

THE FUTURE OF SCHOLARSHIP

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PSYCHOLOGY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Mary J. Stanford
February 2025

© Copyright by Mary J. Stanford 2025
All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Advisor T. Greatest) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(L. O. Sunshine)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Carl Friedrich Gauss)

Approved for the Stanford University Committee on Graduate Studies

Abstract

I'm an abstract right here, look at me!

Dedication

You can have a dedication here if you wish.

Acknowledgments

I want to thank a few people.

Contents

Abstract	v
Dedication	vi
Acknowledgments	vii
Introduction	1
1 The big fat chapter: Interaction structure constrains the emergence of conventions in group communication	2
1.1 Introduction	2
1.2 Results	5
1.2.1 Overview of experiments	5
1.2.2 Smaller and higher-coherence groups are more accurate	6
1.2.3 Smaller and higher-coherence groups are more efficient	8
1.2.4 Larger groups make greater use of matcher contributions	9
1.2.5 Descriptions converge faster in groups with thicker channels .	10
1.2.6 Games with thicker channels diverge from one another more quickly	12
1.3 Discussion	13
1.4 Materials and Methods	15
1.4.1 Participants	16
1.4.2 Materials	16
1.4.3 Procedure	17
1.4.4 Data pre-processing and exclusions	19

1.4.5	Modelling strategy	20
2	Kids	26
2.1	Introduction	26
2.2	Experiment 1	29
2.2.1	Methods	29
2.2.2	Results	31
2.2.3	Discussion	35
2.3	Experiment 2	35
2.3.1	Methods	35
2.3.2	Results	37
2.4	Joint analysis	39
2.5	General discussion	40
3	Processing stuff	43
	Conclusion	44
	References	45

List of Tables

2.1	Example descriptions children successfully used to identify different target images in Experiment 2.	36
-----	--	----

List of Figures

1.1	(A) Participants played a repeated reference game in groups of size 2 to 6. On each trial, a describer described the target image to the group of matchers. Each image appeared once per block for six blocks. (B) Experiments varied along 3 dimensions: Group size, group coherence, and matcher contributions. (C) Experiment 1 (pink) varied group size from 2 to 6 players while holding group coherence and matcher contributions constant. Experiment 2 (blue) held group size constant at 6 and manipulated the other dimensions. Experiment 3 (green) tested 4 corners of the space, crossing group size (2 vs. 6 players) with the thickness of interaction structure (high vs. low coherence and matcher contributions).	22
1.2	Behavioral results across all three experiments. (A-C). Matcher accuracy at selecting the target image. (D-F). Number of words produced by the describer each trial. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a few outliers points are not visible.	23

1.3	Example utterances describing the shown tangram figure produced by two 3-player games in Experiment 1. To measure convergence within a game (blue), we measured the cosine similarity between SBERT embeddings of descriptions and the embedding of the round 6 utterance (taken to be the convention). Higher cosine similarity indicates more similar meaning. To measure divergence between games (green), we measured the similarity between embeddings of utterances from the same round across games.	24
1.4	Language similarity results measured with pairwise cosine similarity between embeddings of two utterances. (A-C). Convergence of descriptions within games as measured by similarity between an utterance from block 1-5 to the block 6 utterance in the same game for the same image. (D-F). Divergence of descriptions across games as measured by the similarity between two utterances produced for the same image by different groups in the same block. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a few outliers points are not visible.	25
2.1	Experimental setup and procedure. Panel A shows the experimental setup. 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.	28
2.2	Children’s accuracy at selecting the correct target over time. Error bars are bootstrapped 95% CIs with a linear trend line overlaid. . .	32
2.3	Length of description produced by the teller each trial. Grey dots are individual data points, colored dots are per trial means with bootstrapped 95% CIs.	33
2.4	Semantic similarity between pairs of descriptions from different sources. Dots are means and lines are bootstrapped 95% CIs.	34

2.5	Semantic similarity between descriptions from earlier blocks (1-3) and the last block in Experiment 2. Heavy dots are means with bootstrapped 95% CIs; light dots are individual values.	39
-----	--	----

Introduction

blah blah goes here

Chapter 1

The big fat chapter: Interaction structure constrains the emergence of conventions in group communication

1.1 Introduction

Much of human social life revolves around communication in groups. At school, teachers address large classrooms of children (Cazden, 1988); at home, we chat with groups of friends and family members over dinner (Tannen, 2005); and at work, we attend meetings with colleagues and managers (Caplow, 1957; Zack, 1993). Such settings present considerable challenges that do not arise in the purely two-party (dyadic) settings typically studied in psychology (H. Branigan, 2006; Ginzburg & Fernandez, 2005; Traum, 2004). For example, producers need to account for the fact that different comprehenders in the group may have different mental states or levels of background understanding (Fox Tree & Clark, 2013; Horton & Gerrig, 2002; Horton & Gerrig, 2005; Weber & Camerer, 2003; Yoon & Brown-Schmidt, 2014, 2018), while comprehenders must account for the fact that utterances are not necessarily tailored to them (Carletta et al., 1998; Cohn-Gordon et al., 2019; Fay et al., 2000; Metzling & Brennan, 2003; Rogers et al., 2013; Tolins & Fox Tree, 2016; Yoon & Brown-Schmidt,

2019). What enables producers and comprehenders to nevertheless overcome these challenges and navigate multi-party settings with relative ease?

One promising set of hypotheses centers on the group’s *interaction structure*, the set of constraints placed on the group’s shared communication channel. Many different aspects of interaction structure have been implicated in the effectiveness of dyadic communication, including the availability and quality of concurrent feedback (Krauss & Bricker, 1967; Krauss & Weinheimer, 1966; Kraut et al., 1982), the bandwidth of the communication modality (Dewhirst, 1971; Krauss et al., 1977), and the group’s access to a shared workspace (Clark & Krych, 2004; Garrod et al., 2007). Yet larger groups introduce qualitatively different dimensions of interaction structure, leading to a large but often inconsistent body of findings even for these well-understood factors (Hiltz et al., 1986; Swaab et al., 2012). While communication is generally expected to deteriorate as groups get larger (MacMillan et al., 2004; Seaman & Basili, 1997), the structural “thickness” of the feedback channel may slow such deterioration (Ahern, 1994; Parisi & Brungart, 2005).

In this paper, we develop an experimental paradigm for evaluating the relative contribution of these factors: a *multi-party repeated reference game*. The ability to distinguish one particular entity from other possible entities, known as *reference*, is one of the most primitive and ubiquitous functions of communication. Reference games (Lewis, 1969; Wittgenstein, 1953) have been widely used to study dyadic communication under controlled conditions in the lab. They provide a clear metric of communicative effectiveness: how many words are required before a matcher successfully chooses a target image from a context of distractors? *Repeated* reference games, where the same target images appear multiple times in succession, were introduced to examine how interlocutors establish shared reference in the absence of conventional labels (Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964). At the beginning of the game, long and costly descriptions are typically required to succeed. A key finding, however, is that dyads become increasingly efficient over the course of interaction. Fewer words are required to achieve the same accuracy, but referring expressions also become more impenetrable to outsiders (Schober & Clark, 1989; Wilkes-Gibbs & Clark, 1992). The evolution of referring expressions over repetitions shows the characteristic dynamics of conventions: *stability*, or convergence

on labels within a group, and *arbitrariness*, or divergence to different across groups, suggesting that dyads leverage their shared communication history to coordinate on expectations about how to label the target images (Hawkins et al., 2023).

In principle, repeated reference games provide a strong operationalization of communicative effectiveness for the problem of multi-party communication: describers must simultaneously achieve shared reference with multiple matchers. However, empirically studying multi-party communication raises a number of difficulties in practice. A much larger pool of participants must be recruited to achieve sufficient power at the relevant unit of analysis – the group – spanning a very high-dimensional space of possible parameter settings (Almaatouq et al., 2022). We address this problem by drawing on recent technical advances that have made it newly possible to achieve such samples using interactive web-based platforms (Almaatouq et al., 2021; Haber et al., 2019; Hawkins et al., 2023). Repeated reference games in web-based platforms have previously replicated earlier results from face-to-face studies, although people produce fewer words in text modalities than oral modalities (Hawkins et al., 2020b). The text-based chat modalities arguably more closely resemble the interfaces used by modern teams who increasingly communicate through group text threads or popular platforms like Slack or Discord.

