the screens.

Description length

The average length of descriptions on the first trial was 3.45 [2.2, 4.81] words and description length was relatively stable over time (0.02 [-0.04, 0.09], Figure 2). This is comparable to Experiment 1, again finding that children produce short utterances without much change in length over time.

TODO could include examples of cute things kiddos said here

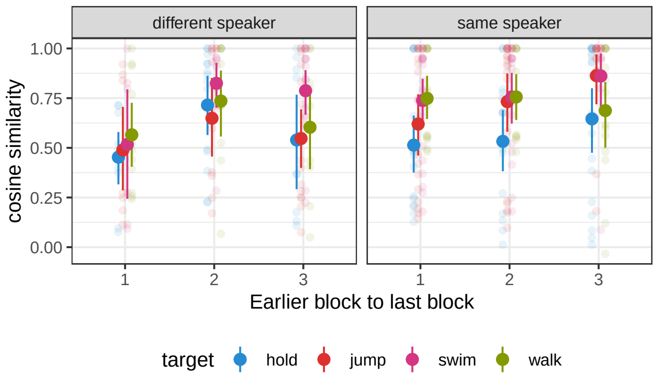

Figure 4: Semantic similarity between earlier blocks (1-3) with last block (4) for descriptions to the same image within the same group. SImilarity measured as cosine similarity between S-BERT embeddings of referential descriptions. Heavy dots are means with bootstrapped 95% CIs; light dots are individual values.

Convergence

To look at semantic distance between utterances, we again operationalize similarity between pairs of utterances as the cosine similarity between their SBERT embeddings (CITE SBERT).

As a coarse comparison, we repeated the analysis from Experiment 1, comparing the similarity of descriptions to the same target for the same-speaker, same-game, or different-game. We modeled this as sim∼ same_game + same_speaker + (1|target). Utterances were more similar if they came from the same partnership (0.27 [0.243, 0.297]) and were slightly more similar still if they came from the same person with the partnership (0.097 [0.059, 0.132]). The big differences in descriptions between games compared to within games is a measure of partner sensitivity – children are more likely to use descriptions semantically similar to their partner's than to another child's.

The greater number of trials in Experiment 2 makes it possible to look for changes over time that could be indicative of convergence to shared descriptions within a game and divergence between games.

To look for convergence to shared descriptions within games, we compared the utterances from the first three blocks to the descriptions in the last block: sim∼ earlier_block.num + same_speaker + (1|game1) + (1|target). Over the first three blocks, descriptions become increasingly similar to the last block description (0.042 [0.007, 0.078]). Descriptions are more similar to the last block if they come from the same child who gave the description in the last block (0.067 [0.006, 0.127]).

Another way to look for convergence is to look at the semantic distance between utterances in adjacent blocks: sim∼ earlier_block.num + same_speaker + (1|game1) + (1|target). Although over time descriptions do get more similar to the last block utterance, the distance between adjacent block utterances is relatively constant: 0.009 [-0.026, 0.044].

As each partnership converges to their shared nicknames, partnerships often diverge from one another as groups focus on distinct aspects of the image. We tested whether descriptions in different games to the same target diverged over time: sim∼ block.num + (1|target). As the games progress, descriptions from different games became slightly further apart in semantic space (-0.013 [-0.018, -0.008]).

TODO question for the group: do we want to include any meta analysis across the two expts?

TODO do we want to try to look at whether successful descriptions are more likely to "stick" (might have low sample b/c accuracy is high)

General Discussion

Summary of experiments

Limitations. The population of children at university nursery schools is non-representative, and the set of materials we used was also not that varied. This set of tangram images may be easier to distinguish and refer to than some sets used with adults, leading to overall shorter utterances. Probably shouldn't say based on this that children can "do reference" at age 4, but it is evidence that under supportive circumstances, a number of children at this age are able to.

We also specifically target the construction of referring expressions that can be jointly understood. There are other parts of the coordination where help was provided in children seemed stuck or confused, such as when to make a choice or ask for more information.

Broader implications

This work (along with other work on children's referential communicaiton) suggests that there's a more gradual development Has implications for how we think about children's language development. There's debate over how that is ordered and whether communication/pragmatics is a final stage, or how all the stages are boostrapped (CITATIONS). This early ability to use language for communicative purposes is more consistent with the early pragmatics viewpoint.

Suggestive of a gradual development where children's capabilities are increasing for a wide amount of childhood, as their working memory capacity and executive function improve and they are able to better track other's states of knowledge and keep track of wider arrays of images.

References