TODO Preschoolers form conventional pacts with each other to communicate about novel shapes

Anonymous CogSci submission

Abstract

Learning language requires learning not only the content of language, but also skills that allow us to use language to communicate. Iterated reference games provide a window into such skills, requiring rich communication as participants jointly converge on mutually understandable names for initially novel objects. Some classical experiments with young children are interpreted as showing that 4-5-year old children are incapable of succeeding at iterated reference games. Here, we revisit the question of young children's referential communicative abilities using a simpler, child-friendly paradigm. Across 51 pairs of 4-5-year old children, we find that preschool-aged children were successful in an iterated reference game. Children were Y% accurate, and they often used descriptions similar to their partner's. These findings suggest that children's capacity to construct effective referring expressions in novel contexts emerges earlier than once thought, consistent with a larger picture of the early emerging ability to pragmatically use language in communication. TODO fix wrap up sentence! TODO possibly shorten!

Keywords: development; iterated reference game; TODO keywords

Introduction

Learning a language requires learning not only the content of that language, but also how to use the language to communicate. One test environment for language use is referential communication, the ability to give a description so that an interlocutor can pick out a target referent from a set of possibilities. Adults show sensitivity to both the visual context and their audience during referential communication, calibrating the type of description and the amount of information they provide to their beliefs about the interlocutor's knowledge state.

One test of referential communication skill is *iterated reference games* where one describes abstract shapes from an array of images repeatedly to the same partner (Boyce et al., 2024; Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2020; Krauss & Weinheimer, 1964). Over repetition, features of the initial verbose descriptions are conventionalized as each pair comes to agree on a shared understanding of how to label each image. Success at this task requires mastery of a number of linguistic and communicative skills, including producing adequate initial descriptions, monitoring for comprehension, asking for clarification, and appropriately using the shared conversation history to inform referring expressions in later rounds.

Indeed, even young children must make use of these skills in order to achieve success in the task. However, there is debate on the developmental trajectories of these communicative skills. Studying how children play iterated reference games can provide insight into the developmental trajectory of the ability to produce referential expressions in order to achieve joint understanding.

A study on children's referential communication suggests that 4-5-year old preschoolers struggle to succeed at child-child referential communication (Glucksberg et al., 1966). In their paradigm, one child was given a set of 6 blocks in a specific order. Their task was to describe the images on the block so their partner could pick out the corresponding block. As children described and selected blocks, they stacked them on pegs, so that the order of blocks on the peg was a measure of correct matching. While 4-5-year old children were successful on practice trials where the blocks had drawings of farm animals and they had visual access to each other's blocks; children failed on critical trials where the blocks had drawings of abstract shapes and there was no visual access. Even after multiple rounds with the same images, children were not able to produce referential descriptions that allowed for success at the matching task. Glucksberg et al. (1966) attributed young children's failure to the children's use of idiosyncratic referring expressions unique to their own experiences, making it difficult for the pair to converge on shared descriptions.

Similar experiments with older children indicated a gradual improvement through adolescence both at their initial accuracy and on their improvement over repetitions. Still, even the 9th grade sample was noticeably worse than the sample of adult college students (Glucksberg & Krauss, 1967; Krauss & Glucksberg, 1969). That the task is difficult even among teenagers suggests that the complexity of the stacking paradigm and the large number of potential targets presents significant task demands that might be masking the referential communication abilities of even the youngest children.

Since then, a number of studies have revealed early emerging skills in the preschool years that support the use of referential communication. Studies show that preschool-aged children (3-5-years old) are faster to select a target when it is referred to in a consistent way (Graham et al., 2014; Matthews et al., 2010) and by 6-years more consistently produce partner-specific referential expressions (Köymen et al., 2014), suggesting that young children are sensitive to consistent use of descriptions in cooperative communication. Further, preschool-aged children adapt the informativeness of their referential expressions based on the visual content available to their interlocutor (Matthews et al., 2006; Nadig & Sedivy, 2002; Nilsen & Graham, 2009), showing that children at this age can reason about others' perspectives to more efficiently communication.