We leverage our platform to explore effects of group size and interaction channel thickness in a series of three experiments. While we find that small groups reliably converge on group-specific “shorthand” regardless of the interaction structure, larger groups require thicker channels – richer conversational feedback among members – to achieve the same degree of coherence. Thus, increasing group size alone does not impede communication; rather, larger groups may require stronger social and linguistic cues to establish common ground among all members. More broadly, our work suggests that studying communication in larger groups is necessary to unveil critical aspects of interaction structure that have not been evident in typical dyadic settings.

1.2 Results

We recruited 1319 participants through Prolific, an online crowd-sourcing platform. Participants were organized into 313 groups of size two to six for a communication game (Figure 1.2A). On each trial, everyone in the group was shown a gallery of 12 tangram images (Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2020b; Ji et al., 2022). One player was designated the *describer* and the others were designated the *matchers*. The describer was asked to use a chat box interface to describe a privately indicated *target* image. After all matchers guessed which of the 12 images was the target, they received task feedback and proceeded to the next trial. The game consisted of 72 trials structured into 6 repetition blocks, where each image appeared as the target exactly once per block.

We manipulated the interaction structure of this game across 11 distinct conditions in 3 distinct pre-registered experiments (Figure 1.2B). We systematically sampled points along four dimensions parameterizing different aspects of the interaction space. We manipulated *group size* (ranging from two to six), *role stability* (whether or not participants took turns in the describer role), richness of *task feedback* (whether or not matchers were able to see each other’s responses), and richness of the *matcher contributions* (whether matchers were able to freely respond through a chatbox or could only use emojis; Figure 1.2C). Other factors, such as the set of stimuli and background knowledge about one’s partners, were held constant across games.

1.2.1 Overview of experiments

Experiment 1 began by investigating how performance scaled with group size. Based on prior qualitative work, we predicted that larger groups face a more challenging coordination problem. We continuously varied the number of players from 2 to 6 while keeping other factors constant. For these conditions, the describer role rotated after each block, so that all players had at least one turn as describer. Matchers had access to an unrestricted chat box, but only received binary task feedback about whether their individual selection was correct without revealing others’ selections or the intended target.

Experiment 2 focused on the most challenging 6-player groups and explored the

role of interaction structure. Each condition in Experiment 2 varied one aspect of the experiment relative to the Experiment 1 6-player baseline. We tried two variants that we expected to increase group coherence and improve performance, and a third variant we expected to interfere with the ability to establish mutual understanding and thus impede performance. In the first variant, we maintained the same describer throughout rather than a rotating describer, such that the same individual has the opportunity to aggregate feedback across trials and track which matchers are struggling with which targets. In the second variant, we gave the group of matchers full feedback about what every other member of the group had selected, and we showed the intended target. In the third variant, we changed how matchers could make contributions to the group. In contrast to prior experiments, where matchers could contribute freely to the chat; here, we limited matchers to sending four discrete emojis (green check, thinking face, red x, and laughing-crying face) that could convey simple valence and level of comprehension, but not any referential content.

Experiment 3 crossed the extremes of group size from experiment 1 (2 vs. 6 people) with the extremes of group interactions from Experiment 2 (*thick* vs. *thin* interaction structure). In the *thick* condition, we maintained a consistent describer, gave all matchers full task feedback, and allowed them to freely use a chat box. In the *thin* condition, we forced the describer to rotate on each block, restricted feedback to their own binary correctness, and restricted matcher contributions to the four emojis. Note that the 2-player thick game most closely resembles the design of classic repeated reference games (Clark & Wilkes-Gibbs, 1986).

1.2.2 Smaller and higher-coherence groups are more accurate

Our first set of hypotheses focused on group performance: how accurately and efficiently groups were able to perform the referential task. We characterize group performance along two complementary metrics: (1) matcher accuracy and (2) describer efficiency. Matcher accuracy is given by the percent of matchers on each trial who successfully selected the target referent. Describer efficiency is given by the number of words produced by the describer to achieve that degree of matcher accuracy in the group. The degree to which describers are able to communicate more efficiently

without negatively impacting matcher accuracy is indicative of convergence on a more effective shared communication protocol within the group.

We begin by examining matcher accuracy, the extent to which the intended target was reliably transmitted to all matchers. We constructed a series of 5 logistic mixed-effects regression models predicting accuracy as a function of condition and repetition block (separate models were run for experiment 1, each condition in experiment 2, and experiment 3). For this and other effects, there was substantial variation at the tangram and game levels, with some tangrams being markedly easier than others and some groups performing differently than others. This wide variation made it difficult to precisely estimate population-level main effects, leading to wide credible intervals. See SI Figure 11 for a visualization of the relative magnitudes of population effects and game and tangram level variations.

Across all conditions, we observed strong positive effects of repetition block, indicating improved performance over time (Figure 1.2A-C, SI Tables 4-8). In Experiment 3, larger games began with lower initial accuracy ($\beta = -0.64$, 95% CrI = $[-1.05, -0.25]$) and improved more slowly ($\beta = -0.34$, 95% CrI = $[-0.43, -0.25]$) than smaller games, although group size differences were not reliable in Experiment 1 (SI Table 4), and these experiment 3 differences were not robust in a sensitivity analysis (SI Figure 4C and SI Table 58). Among large groups in Experiment 2, accuracy was higher in the thicker conditions than in the condition with thin interaction structure (SI Tables 5-7), although effects of game thickness were not reliable in Experiment 3 (SI Table 8).

Because each experiment only explored a slice of the full parameter space, we also considered an exploratory analysis that pooled data across experiments, aiming to mitigate the loss in power from running entirely separate regression models. Specifically, we aggregated data from all experiments into a post-hoc mega-analytic model predicting accuracy as a function of repetition block, game thickness (thin v. not-thin) and game size. Overall, we found evidence that accuracy increased over time ($\beta = 0.46$, 95% CrI = $[0.4, 0.52]$) but the rate of increase was reduced for thin games ($\beta = -0.12$, 95% CrI = $[-0.21, -0.02]$) and larger games ($\beta = -0.07$, 95% CrI = $[-0.09, -0.05]$) compared to smaller or thicker games. That is, smaller groups and groups with higher coherence tended to be more accurate,

though the magnitude and reliability of these effects varied across individual experiments.

1.2.3 Smaller and higher-coherence groups are more efficient

After establishing that groups were able to communicate accurately, we turned to the challenges faced by describers when deciding how much information to provide. Specifically, we predicted that larger and more heterogeneous groups may initially require more information, but that thicker interaction structure may similarly allow describers to communicate more effectively over time. We tested these predictions using linear mixed-effects models predicting the number of words a describer produced on each trial as a function of condition and block. These models counted all words the describer produced, including after matcher contributions (similar effects were found in models predicting the length of describer's utterances before any matcher contributions, see SI Tables 21-24).

First, as predicted, describers in larger groups produced longer descriptions at the outset than describers in smaller groups (Figure 1.2D-F). This effect held for the continuous manipulation of group size for Experiment 1 ($\beta = 1.6$, 95% CrI = $[0.62, 2.6]$) as well as the 2-person versus 6-person manipulation in Experiment 3 ($\beta = 7.51$, 95% CrI = $[3.63, 11.3]$). Smaller groups also continued to use shorter descriptions than larger groups over the course of the game. In Experiment 1, the rate at which efficiency increased was similar across different size groups ($\beta = -0.09$, 95% CrI = $[-0.37, 0.18]$). In Experiment 3, larger groups reduced faster than smaller ones ($\beta = -1.22$, 95% CrI = $[-2.06, -0.29]$), but the faster reduction did not fully make up for the longer initial starting point, and was not robust to a sensitivity analysis (SI Figure 4F and SI Table 63).

