

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 establishing reference. Children were 85% 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 the view that children show early pragmatic competence in supportive contexts.

Keywords: development; iterated reference game; pragmatics; preschoolers

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 describe a target so an interlocutor can pick it out from a set of possibilities. Adults show sensitivity to both the visual context and their audience during referential communication, calibrating the description 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. 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 classic study suggests that 4-5-year old preschoolers struggle with 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. While 4-5-year old children were successful on practice trials with familiar shapes and 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 correctly order the blocks. Similar experiments with older children indicated a gradual improvement through adolescence both for initial accuracy and for the increase in accuracy across 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 was difficult even among teenagers suggests that the complexity of the stacking paradigm and the large number of potential targets present significant task demands that might prevent accurate measurement of children's abilities.

More recently, a number of studies have revealed early emerging skills in the preschool years that support the use of referential communication. Preschool-aged children are faster to select a target when it is referred to in a consistent way (Graham et al., 2014; Matthews et al., 2010) and 6 year olds are more likely to use consistent referential expressions when their partner is consistent (Köymen et al., 2014), suggesting that young children are sensitive to the norm of consistent 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 communicate.

While children are sensitive to others' perspectives, they still seem to struggle with appropriately tailoring the specificity of their utterances to the visual context, often 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

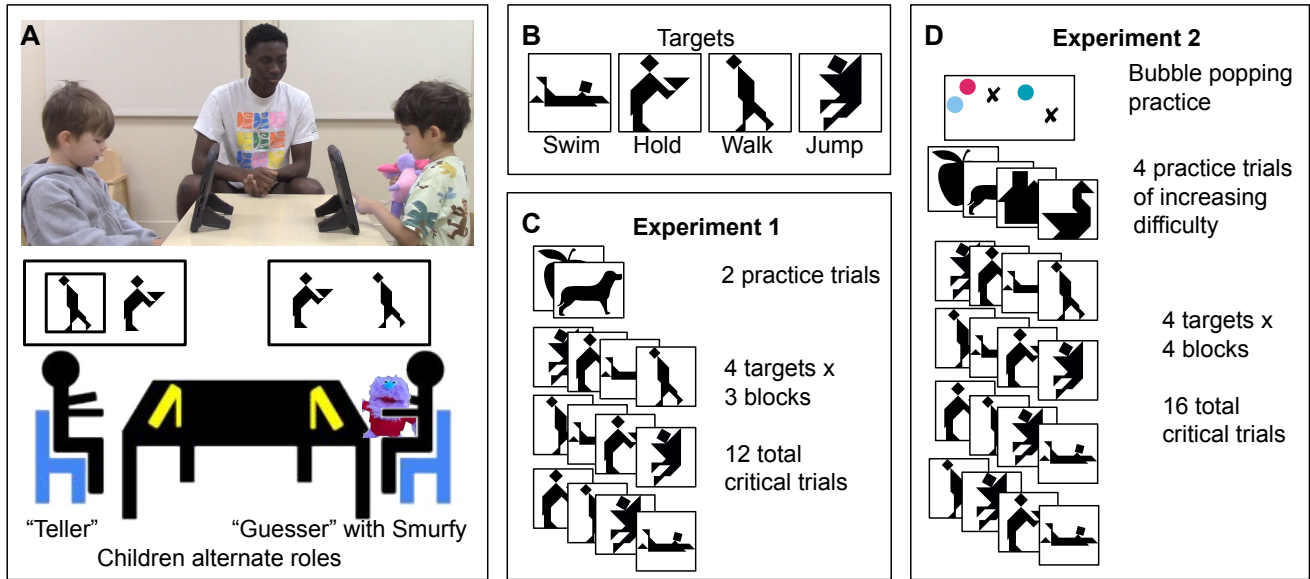


Figure 1: Experimental Setup and Procedure. Panel A shows the experimental setup with the teller and guesser across the table from each other. Panel B shows the 4 possible targets; names for targets are for cross-reference with later figures only. Panel C shows the procedure for Experiment 1; within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

& 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. when there are fewer possible referents to consider (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 adult-like patterns of increasing accuracy, increasing speed, and shorter descriptions across repetitions, but children’s accuracy was still far below adult performance and highly variable between dyads (Branigan et al., 2016). 4-6-year olds have also been shown to rely on conventionalized gestures with their partner in an iterated reference game where children could only use gestures to communicate (Bohn et al., 2019). Recent work has established that young children can succeed at iterated reference games when playing with their parents (Leung et al., 2024). 4-8 year olds’ played 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 about novel referents under the right conditions.