While children are sensitive to others' perspectives, they still seem to struggle with appropriately tailoring the specificity of their utterances to the visual context, resulting in utterances that are under-informative (Leung et al., 2024; Matthews et al., 2012). When children must describe one of two very similar images, 4 and 5-year olds struggle to mention the relevant features, although they do better when playing in an interactive game than when not (Grigoroglou & Papafragou, 2019). By 5-years old, children are sensitive to utterances that are over- or under-informative, asking for clarification on some under-informative utterances and taking longer to make selections on over-informative utterances (Morisseau et al., 2013).

A common thread between many studies in referential expression production is that children perform better when the cognitive load of the task is reduced, i.e. fewer distractor referents (Abbot-Smith et al., 2016). This paired with evidence of emerging communicative skills in young children suggests that tailored referential expression production is possible in children, but it is likely to be masked when the task demands are too high.

Since Glucksberg et al. (1966), few studies have revisited the question of whether children can successfully communicate in an iterated reference game. Recent work shows that 8-10-year olds exhibit classic patterns of increasing accuracy, increasing speed, and shorter descriptions across repetitions in an integrated reference game, but children's accuracy was still far below adult performance and varied dramatically between dyads (Branigan et al., 2016). 4-6-year olds have also been shown to rely on conventionalized gestures with their partner in an integrated reference game where children could only use gestures rather than speech to communicate (Bohn et al., 2019). A recent study with younger children from 4-8-years old examined children's abilities to play an iterated reference game with a parent using a simple, child-friendly tablet-based task. In this task, even 4-year olds had an initial accuracy above 80%, which rose to above 90% in later repetitions (Leung et al., 2024). Together, these findings provide evidence of young children's ability to successfully communicate in an iterated reference game under the right conditions.

Given the task demands in Glucksberg et al. (1966) and work showing that children can succeed in a simplified, less demanding paradigm, we revisit the question of preschool-aged children's ability to converge on appropriate referential expressions with each other. In the present study, we re-examine young children's ability to establish effective referring expressions with each other in an iterated reference game using a simplified tablet-based paradigm. Across 51 pairs of 4-5-year old children, we find that preschool-aged children were successful in an iterated reference game, suggesting that children's capacity to construct effective referring expressions in novel contexts emerges earlier than once thought.

Experiment 1 Methods

(Note: we should talk about whether it makes more sense to do all the methods and then all the results versus the current expt 1 than expt 2 layout.)

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 [CITATION WANTED], so we adapted the experimental framework from Leung et al. (2024), 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 the 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 et al. (2024), chosen based on visual dissimilarity (Figure 1B). We coded the matching game using Empirica and hosted it on a lab server (Almaatouq et al., 2020). 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 (Figure 1A). 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). 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