While thin 6-person games showed a flatter reduction trajectory than thicker 6-person games in Experiment 2 (SI Tables 10-12), there was no reliable effect of game thickness on reduction in Experiment 3 (SI Table 13).

The reduction patterns of description lengths is paralleled by how long matchers took to make selections; across conditions, matchers selected faster in later conditions (SI Figure 9), and the correlation between speed and description length was consistent across experiments (SI Figure 10).

Aggregating across experiments with a mega-analytic model, however, suggested that larger games were associated with steeper reduction ($\beta = -0.36$, 95% CrI = $[-0.51, -0.2]$) from a more verbose starting point ($\beta = 2.12$, 95% CrI = $[1.5, 2.75]$) than smaller games, and thin games had shallower reduction curves ($\beta = 0.79$, 95% CrI = $[0.04, 1.52]$) than thicker games. Overall, then, smaller games used shorter descriptions than larger games across various time points in the experiment, and thinner games reduced less than thicker games.

1.2.4 Larger groups make greater use of matcher contributions

As a final measure of group performance, we examined the back-and-forth interactions between the describer and the group of matchers. Matchers use their chat contributions to actively provide feedback, ask questions, offer alternative descriptions, and seek clarification about the describer’s referring expressions. Example transcripts from successful games, one in the 6-thick condition and one in the 6-thin condition, are shown in Table ?? . Additional examples are in the SI Tables 1 and 2. Overall, we found that larger groups displayed a higher proportion of trials where at least one matcher produced utterances (SI Figure 6A, $\beta = 0.79$, 95% CrI = $[0.58, 0.98]$), which declined across repetition blocks ($\beta = -0.8$, 95% CrI = $[-0.97, -0.62]$). On an individual level, a matcher in a larger group was more likely to make contributions than a matcher in a smaller group, although each contribution tended to be shorter (SI Figure 7, SI Tables 18, 20). The length of matcher interjections also decreased over time, especially for large groups (SI Figure 6D, $\beta = -0.41$, 95% CrI = $[-0.72, -0.11]$) consistent with the need for early matcher involvement in establishing referential conventions. Emoji use in Experiment 3 followed similar trends (SI Figure 8). Overall, describers in larger groups receive more total input from matchers, suggesting larger groups may require greater participation by matchers to reliably establish common ground.

1.2.5 Descriptions converge faster in groups with thicker channels

In the previous sections, we examined three metrics of communicative performance in groups of different sizes and interaction structures. We confirmed that groups in all conditions replicated the classic patterns of increasing accuracy and decreasing description length. We also found some initial evidence that larger groups may struggle to improve performance in the absence of thick communication channels. Here, we aim to better understand the mechanisms that allow describers to use shorter descriptions without sacrificing accuracy. In particular, we explore the hypothesis that interaction structure and group size affect performance through a *convention formation* process (Clark & Wilkes-Gibbs, 1986). Under a recent model of convention formation (Hawkins et al., 2023), groups are able to leverage their shared history to coordinate on stable expectations about how to refer to particular images. This model makes specific predictions about how interaction structure affects the ability to coordinate, in terms of the available feedback.

First, due to heterogeneity in the group – 6 individuals who may have diverging conceptualizations — a rational describer should provide a strictly more detailed initial description to hedge against multiple possible misunderstandings, as we previously observed. Second, all groups should display the characteristic dynamics of conventions: *stability*, or convergence within group, and *arbitrariness*, or divergence to multiple equilibria across groups. Third, convergence should be faster when a single individual is consistently in the describer role and when matchers are able to freely respond in natural language, as describers are able to aggregate feedback about the effectiveness of their own utterances from block to block and also immediately correct specific misunderstandings within a given trial.

To assess the dynamics of describer descriptions, we examine the *semantic similarity* of descriptions within and across games. We quantified description similarity by concatenating describer messages together within a trial and embedding this description into a high-dimensional vector space using SBERT. SBERT is a BERT-based sentence embedder designed to map semantically similar sentences to embeddings that are nearby in embedding space. Semantically meaningful comparisons between sentences are made by taking pairwise cosine similarities between the embeddings

(Reimers & Gurevych, 2019).

To measure stability, or convergence within groups, we compared utterances from blocks one through five to the final (block six) description for the same image from the same game. To measure arbitrariness, or divergence across groups depending on group-specific history, we compared utterances produced by different describers for the same image in the corresponding blocks. Figure 1.3 illustrates these two measures with example utterances and their within-game and between-game cosine similarities.

We modeled semantic convergence with a mixed effects linear regression model predicting the similarity between a block 1-5 utterance and the corresponding block 6 utterance as a function of the earlier block number and condition (Figure 1.4A-C; SI Tables 25-29). All conditions showed some convergence toward a conventional “short-hand” for the picture, but the speed of convergence was affected both by group size and channel width. First, we found that smaller groups reached stable descriptions faster than larger games. In Experiment 1, initial similarity was invariant across group size ($\beta = -0.008$, 95% CrI = $[-0.021, 0.005]$), but smaller groups converged faster (Figure 1.4A, $\beta = -0.008$, 95% CrI = $[-0.011, -0.005]$). In Experiment 3, 6-person thick games started off further from their eventual convention than 2-person thick games ($\beta = -0.069$, 95% CrI = $[-0.113, -0.025]$) but closed the gap over time (Figure 1.4C, $\beta = 0.009$, 95% CrI = $[0.001, 0.017]$, this effect was not robust to sensitivity analysis, SI Figure 5C and SI Table 68). Second, thicker games tended to converge faster than thin games (Figure 1.4B-C). In Experiment 3, small thin games started off slightly further from their convention than small thick games, and this gap widened over time ($\beta = -0.025$, 95% CrI = $[-0.033, -0.017]$). Finally, the combination of thin interaction structure and larger group hindered convergence more than either factor individually. Beyond the generally slower convergence in thin games, 6-person thin games showed substantially slower convergence even compared to 2-person thin games in Experiment 3 ($\beta = -0.035$, 95% CrI = $[-0.047, -0.025]$).

Pooling across experiments in a mega-analysis confirms this pattern. Thin games converge less than thick games overall ($\beta = -0.016$, 95% CrI = $[-0.025, -0.008]$), and *large* thin games are especially slow to converge ($\beta = -0.007$, 95% CrI = $[-0.01, -0.004]$). Across games, convergence towards the last utterance was driven by cumulative increasing similarity between pairs of utterances in adjacent blocks (SI

Figure 12D-F, SI Tables 40-44). In early rounds, descriptions could change substantially between rounds, but by later rounds, many descriptions had already reduced and solidified and varied little round to round. In summary, we found that stable descriptions emerged earlier if the group was smaller, or if the group had a thick interaction structure.

1.2.6 Games with thicker channels diverge from one another more quickly

While groups may initially overlap in their descriptions, including details of shapes or body parts, we predicted that their descriptions would become increasingly dissimilar as groups increasingly adapt to their own idiosyncratic shared history. To test this effect, we constructed a mixed-effects linear regression model predicting the cross-game similarity between a pair of utterances for the same image. A decrease in the similarity between different groups descriptions occurred in every condition, indicating increasing arbitrariness and group-specificity of descriptions (Figure 1.4D-F, SI Tables 30-34). However, different game sizes and interaction structures revealed very different strengths of divergence.

First, smaller games used more group-specific language. In Experiment 1, smaller games diverged more quickly than larger games ($\beta = 0.001$, 95% CrI = [0.001, 0.002]). In Experiment 3, 2-person thick games started off more dissimilar than 6-person thick games, although 6-person games diverged faster and eventually approached the dissimilarity levels of 2-person thick games (SI Table 34). Second, thicker interaction structure was associated with stronger group-specific divergence. In Experiment 3, 2-person thin games diverged more slowly than 2-person thick games ($\beta = 0.004$, 95% CrI = [0.002, 0.005]). As with the convergence patterns, large games with thin interaction structures had the flattest trajectories, as thinness and largeness compounded. In Experiment 3, 6-person thin games diverged even less than 2-player thin games (Figure 1.4F, $\beta = 0.017$, 95% CrI = [0.015, 0.019]), and in Experiment 2, 6-person thin games barely diverged at all (Figure 1.4E, $\beta = -0.004$, 95% CrI = [-0.006, -0.001]). A mega-analytic model confirms this pattern: thin games differentiate less between groups ($\beta = 0.005$, 95% CrI = [0.004, 0.007]) and large thin groups differentiate even less ($\beta = 0.004$, 95% CrI = [0.004, 0.005]).