Given the task demands in Glucksberg et al. (1966) and work showing that children can succeed in a less demanding paradigm, we revisit the question of child-child referential communication. 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 found 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 claimed.

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 (Carruthers, 2013; Keen, 2003; Turan-Küçük & Kibbe, 2024), so we adapted the experimental framework from Leung et al. (2024), and further simplified it by reducing the total pool of targets and the number of trials. This experiment was pre-registered at <https://osf.io/kcv8j>.

Participants 4 and 5-year old children were recruited from a university preschool 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 included incomplete 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 away from the game.

Procedure Once a pair of children agreed to play the game, a research assistant took them to a quiet testing room. 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 was the “teller” and saw a black box around one of two images on their screen and was asked to “say what they saw” in the black box. The “guesser” saw the same two images in a randomized order and tried to select the described image to help Smurfy make a match. When the guesser selected an image, both children received feedback in form of a smiley or frowny face and an audible sound. After each trial, children switched roles. 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 target images (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 they did scaffold the interaction, prompting children to describe the images, and sometimes repeating children’s statements (especially when utterances were inaudible or the child did not respond immediately; this aspect of the procedure was modified in Experiment 2). 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 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 brms (Bürkner, 2018) with weakly informative priors. We present the estimate and 95% credible intervals.

Results

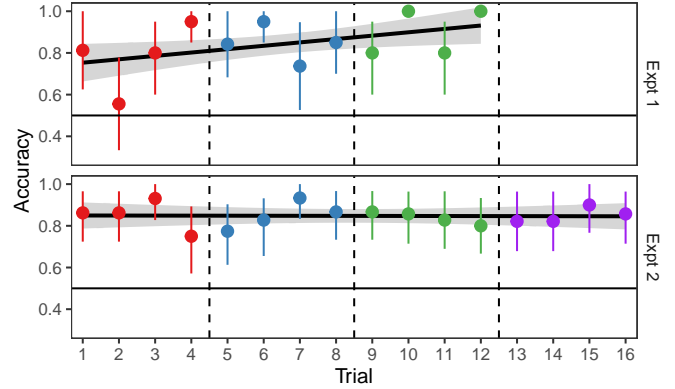


Figure 2: Children’s accuracy at selecting the correct target over time. Experiment 1 had 3 blocks (12 total critical trials) and experiment 2 had 4 blocks (16 critical trials). Error bars are bootstrapped 95% CIs with a linear trend line overlaid.

Accuracy and Speed Our primary measure of interest was whether children could accurately communicate the intended target. As shown in Figure 2, children generally accurate in their selections. To test for changes in accuracy over time, we fit a Bayesian mixed effects logistic regression predicting accuracy.¹ Children’s accuracy was above chance (Odds Ratio: 3.00 [1.14, 8.09]), and their accuracy slightly increased over the game (OR of one trial later: 1.17 [1.03, 1.39]). This level of accuracy is generally in-line with accuracies from 4-year-olds playing with their parents in Leung et al. (2024) and indicates that children can understand and succeed at the task.

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 linear regression predicting the time to selection in seconds.² The first critical trial averaged 27.48 [20.48, 34.53] seconds, and children got faster over time (-1.22 [-1.90, -0.53] seconds / trial). This is a measure of efficiency as children were able to achieve the same accuracy levels in less time.