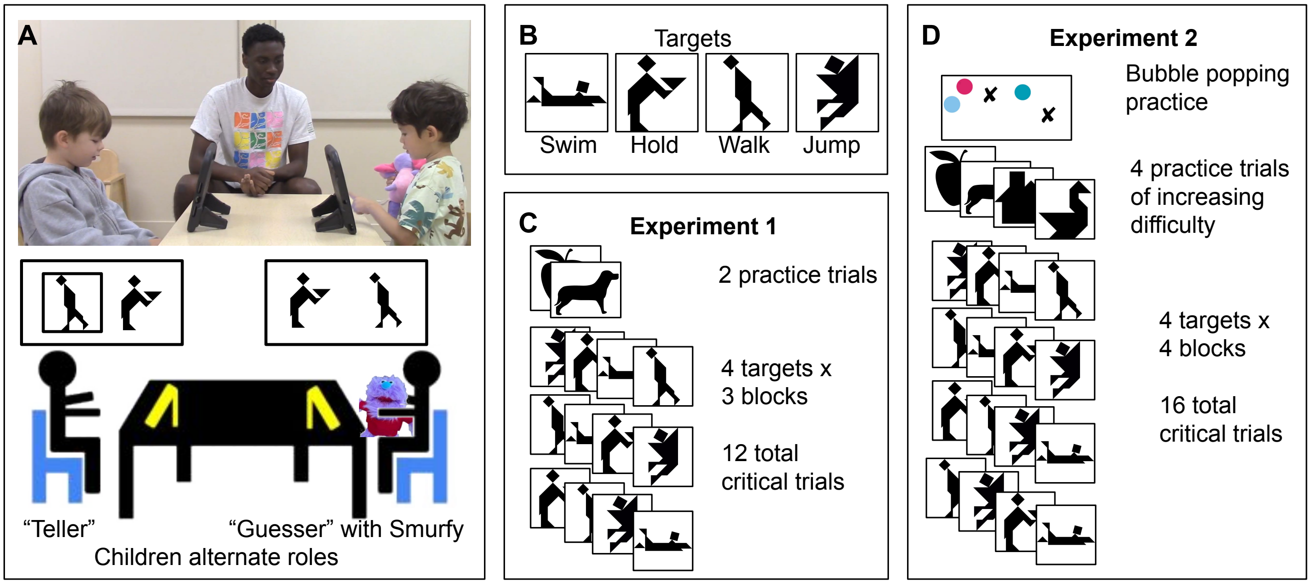

Figure 1: Experimental Setup and Procedure. Panel A shows the experimental setup with the teller and guesser across the table from each other. Panel B shows the 4 possible targets; names for targets are for cross-reference with later figures only. Panel C shows the procedure for Experiment 1; within critical block targets were ordered randomly. Panel D shows the procedure for Experiment 2.

4 targets (Figure 1C). 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.

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 (Radford et al., 2022) 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 impossible to transcribe. After exclusions, we had 466 trials remaining.

Statistical analyses were run in TODO CITE 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

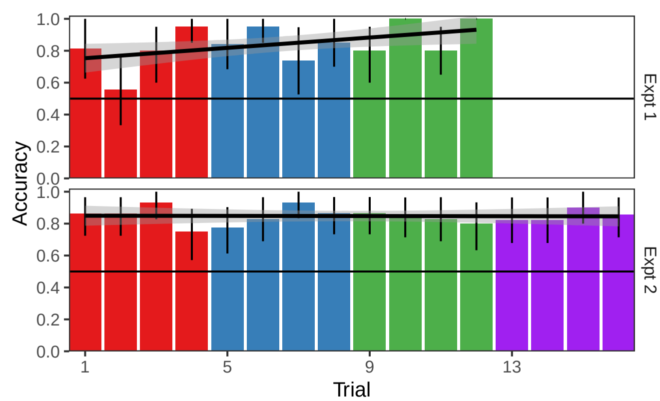

Figure 2: Children's accuracy at selecting the correct target over time. Experiment 1 had 3 blocks (12 total critical trials) and experiment 2 had 4 blocks (16 critical trials). Error bars are bootstrapped 95% CIs with a linear trend line overlaid.

(Glucksberg et al., 1966). As shown in Figure 2, 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

[less important, could drop] As another measure of children's performance, we looked at how long each trial took to see

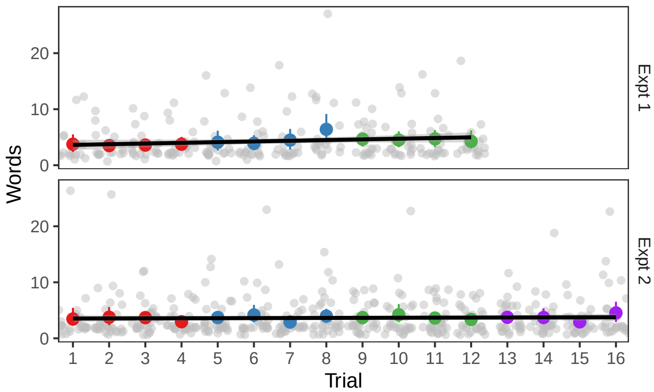

Figure 3: Length of referential expression (in words) produced by the speaker each trial. Grey dots are individual data points, colored dots are per trial means with bootstrapped 95% CIs.

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 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 (Boyce et al., 2024; Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2020). 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: 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 3). 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. (The adult controls in Leung et al. (2024) use shorter initial descriptions than adults in studies with larger arrays of harder to distinguish images.) However, young children may also choose to provide shorter descriptions than adults would, and young children may accept shorter descriptions as adequate whereas adults may ask for more information to increase their confidence level before making a guess.

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 (Boyce et al., 2024). Following 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 compared the semantic similarities between descriptions of the same target based on who produced the description. Our question was whether the two children within the same game produced descriptions that are more similar to each other than two children in different games.