As a complement to the embedding analysis, we also examined the frequency of a few classes of words in the descriptions. Literal geometric words (ex. square, triangle, etc) and words for body parts (leg, arm, etc) are common early in games, but decline over repetition in most conditions, to be replaced by more abstract descriptions that do not contain these classes of words (SI Figure 2). The 6-person thin condition, however, retains a higher level of literal geometric and body part words, along with high levels of positional words (above, left, below, etc) and posture words (kicking, standing, seated, etc), with a lower level of utterances that do not contain any of these classes of words.

1.3 Discussion

From classrooms to boardrooms, human communication often takes place in multi-party settings. However, experimental research rarely focuses on such settings, largely due to practical obstacles. In the current work, we asked how convention formation processes, typically studied in dyadic reference games, unfold in larger groups and under varying interaction structures. Across 3 online experiments and 11 experimental conditions, we varied multiple features of interaction structure including group size, modality of matcher contributions, and degree of group coherence. All conditions replicated classic dyadic phenomena: increasing accuracy and efficiency, semantic convergence within games, and differentiation of descriptions between groups. However, we also found that the interaction structure substantially affects how rapidly groups develop partner-specific conventions. Small groups may be able to successfully form conventions under limited feedback, but larger groups require thicker interaction structure. Multi-player groups may therefore reveal key factors which are masked in purely dyadic settings.

Increasing efficiency, for example, has often been taken as an index of group-specific convention formation (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Yoon & Brown-Schmidt, 2014, 2018). In our work, however, we observe distinct patterns for measures of raw utterance length compared to the dynamics of semantic content. In Experiment 3, thin 6-person games showed much less group-specific divergence despite comparable accuracy and efficiency. This gap raises the possibility

that it is possible to become more efficient and accurate without negotiating a unified group-specific label. Instead, they may be relying more strongly on the group’s priors (Guilbeault et al., 2021). Thus, we encourage measures of semantic content (and not just performance) when evaluating convention formation. The transcripts for these games provide a rich dataset for exploring different ways language is used to form referential conventions.

The causal mechanisms driving group size effects remain unclear. There are many differences between a 2-person group and a 6-person group that could plausibly lead to different outcomes. For example, in a dyad, producers can tailor their utterances to the one matcher, but in large groups, producers must balance the competing needs of different comprehenders (Schober & Clark, 1989; Tolins & Fox Tree, 2016; Yoon & Brown-Schmidt, 2018). These effects likely vary with the knowledge state and the communication channels available to comprehenders (Fox Tree & Clark, 2013; Horton & Gerrig, 2002; Horton & Gerrig, 2005). Further work digging into the language used and the interactions between participants might unearth plausible mechanisms for how differences in group size and interaction structure influence outcomes, pointing towards future experimental conditions.

Even within the boundaries of the repeated reference game paradigm, there is a high-dimensional space of possible experiments. We sampled only a few points along a few salient dimensions. In our experiment 3, we grouped some factors together in order to run more games in each condition: a fully factorial design would have been too expensive to power adequately. We instantiated a “thin” channel by limiting matchers to 4 discrete utterances (emojis), but there are other possible restrictions that could be placed on the channel, such as rate-limited typing or explicit time pressure. Future work could explore other dimensions of the interaction structure, introducing pre-existing relationships or familiarity among group members, alternative incentives involving competition and power, or alternative referential targets involving more complex concepts.

A particularly important dimension shaping interaction is the modality of communication, including whether whether the participants use oral or written language, whether they are co-present in the same space, and whether they have visual access to each others’ faces and gestures. Distinct modalities carry distinct affordances

and norms. In this work, we relied on a text-based chat modality without allowing co-presence or visual access.

We suspect that the general pattern of effects we see, in terms of group size and coherence, are likely to extend to other modalities. However, different modalities may allow for different strategies that may be more or less sensitive to group size, describer rotation, or different levels of matcher contributions. For instance, in face-to-face oral settings, it may be easier for describers to continuously talk until interrupted, or to monitor the comprehension of individual group members from their facial expressions.

In conclusion, narrowly focusing on the settings that are easy to study in the lab – dyads with rich communication channels – can lead to theories that mispredict how interactions play out in multi-party groups. By studying common ground and coordination across a wider range of interaction structures, we can develop a more nuanced understanding of the obstacles that stand in the way of successful communication and how groups can overcome them. This understanding can inform the design of policies and collaborative platforms that promote effective communication in various contexts, from small-scale conversations to large-scale civic discourse. As remote work and online communication become increasingly prevalent, it is increasingly crucial to understand how the structure of group communication environments shapes the effectiveness of human communication.

1.4 Materials and Methods

Our iterated reference task was implemented with Empirica (Almaatouq et al., 2021), a React-based web development framework for real-time multi-player tasks. Our experiments were designed sequentially and pre-registered individually.¹ We followed the pre-registered analysis plan for each experiment, although accuracy models were not explicitly specified until Experiment 3, and linguistic analyses were only verbally described starting with Experiment 2b. Results from some pre-registered models are omitted from the main text for brevity but are shown in the SI. Exploratory

¹Experiment 1: <https://osf.io/cn9f4> for the 2-4 player groups, and <https://osf.io/rpz67> for the 5-6 player data run later. Experiment 2: same describer at <https://osf.io/f9xyd>, full feedback at <https://osf.io/j5zbn>, and thin at <https://osf.io/k5f4t>. Experiment 3: <https://osf.io/untzy>

mega-analytic models pooling across the three experiments were not pre-registered.

All materials, data, and analysis code is available at <https://github.com/vboyce/multiparty-tangrams>.

1.4.1 Participants

This research was covered by the Stanford IRB under protocol 20009 “Online investigations of language learning”. Participants were recruited using the Prolific platform. All participants self-reported as fluent native English speakers on Prolific’s demographic prescreen. Experiment 1 took place between May and July 2021, Experiment 2 between March and August 2022, and Experiment 3 in October 2022. Each participant took part in only one experiment and was blocked from participating in subsequent experiments. As games with more participants tended to run longer, we paid participants different rates based on group size, with the goal of a consistent \$10 hourly rate. Participants were paid \$7 for 2-player games, \$8.50 for 3-player games, \$10 for 4-player games, and \$11 for 5- and 6-player games. When one player occupied the describer role for the entirety of a 6-player game, they were rewarded an additional \$2 bonus. Across all games, participants could earn up to \$2.88 in performance bonuses.

A total of 1319 people participated across the 3 experiments. We recruited enough participants for 20 games in each condition in experiments 1 and 2 and 40 games per condition in experiment 3. However, due to attrition in filling the games initially and due to participants dropping out of the games, we ended up with fewer games in some conditions. For logistical reasons of matching participants into real-time games, we had to recruit participants in fairly large batches, and so did not have precise control to add new games to replace games that did not fill or had participants drop out early. A breakdown of number of games and participants in each condition is shown in SI Table 3 along with further discussion of recruitment logistics.

1.4.2 Materials

The same 12 tangram images, drawn from Hawkins et al. (2020b) and Clark & Wilkes-Gibbs (1986), were used every block. These images were displayed in a 4×3 grid

with the order randomized across participants to disincentivize spatial descriptions such as “top left,” as the image might be in a different place on the describer’s and matchers’ screens. To reduce cognitive load from visual search, the locations were fixed for each participant across trials.

1.4.3 Procedure

The experimental procedure was very similar across the three experiments. We first describe the procedure used in Experiment 1 and then describe the differences in later experiments.