Description length In iterated reference games with adults, description lengths usually shorten over repeated references (Boyce et al., 2024; Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2020). We were curious if children’s descriptions would display the same trend, so we ran a Bayesian mixed effects linear regression predicting the number of words in the description the “teller” produced.³ On the first critical trial, descriptions averaged 3.66 [2.56, 4.84] words and description length was relatively stable over time (change of 0.12 [-0.05, 0.27] words per trial, Figure 3). Thus, while children were taking less time per trial, their speed-up was not from shorter utterances, instead they were understanding the game better and taking less time to initiate their descriptions and make their selections.

¹correct.num~ trial.num + (trial.num|game) + (1|target)

²time.sec~ trial.num + (trial.num|game) + (1|target)

³words~ trial.num + (trial.num|game) + (1|target)

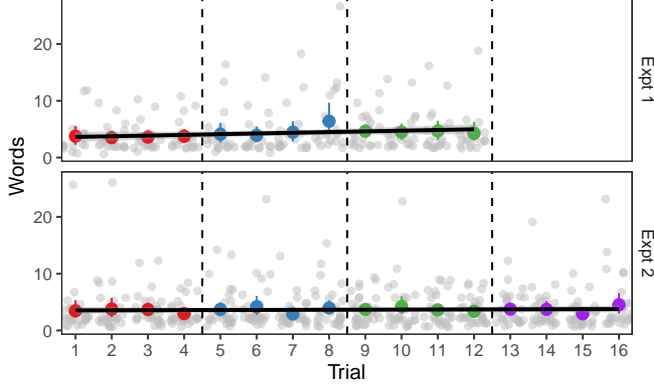


Figure 3: Length of referential expression (in words) produced by the teller each trial. Grey dots are individual data points, colored dots are per trial means with bootstrapped 95% CIs.

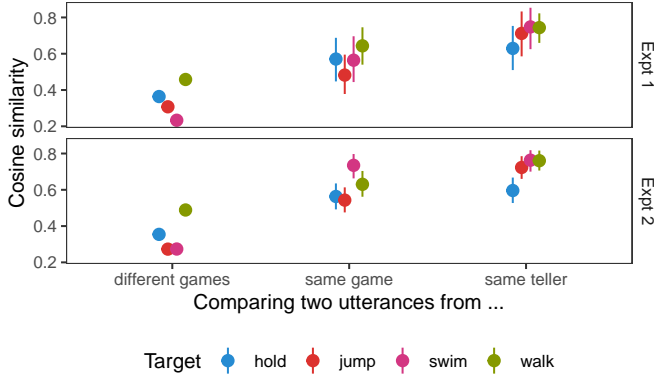


Figure 4: Semantic similarity between referential descriptions from the teller to the same target under different circumstances. Different games refers to similarities across two children from different games, same game to similarities across the two participants in one game, and same teller to descriptions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





Some examples to illustrate the variety of effective descriptions children employed are shown in Table 1; these specific examples are from Experiment 2, but are also representative of the types of descriptions produced in Experiment 1.

Convergence While description length is used as a proxy for measuring convention formation, semantic overlap between utterances does not always track description length. Boyce et al. (2024) introduced a more sensitive measure of semantic convergence. By comparing the content of utterances using word embeddings, Boyce et al. (2024) traced when how similarities between utterances in different rounds and games

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 (Reimers & Gurevych, 2019).

We compared the semantic similarities between descrip-

Table 1: Example descriptions children successfully used used to identify different target tangram images in Experiment 2.

<ul style="list-style-type: none"> • 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 	
<ul style="list-style-type: none"> • vampire • hopping • a person flying • a person • a kite • a triangle with a head on it with feet • somebody skydiving, not in the airplane 	
<ul style="list-style-type: none"> • racecar • airplane • alligator • a person fell down • a boat • a person that's in a race car that has one triangle and two triangles 	
<ul style="list-style-type: none"> • person • a person walking • a person looking down • a people, but it doesn't have any arms 	

tions of the same target based on who produced the description. 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 (Figure 4). We modeled this as a Bayesian mixed effects linear regression predicting similarity.⁴ Utterances were more similar if they came from the same partnership (0.222 [0.184, 0.260]) 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.