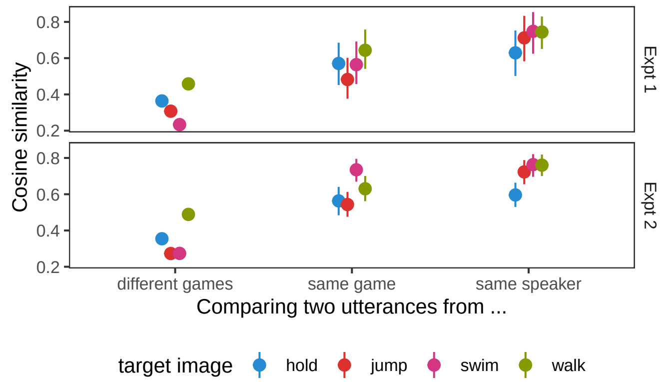

Figure 4: Semantic similarity between referential descriptions from the speaker to the same target under different circumstances. Different games refers to similarities across two speakers from different games, same game to similarities across the two participants in one game, and same speaker to descriptions from the same person in different blocks. Dots are means and lines are bootstrapped 95% CIs.

As shown in Figure 4, different children in the same game produced more similar descriptions than children in different games, although the descriptions were 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.

Experiment 1 Discussion

In the first experiment, we adapted the paradigm of Leung et al. (2024) for pairs of children, taking an already simple set up and making it shorter. Our goal was to see if young children were at all able to provide descriptions to each other adequate to identify abstract figures. so children received a lot of scaffolding around the experimental interaction. Sometimes, this included experimenters echoing children's descriptions which could potentially influence children's responses. In Experiment 2, we repeated the same paradigm, with tighter experimenter controls and a larger sample size.

Experiment 2 Methods

As Experiment 2 was very similar to Experiment 1, we focus on the changes made compared to Experiment 1. 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 to provide a gradient from more recognizable shapes to less recognizable, blockier shapes (Figure 1D).

Procedure

The procedure was much the same (Figure 1D). 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 experimental script was fully written out so children all received the same instructions. We wrote up contingency statements that the experimenter could use to prompt children who were not giving descriptions or making selections. Experimenters would help with game mechanics such as who's turn it was to tell and who should press the screen to move the game along, but would not provide or repeat any content about the images, what to ask, or whether a description was adequate.

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 2). 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 3). This is comparable to Experiment 1, again finding that children produce short utterances without much change in length over time. Children used a wide range of descriptions successfully, and some example descriptions are shown in Table 1. Often children would use the same description as their partner, although there were some transitions such as from "a person holding a flute" to "a person holding a trumpet".

(not sure how much to say or where to put) Pairs of children varied significantly in how well they could scaffold their own interaction versus needed reminders from experimenters about the structure of the game. In one game, the experimenter provided no input after the practice trials because the children rapidly took turns producing referring expressions and selecting the targets. Their advantage could partially be due to being best friends with each other, who may have been more comfortable with each other and had more common ground to start with. All children were paired with another child from the same classroom they were willing to play with, but friendship levels varied.

Some children asked clarifying questions of their partner, but knowing how much information to provide was not always consistent. In one game, child A described the figure as "A human", prompting child B to note "There's two humans." Later in the game, child B used the description "A human", leading child A to ask "Which one?", which child B clarified with"The one that is walking". This anecdote illustrates how emerging abilities can be inconsistent – finding a description inadequate as a guesser does not always translate into providing a more informative description as a teller.

Convergence

To look at semantic distance between utterances, we again operationalize similarity between pairs of utterances as the cosine similarity between their Sentence-BERT embeddings (Reimers & Gurevych, 2019).

As a coarse comparison, we repeated the analysis from Experiment 1. 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 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.

Table 1: Example descriptions children used for different tangrams. TODO may want to shorter or pair with the tangram images

Descriptions
”Hold”:
• person
• a person holding a sandwich
• a people carrying a box of dirt
• a monster
• someone holding a plate and giving it to a restaurant and has watermelon
”Jump”:
• vampire
• hopping
• a person flying
• somebody skydiving, not in the airplane
• a person
• a kite
• a triangle with a head on it with feet
”Swim”:
• racecar
• airplane
• alligator
• a person fell down
• a boat
• a person that’s in a race car that has one triangle and two triangles
”Walk”:
• person
• a person walking
• a person looking down
• a people, but it doesn’t have any arms