Experiment 1 Participants were directed from Prolific to our custom web application, where they were presented with a consent form and a series of instruction pages explaining the protocol. After finishing the instructions, they needed to pass a quiz to proceed. They were then directed to a “waiting room” lobby. Once the lobby filled to the required number of players, the game began. One lobby was filled before another was started; if a participant was waiting for 5 minutes, that lobby timed out, and the participant was paid without completing the experiment. Due to technical constraints with assigning participants to lobbies and games, only games of a single experimental condition could be active at a time. Thus, different conditions were run on different days or times of day.

One of the participants was randomly selected to begin in the role of describer, and the other participants were assigned to the role of matchers. On each trial, the describer saw a fixed array of tangrams with one tangram (privately) highlighted as the *target*. They were given a chat interface to communicate the target to the matchers, who were asked to determine which of the 12 images was the referential target. All participants were free to use the chat box to communicate at any time, but matchers could only make a selection after the describer had sent a message. Once a matcher clicked, they could not change their selection. There was no signal to the describer or other matchers about who had already made a selection. We recorded what all participants said in the chat, as well as who selected which image and how long they took to make their selections.

Once all matchers had made a selection (or a 3-minute timer ran out), participants

were given feedback and proceeded to the next trial. Matchers only received *binary* feedback about whether they had chosen correctly or not; that is, matchers who made an incorrect choice were not shown the correct answer (see SI Figure 1 for example feedback). The describer saw which tangram each matcher selected, but matchers did not see one another’s selections. Matchers got 4 points for each correct answer; the describer got points equal to the average of the matchers’ points. These points were translated into performance bonuses at the end of the experiment (1 point = 1 cent bonus). After the describer had described each of the 12 images as targets, in a randomized sequence, the process repeated with the same set of targets, for a total of 6 such repetition blocks (72 trials).

The same person was the describer for an entire block, but participants rotated roles between blocks. Thus, over the course of the 6 blocks, participants were describers 3 times in 2-player games, twice in 3-player games, once or twice in 4 and 5-player games, and once in 6-player games. Rotating the describer was chosen in this first experiment to keep participants more equally engaged (the describer role is more work), and to provide a more robust test of our hypotheses regarding efficiency and convention formation. After the game finished, participants were given a survey asking for optional demographic information and feedback on their experience with the game.

If a participant disconnected from the experiment, the game would stop.

Experiment 2 Experiment 2 consisted of three different variations on Experiment 1, all conducted in 6-player games. Each of these conditions differed from the Experiment 1 baseline in exactly one way. In the *same describer* condition, one person was designated the describer for the entire game, rather than having the describer role rotate. In the *full feedback* condition, all participants were shown what all others had selected as well as the identity of the correct target. This condition was similar to previous dyadic work, such as Hawkins et al. (2020b), where the correct answer was indicated during feedback. In the *thin* condition, we altered the chatbox interface for matchers. Instead of a textbox, matchers had 4 buttons, each of which sent a different emoji to the chat. Matchers were given suggested meanings for the 4 emojis during the instruction phase. They could send as many emojis as desired; for instance, they

might initially indicate confusion, and later indicate understanding. In addition, for the thin condition, we added notifications that appeared in the chat box marking the time when each player had made a selection.

Experiment 3 The thin channel condition in Experiment 3 was the same as the thin condition in Experiment 2. The thick condition combined the two coherency-enhancing variations from Experiment 2: the same participant remained in the describer role throughout, and full feedback was given about the correct answer and what all other players had selected. Across both conditions in Experiment 3, notifications were sent to the chat to indicate when a participant had made a selection. For experiment 3, game lobbies worked slightly differently, and 5 minutes after the first participant had joined the lobby, the game started if there were at least two participants. Correspondingly, in experiment 3, games did not stop if a player disconnected, instead if there were at least two players still active, the game continued, swapping a player into the role of describer if necessary to continue the game.

1.4.4 Data pre-processing and exclusions

Participants could use the chat box freely, which meant that the chat transcript contained some non-referential language. The first author skimmed the chat transcripts, tagging utterances that did not refer to the current tangram. These were primarily pleasantries (“Hello”), meta-commentary about how well the task was going, and bare confirmations or denials (“ok”, “got it”, “yes”, “no”). We excluded these utterances from our analyses. Note that chat lines sometimes included non-referential words in addition to words referring to the tangrams (“ok, so it looks like a zombie”, “yes, the one with legs”); these lines were retained intact.

In Experiments 1 and 2, games did not start if there were not enough participants and ended if any participant disconnected. In Experiment 3, games started after a waiting period even if they were not entirely full and continued even in the event that a participant disconnected (with describer role reassigned if necessary), unless the game dropped below 2 players. The distribution of player counts in games that were initially recruited to be 6 player games is shown in SI Figure 3. The realities of online recruitment and disconnection meant that the number of games varied between

conditions. We excluded incomplete blocks from analyses, but included complete blocks from partial games (See SI Table 3). When skimming transcripts to tag non-referential utterances, we noticed that one game in the 6-player thick condition had a describer who did not give any sort of coherent descriptions, even with substantial matcher prompting. We excluded this game from analyses.

1.4.5 Modelling strategy

We fit all regression models in brms (Bürkner, 2018) with weakly regularizing priors. We were unable to fit the full pre-registered mixed effects structure in a reasonable amount of time for some models, so we included the maximal hierarchical effects that were tractable. All model results and priors and formulae are reported in the SI. Models of accuracy used by-group random intercepts only, models of word count used full mixed effect structure, and models of S-BERT similarities used by-group and by-target random intercepts as applicable (see SI Figure 11). Models of matcher accuracy were logistic models with $\text{normal}(0,1)$ priors for betas and sd. Models of describer efficiency were run as linear models with an intercept prior of $\text{normal}(12,20)$, a beta prior of $\text{normal}(0,10)$, an sd prior of $\text{normal}(0,5)$ and a random-effect correlation prior of $\text{lkj}(1)$. For all of the models of SBERT similarity, we used linear models with the priors $\text{normal}(.5,.2)$ for the intercept, $\text{normal}(0,.1)$ for betas, and $\text{normal}(0,.05)$ for sd. As an additional post-hoc analysis, we ran mega-analytic models combining data across all experiments. For these models, we grouped the 3 thin-ish conditions (2c, and the two thin conditions of experiment 3) as one level, and coded the rest of the conditions as thick-ish. Game size was coded as a continuous measure (2 through 6). The priors for the mega-analytic models were the same as for the per-experiment models described above.

As a sensitivity analysis, we re-ran the primary models on the subset of the data from games that a) completed all 72 trials and b) had the full complement of players the entire time (relevant to 6-player experiment 3 games where games could start or continue with fewer players). Discrepancies are mentioned in the results, and these analyses are depicted in SI Figures 4 and 5 and SI Tables 54-73. We also needed to decide how to handle dropout in Experiment 3, as some of the 6-player games did not retain all 6 players for the entire game. Our decision was to follow an intent-to-treat

analysis and treat data as missing completely at random. Note that this choice underestimates differences between 2-player and (genuine) 6-player games by labeling some smaller groups as 6-player groups. We do not know exactly what leads some participants to drop out, but it is possible that some factors may be random (ex. connection issues) and others may be correlated with performance (ex. frustration because group is struggling).

We do not know whether groups that start and continue at the full size differ from games where some participants drop out. This is potentially an issue across all experiments; in experiments 1 and 2, groups stopped playing if anyone dropped out, and in experiment 3 they kept playing as a smaller group. The number of games in each condition and rates of dropoff are shown in SI Table 3 and SI Figure 3.