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 adequate descriptions, 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 a tighter experimental script and a larger sample size.

⁴ $\text{sim} \sim \text{same_game} + \text{same_speaker} + (1|\text{target})$

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>.

The biggest change between the experiments was increasing the number of repetitions of target stimuli from 3 to 4 (from 12 to 16 trials). The greater number of trials in Experiment 2 made it possible to look for changes over time that could be indicative of convergence to shared descriptions within a game and divergence between games.

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 older 3-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. In response to some children struggling with the abrupt switch from familiar 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 familiar shapes to less recognizable, blockier shapes (Figure 1D).

Procedure The procedure was much the same as Experiment 1 (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 Experiment 1). The experimental script was fully written out and memorized by experimenters 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 helped with game mechanics such as whose turn it was to tell and who should press the screen to move the game along, but avoided contributing or repeating any content about the images or the descriptions.

Data processing Data were 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.

Results

Accuracy In Experiment 2, children’s accuracy was above chance (Odds Ratio: 5.95 [3.07, 11.89]) and accuracy was relatively stable 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 (-0.70 [-0.99, -0.41] seconds / trial). 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 [2.20, 4.81] words and description length was relatively stable over time (change of 0.02 [-0.04, 0.09] words / trial, Figure 3). This is comparable to Experiment 1, again finding that children produce short utterances without much change in length over time. As with Experiment 1, children used a variety of descriptions successfully (Table 1).

Convergence As a coarse comparison, we repeated the analysis from Experiment 1. Utterances were more similar if they came from the same partnership (0.270 [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]).

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 using a Bayesian mixed effects linear regression predicting similarity.⁵ Over the first three blocks, descriptions became increasingly similar to the last block description (0.042 [0.007, 0.078]). Descriptions were more similar to the last block if they came from the same child who gave the description in the last block (0.067 [0.006, 0.127]). Although over time descriptions do get more similar to the last block utterance, the semantic distance between adjacent block utterances was relatively constant: 0.009 [-0.026, 0.044].

As each partnership converged to their shared nicknames, partnerships often diverged from one another as groups focused on distinct aspects of the image. We tested whether descriptions in different games to the same target diverged over time using a Bayesian mixed effects linear regression.⁶ As the games progressed, descriptions from different games became slightly further apart in semantic space (-0.013 [-0.018, -0.008]).

Qualitative variation in interaction structure 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; all children were paired with another child from the same classroom who 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 al-

⁵ $\text{sim} \sim \text{earlier_block.num} + \text{same_speaker} + (1|\text{game1}) + (1|\text{target})$

⁶ $\text{sim} \sim \text{block.num} + (1|\text{target})$

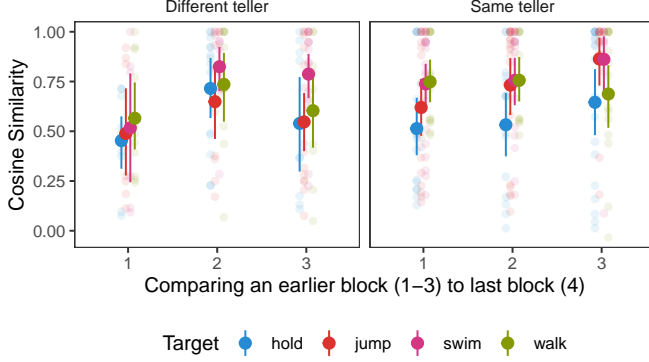


Figure 5: Semantic similarity between earlier blocks (1-3) with last block (4) for descriptions to the same image within the same group. Similarity measured as cosine similarity between S-BERT embeddings of referential descriptions. Heavy dots are means with bootstrapped 95% CIs; light dots are individual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

ways 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.

Joint analysis

As the two experiments were similar to one another, we re-ran models pooling the data across the two experiments, using experiment number as a random effect. Pooling the two experiments, children’s accuracy was above chance (Odds Ratio: 4.64 [1.42, 12.16]) 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 3.79 [-5.74, 14.14] words and description length was relatively stable over time (change of 0.04 [-0.02, 0.10] words per trial).