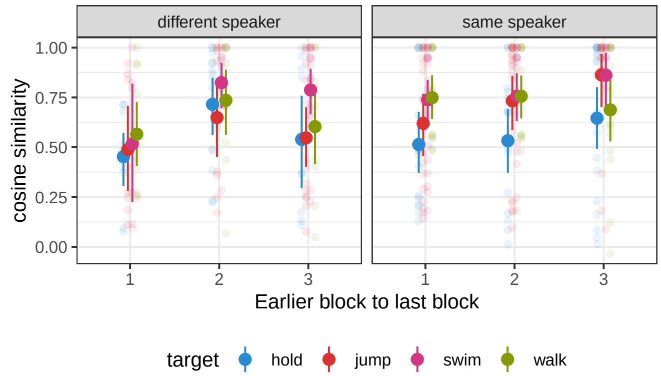

Figure 5: Semantic similarity between earlier blocks (1-3) with last block (4) for descriptions to the same image within the same group. SImilarity measured as cosine similarity between S-BERT embeddings of referential descriptions. Heavy dots are means with bootstrapped 95% CIs; light dots are individual values.

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]).

Post hoc results

Meta-analytic placeholder

Might want to split out and incorporate in along with expt 2 analyses?

As the two experiments were similar to one another, we re-run models pooling across the two experiments. Pooling the two experiments, children’s accuracy was above chance (Odds Ratio: 5.01 [2.78, 9.42]) and accuracy slightly increased over the course of the game (OR of one trial later: 1.04 [0.98, 1.12].

The average length of descriptions on the first trial was 3.59 [2.69, 4.45] words and description length was relatively stable over time (0.04 [-0.02, 0.1], Figure 3).

Utterances were more similar if they came from the same partnership (0.257 [0.235, 0.279]) and were slightly more similar still if they came from the same person with the partnership (0.109 [0.079, 0.139]). Since the two experiments had different length games, we do not want to pool them in comparing to the last block, but we can pool them comparing to the next block.

The distance between adjacent block utterances is relatively constant: 0.017 [-0.014, 0.05].

Query – I did not include experiment # as a random effect here, and maybe should have? I can rerun if we decide we want this analysis and think expt # as a random effect with game nested within it would be useful.

Accuracy is predictive of similarity to next

As a post-hoc measure, we look at whether descriptions that elicit correct selections have more staying power than descriptions that do not elicit correct selections. Here we operationalized "staying power" as similarity to the next block utterance, and we model with sim$\sim$ earlier_block.num $\times$ correct + same_speaker $\times$ correct + (1|game1) + (1|target). Correct utterances have an increase in similarity with the next block by 0.147 [0.058, 0.236] with no substantial interaction with block number or same speaker. Numerically, the size of the boost for being correct is similar to the size of the boost from coming form the same speaker 0.152 [0.053, 0.246].

General Discussion

Across 2 experiments and 51 pairs of 4 and 5-year old children, we tested how well children were at producing referential expressions to allow their partner to find a matching abstract shape. Children varied substantially in what sorts of descriptions they produced, but overall accuracy was high (NUMBER), indicating that children were generally able to produce adequate descriptions. Additionally, children's utterances showed signs of converging toward conceptual pacts. While this task is substantially scaled down relative to measures used for adult competence, and children do not display all the typical trends (i.e no sign of shortening of referential expression), it does suggest that the relevant communication skills are present at least in rudimentary form by the end of the preschool years.

At least some preschoolers, in an environment with low task demands and support for understanding the structure of the task, can succeed at iterated referential communication. Our task is limited by the non-representative population of children at university nursery schools who participated, and by a limited set of target images. This set of tangram images may be easier to distinguish and refer to than some sets used with adults, leading to overall shorter utterances. We specifically targeted children's abilities to construct referring expressions that can be jointly understood. Children were provided scaffolding around the larger coordination problems of taking turns and talking to their partner.

In the broader picture of language acquisition, there's debate over whether children learn the facts of their language first, followed much later, by pragmatic and communicative competency (CITATIONS), or whether sensitivity to communicative intent is an early emerging skill that develops in parallel with linguistic knowledge and may bootstrap language learning (CITATIONS). The findings in the current work are most consistent with a gradual development of children's communicative and linguistic skills, where the skills emerge early and then are refined over time, as children's cognitive capacities increase. At 4-5-years old, children are able to use their linguistic skills communicatively, even as their utterances are not fully adult-like in form.