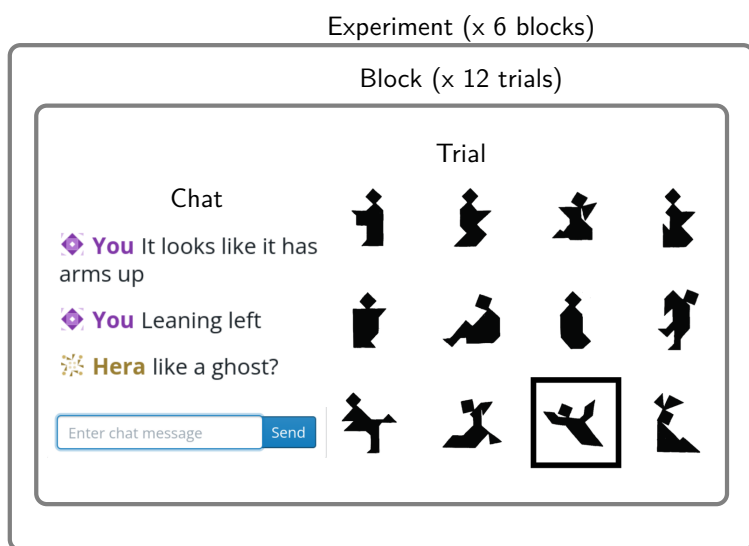
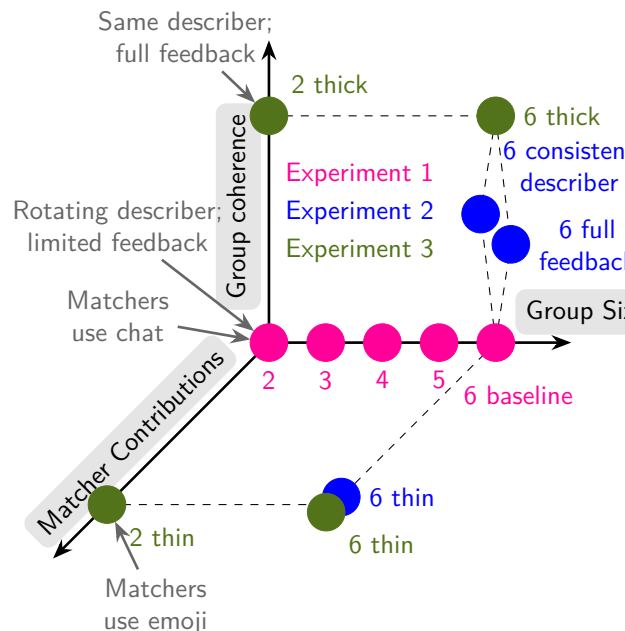
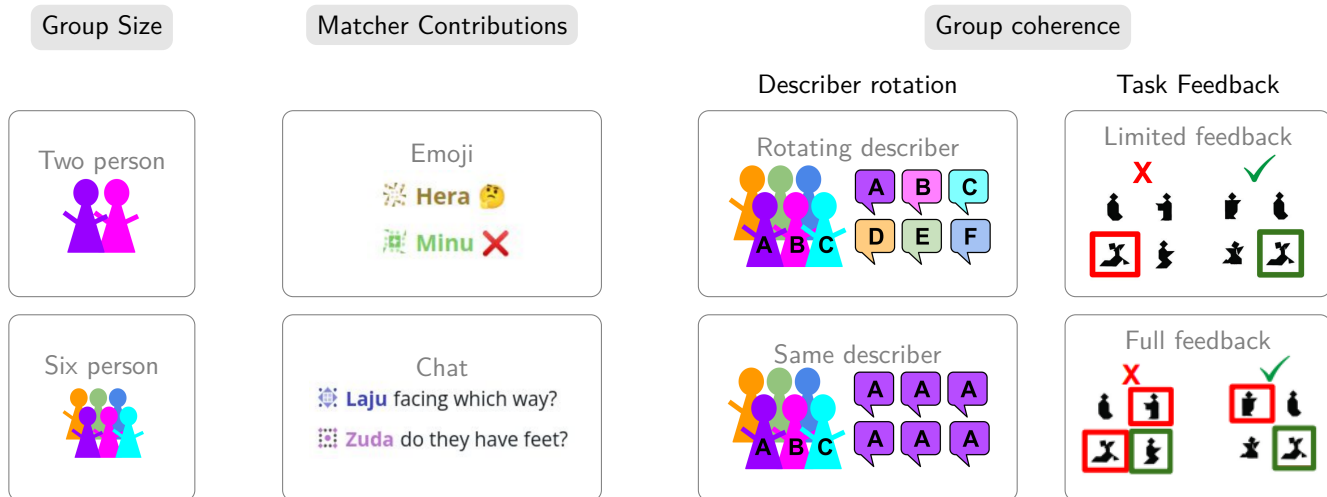
A**B****C**

Figure 1.1: (A) Participants played a repeated reference game in groups of size 2 to 6. On each trial, a descriptor described the target image to the group of matchers. Each image appeared once per block for six blocks. (B) Experiments varied along 3 dimensions: Group size, group coherence, and matcher contributions. (C) Experiment 1 (pink) varied group size from 2 to 6 players while holding group coherence and matcher contributions constant. Experiment 2 (blue) held group size constant at 6 and manipulated the other dimensions. Experiment 3 (green) tested 4 corners of the space, crossing group size (2 vs. 6 players) with the thickness of interaction structure (high vs. low coherence and matcher contributions).

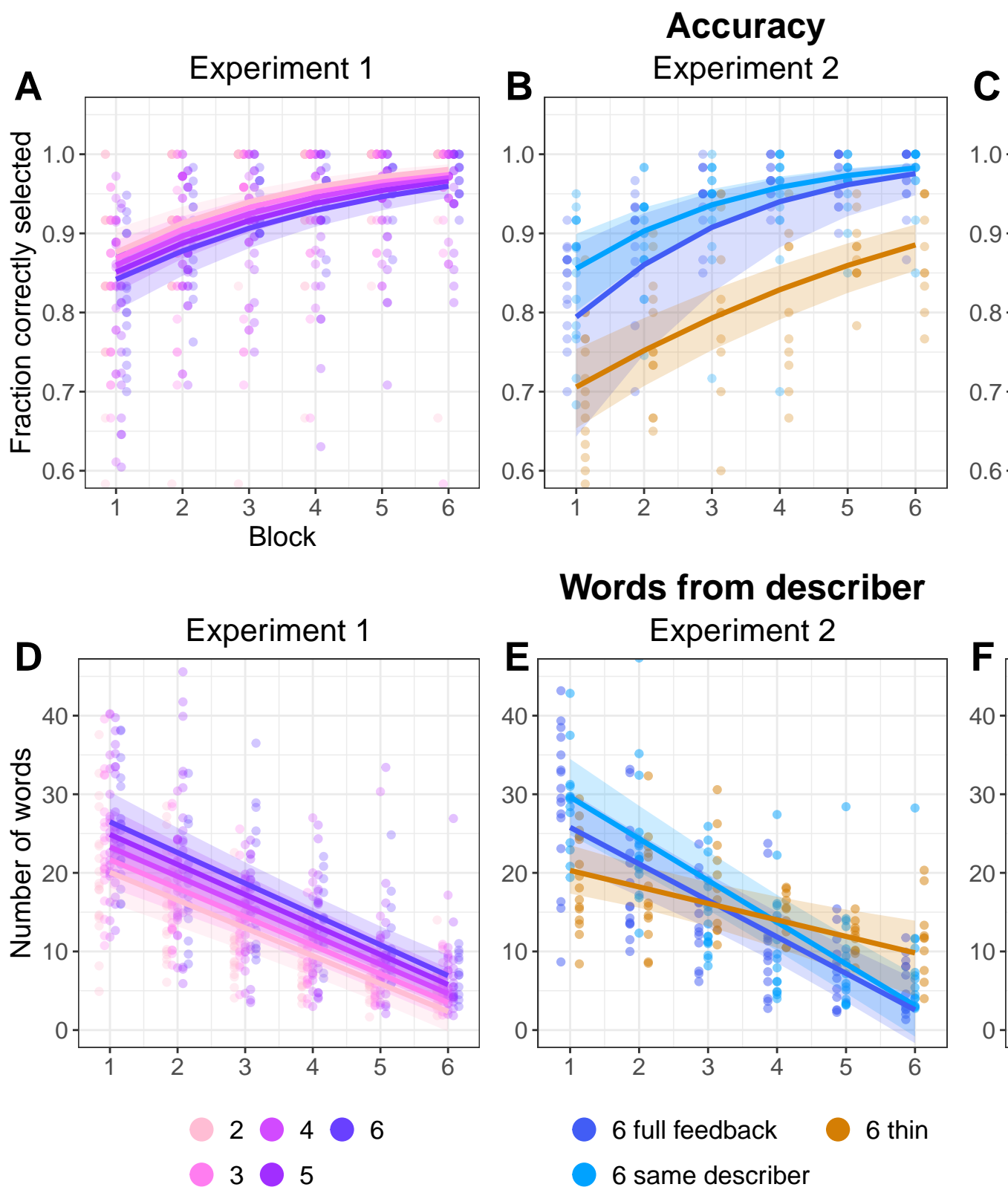


Figure 1.2: Behavioral results across all three experiments. (A-C). Matcher accuracy at selecting the target image. (D-F). Number of words produced by the describer each trial. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a