Utterances were more similar if they came from the same partnership (0.257 [0.234, 0.279]) and were slightly more similar still if they came from the same person within the partnership (0.109 [0.078, 0.140]).

As a post-hoc measure, we looked at whether descriptions that elicited correct selections had more staying power than descriptions that did not elicit correct selections. Here we operationalized “staying power” as similarity to the next block utterance, and we modeled it with $\text{sim} \sim \text{earlier_block.num} \times \text{correct} + \text{same_speaker} \times \text{correct} + (1|\text{game1}) + (1|\text{target}) + (1|\text{expt})$. Correct utterances had an increase in similarity with the next block (0.145 [0.055, 0.236]) with no substantial interaction with block number or same teller. Numerically, the size of the boost for being correct was similar to the size of the boost from coming from the same person 0.150 [0.062, 0.243].

General Discussion

Across 2 experiments and 51 pairs of 4 and 5-year old children, we tested how well children could produce referential expressions that allowed their partner to find a matching abstract shape. Children varied substantially in what sorts of descriptions they produced, but overall accuracy was high (85%), indicating that children were generally able to produce adequate descriptions. Additionally, children’s utterances showed signs of converging toward conceptual pacts. Children did not display a reduction in the length of referential expressions over the course of the game.

It is unclear to what extent the uniformly short descriptions are a product of the simplified task or children’s behavioral differences from adults. The appropriate length for an initial description depends on how easily describable the target is, as well as how large and close the contrast set is. In this case, the low number of options and relatively easy to describe shapes may have meant long initial descriptions were less necessary. The adult controls in Leung et al. (2024) used 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. Especially in light of other work suggesting that conceptual pact formation and reduction in utterance length sometimes decouple in adults (Boyce et al., 2024), further empirical work on the factors driving verbosity in reference games is warranted.

While this task is substantially scaled down relative to measures used for adult competence, it does suggest that the relevant communication skills are present at least in rudimentary form by the end of the preschool years.

Our task is limited by the non-representative population of children at university nursery schools who participated, and by the 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 later by pragmatic competency [for instance, semantics before pragmatic implicatures; Huang & Snedeker (2009), Noveck (2001)], or whether sensitivity to communicative intent is an early emerging skill that develops in parallel with linguistic knowledge and may bootstrap language learning (Bates, 1974; Bohn & Frank, 2019; Tomasello, 2008). 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, even though their utterances are not fully

adult-like in form, children are nonetheless able to use their linguistic skills communicatively.

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>
- Bates, E. (1974). Acquisition of pragmatic competence. *Journal of Child Language*, 1(2), 277–281. <https://doi.org/10.1017/S0305000900000702>
- Bohn, M., & Frank, M. C. (2019). *The Pervasive Role of Pragmatics in Early Language*.
- 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>
- Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the r package brms. *The R Journal*, 10(1), 395–411.
- Carruthers, P. (2013). Mindreading in Infancy. *Mind & Language*, 28(2), 141–172. <https://doi.org/10.1111/mila.12014>
- 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>
- Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, 45(6), 1723–1739. <https://doi.org/10.1037/a0016704>
- Keen, R. (2003). Representation of Objects and Events: Why Do Infants Look So Smart and Toddlers Look So Dumb? *Current Directions in Psychological Science*, 12(3), 79–83. <https://doi.org/10.1111/1467-8721.01234>
- 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>
- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2), 165–188. [https://doi.org/10.1016/S0010-0277\(00\)00114-1](https://doi.org/10.1016/S0010-0277(00)00114-1)
- 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>
- Tomasello, M. (2008). *Origins of human communication*. MIT press.
- Turan-Küçük, E. N., & Kibbe, M. M. (2024). Three-year-olds' ability to plan for mutually exclusive future possibilities is limited primarily by their representations of possible plans, not possible events. *Cognition*, 244, 105712. <https://doi.org/10.1016/j.cognition.2023.105712>