References

Abbot-Smith, K., Nurmsoo, E., Croll, R., Ferguson, H., & Forrester, M. (2016). How children aged 2;6 tailor verbal expressions to interlocutor informational needs. *Journal of Child Language*, *43*(6), 1277–1291. https://doi.org/10.1017/S0305000915000616

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2020). Empirica: A virtual lab for high-throughput macro-level experiments. *arXiv:2006.11398 [Cs]*. https://arxiv.org/abs/2006.11398

Bohn, M., Kachel, G., & Tomasello, M. (2019). Young children spontaneously recreate core properties of language in a new modality. *Proceedings of the National Academy of Sciences*, *116*(51), 26072–26077. https://doi.org/10.1073/pnas.1904871116

Boyce, V., Hawkins, R. D., Goodman, N. D., & Frank, M. C. (2024). Interaction structure constrains the emergence of conventions in group communication. *Proceedings of the National Academy of Sciences*, *121*(28), e2403888121. https://doi.org/10.1073/pnas.2403888121

Branigan, H. P., Bell, J., & McLean, J. F. (2016). Do You Know What I Know? The Impact of Participant Role in Children's Referential Communication. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00213

Clark, H. H., & Wilkes-Gibbs, D. (1986). *Referring as a collaborative process*.

Glucksberg, S., & Krauss, R. M. (1967). WHAT DO PEOPLE SAY AFTER THEY HAVE LEARNED HOW TO TALK? STUDIES OF THE DEVELOPMENT OF REFERENTIAL COMMUNICATION. *Merrill-Palmer Quarterly of Behavior and Development*, *13*(4), 309–316. https://www.jstor.org/stable/23082551

Glucksberg, S., Krauss, R., & Weisburg, R. (1966). *Referential Communication in Nursery School Children: Method and Some Preliminary Findings*.

Graham, S. A., Sedivy, J., & Khu, M. (2014). That's not what you said earlier: Preschoolers expect partners to be referentially consistent. *Journal of Child Language*, *41*(1), 34–50. https://doi.org/10.1017/

S0305000912000530

Grigoroglou, M., & Papafragou, A. (2019). Interactive contexts increase informativeness in children's referential communication. *Developmental Psychology*, *55*(5), 951–966. https://doi.org/10.1037/dev0000693

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *arXiv:1912.07199 [Cs]*. https://arxiv.org/abs/1912.07199

Köymen, B., Schmerse, D., Lieven, E., & Tomasello, M. (2014). Young children create partner-specific referential pacts with peers. *Developmental Psychology*, *50*(10), 2334–2342. https://doi.org/10.1037/a0037837

Krauss, R. M., & Glucksberg, S. (1969). The Development of Communication: Competence as a Function of Age. *Child Development*, *40*(1), 255–266. https://doi.org/10.2307/1127172

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1-12), 113–114. https://doi.org/10.3758/BF03342817

Leung, A., Yurovsky, D., & Hawkins, R. D. (2024). Parents spontaneously scaffold the formation of conversational pacts with their children. *Child Development*, *n/a*(n/a). https://doi.org/10.1111/cdev.14186

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two- and Four-Year-Olds Learn to Adapt Referring Expressions to Context: Effects of Distracters and Feedback on Referential Communication. *Topics in Cognitive Science*, *4*(2), 184–210. https://doi.org/10.1111/j.1756-8765.2012.01181.x

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Applied Psycholinguistics*, *27*(3), 403–422. https://doi.org/10.1017/S0142716406060334

Matthews, D., Lieven, E., & Tomasello, M. (2010). What's in a manner of speaking? Children's sensitivity to partner-specific referential precedents. *Developmental Psychology*, *46*(4), 749–760. https://doi.org/10.1037/a0019657

Morisseau, T., Davies, C., & Matthews, D. (2013). How do 3- and 5-year-olds respond to under- and over-informative utterances? *Journal of Pragmatics*, *59*, 26–39. https://doi.org/10.1016/j.pragma.2013.03.007

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution. *Psychological Science*, *13*(4), 329–336. https://doi.org/10.1111/j.0956-7976.2002.00460.x

Nilsen, E. S., & Graham, S. A. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology*, *58*(2), 220–249. https://doi.org/10.1016/j.cogpsych.2008.07.002

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (No. arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084