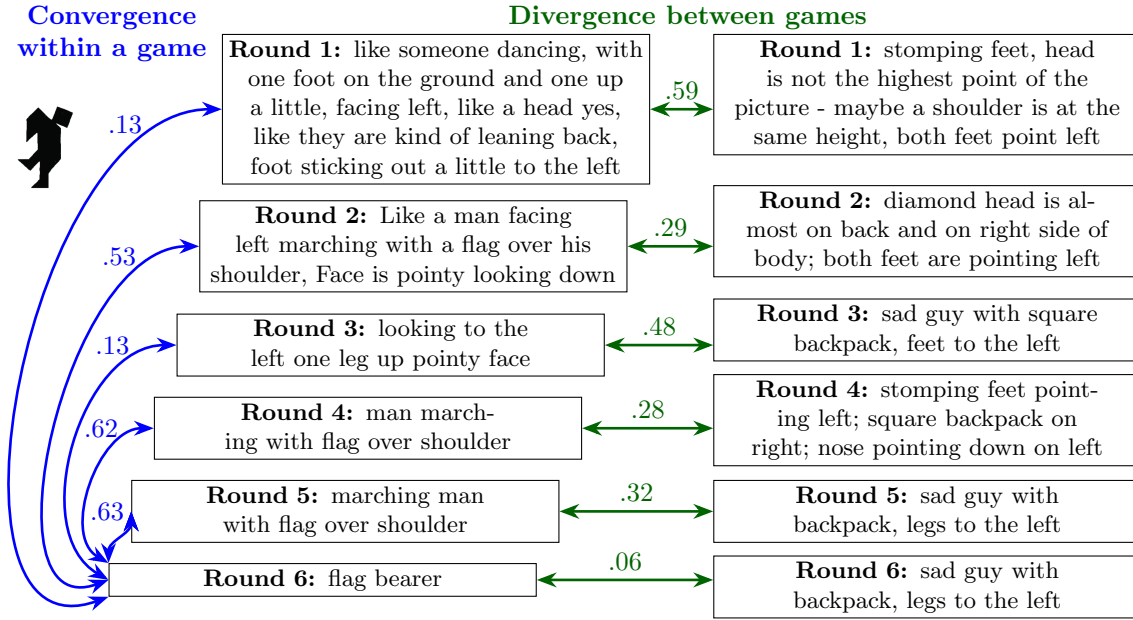


Figure 1.3: Example utterances describing the shown tangram figure produced by two 3-player games in Experiment 1. To measure convergence within a game (blue), we measured the cosine similarity between SBERT embeddings of descriptions and the embedding of the round 6 utterance (taken to be the convention). Higher cosine similarity indicates more similar meaning. To measure divergence between games (green), we measured the similarity between embeddings of utterances from the same round across games.

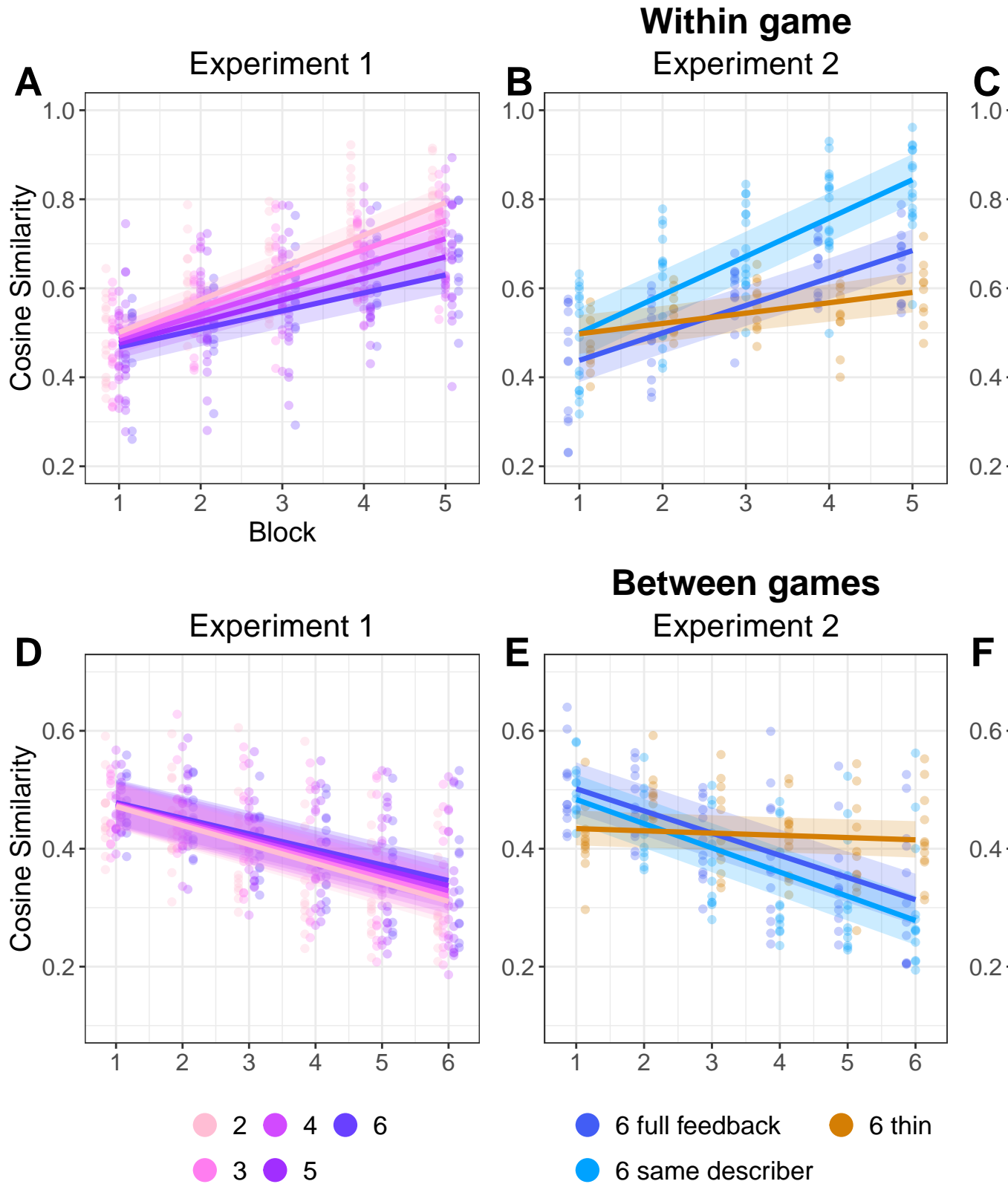


Figure 1.4: Language similarity results measured with pairwise cosine similarity between embeddings of two utterances. (A-C). Convergence of descriptions within games as measured by similarity between an utterance from block 1-5 to the block 6 utterance in the same game for the same image. (D-F). Divergence of descriptions across games as

Chapter 2

Kids

2.1 Introduction

Learning a language requires learning not only the content of that language, but also how to use the language to communicate. One case study 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.

Iterated reference games provide an important paradigm for studying referential communication. In these games, one player repeatedly describes a set of abstract shapes to a partner so they can identify the target images (Boyce et al., 2024; Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2020a; Krauss & Weinheimer, 1964). Over repetition, features of the initial 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 later referring expressions. 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.

One influential early study suggested 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 image on each block so their partner could pick out their corresponding block. As children described and selected blocks, they stacked them on pegs. While 4-5-year-old children succeeded on practice trials with familiar shapes and visual access to each other's blocks, children failed on critical trials where the blocks had abstract drawings and there was no visual access. Even after multiple rounds with the same images, children were not able to correctly order the blocks. Glucksberg et al. (1966) attributed children's communicative failures to their production of ego-centric descriptions that did not account for the other child's perspective.

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 adult college student sample (Glucksberg & Krauss, 1967; Krauss & Glucksberg, 1969). Given that even teenagers had difficulties with the task, the complex stacking paradigm and the large number of potential targets may have posed task demands that prevented accurate measurement of children's abilities.

More recently, though no evidence has directly contradicted Glucksberg et al. (1966), a number of other 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). By age 5, children integrate information about other's perspectives into their comprehension of utterances (San Juan et al., 2015). Overall, by late preschool, children can reason about others' perspectives to communicate more effectively.

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

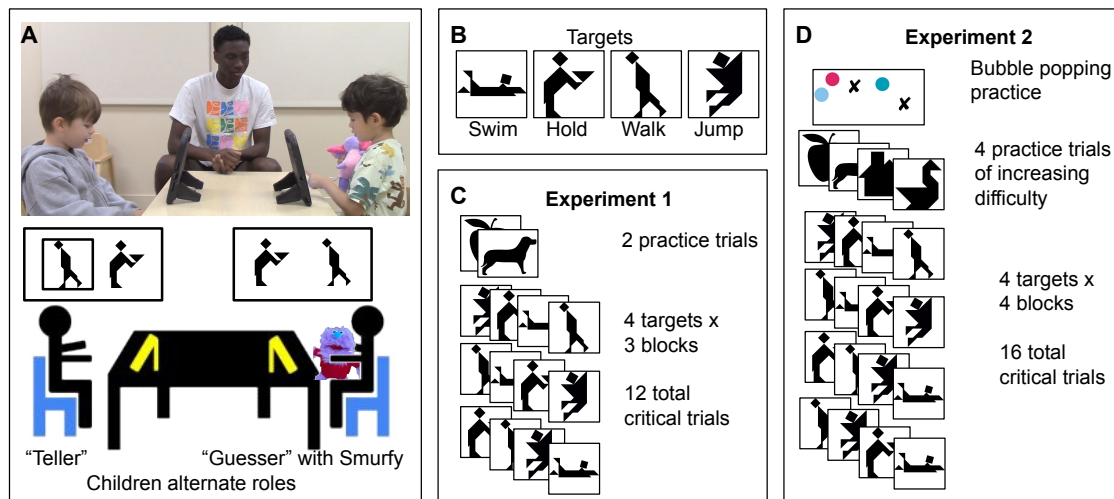


Figure 2.1: Experimental setup and procedure. Panel A shows the experimental setup. Panel B shows the 4 possible targets; names for targets are for cross-reference with later figures only. Panel C shows the procedure for Experiment 1; within critical blocks targets were ordered randomly. Panel D shows the procedure for Experiment 2.

in under-informative utterances (Leung et al., 2024; Matthews et al., 2012). When children must describe one of two very similar images, 4 and 5-year-olds sometimes neglect 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 over- and under-informative utterances, 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 among many developmental studies is that children perform better when the cognitive load of the task is reduced, i.e. when there are fewer possible referents (Abbot-Smith et al., 2016; San Juan et al., 2015). These findings paired with evidence of emerging communicative skills in young children suggest 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. In one study, 8-10-year-olds exhibited 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 (H. P. Branigan et al., 2016). 4-6-year-olds successfully used conventionalized gestures with their partners in an iterated reference game where children could only use gestures to communicate (Bohn et al., 2019). 4-8-year-old children succeeded at playing an iterated reference game with a parent using a simple, child-friendly tablet-based task (Leung et al., 2024). 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 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, here 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 an initial study and a replication including a total of 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.

2.2 Experiment 1

2.2.1 Methods

Our goal for Experiment 1 was to test young children’s ability to coordinate to produce descriptions of 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 [anonymized link](#).

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 games that completed 12 critical trials, and 1 game that completed 11 critical trials. 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 2.1B). We coded the matching game using Empirica (Almaatouq et al., 2020), hosted it on a server, and then accessed the game 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 2.1A). 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 “tell Smurfy what they see” 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 excited or disappointed sound. After each trial, children switched roles. Children passed Smurfy back and forth to 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 2.1C). 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 automatically 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 231 trials remaining.

Statistical analyses were run in brms (Bürkner, 2018) with weakly informative priors. We report estimates and 95% credible intervals. The experimental set-up, analysis code, de-identified transcripts, and performance data for both experiments is available at anonymized repo.

2.2.2 Results

Accuracy and speed

Our primary measure of interest was whether children could accurately communicate the intended target. 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], Figure 2.2). 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 children spent on each trial. We ran a Bayesian mixed effects linear regression predicting the

¹ $\text{correct.num} \sim \text{trial.num} + (\text{trial.num}|\text{game}) + (1|\text{target})$

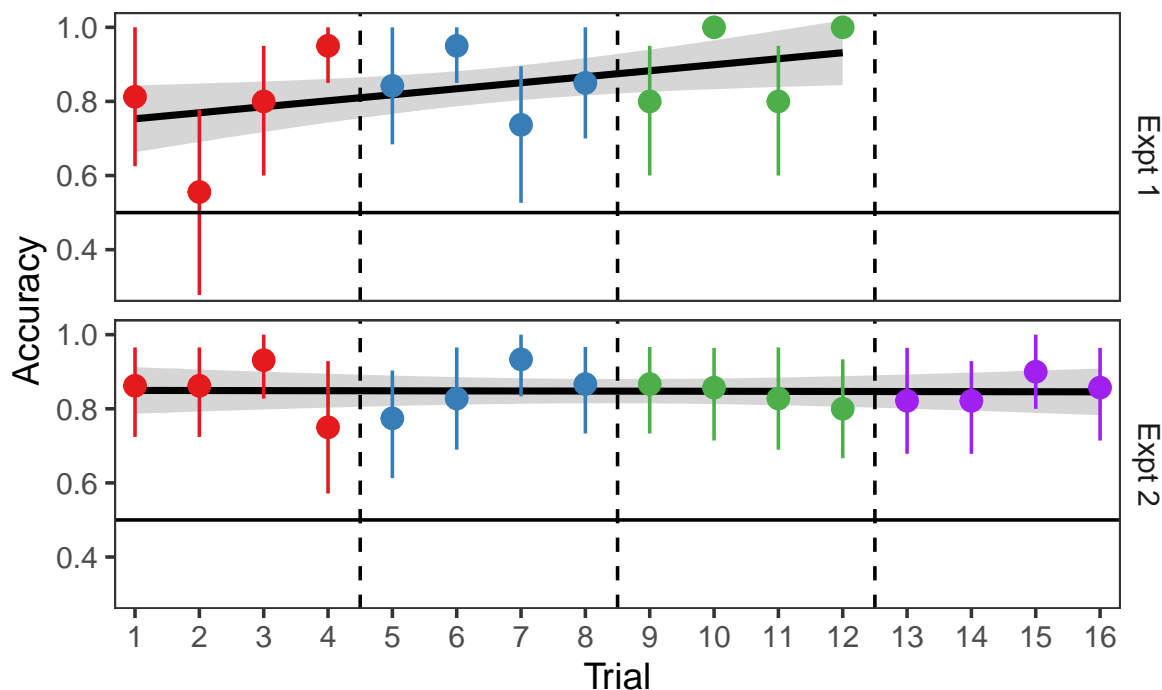


Figure 2.2: Children’s accuracy at selecting the correct target over time. Error bars are bootstrapped 95% CIs with a linear trend line overlaid.

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). Children were able to achieve the same accuracy in less time, suggesting that they were becoming more efficient at completing the task.

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., 2020a). 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, 4.74] words, and description length was relatively stable over time (change

² $\text{time.sec} \sim \text{trial.num} + (\text{trial.num}|\text{game}) + (1|\text{target})$

³ $\text{words} \sim \text{trial.num} + (\text{trial.num}|\text{game}) + (1|\text{target})$

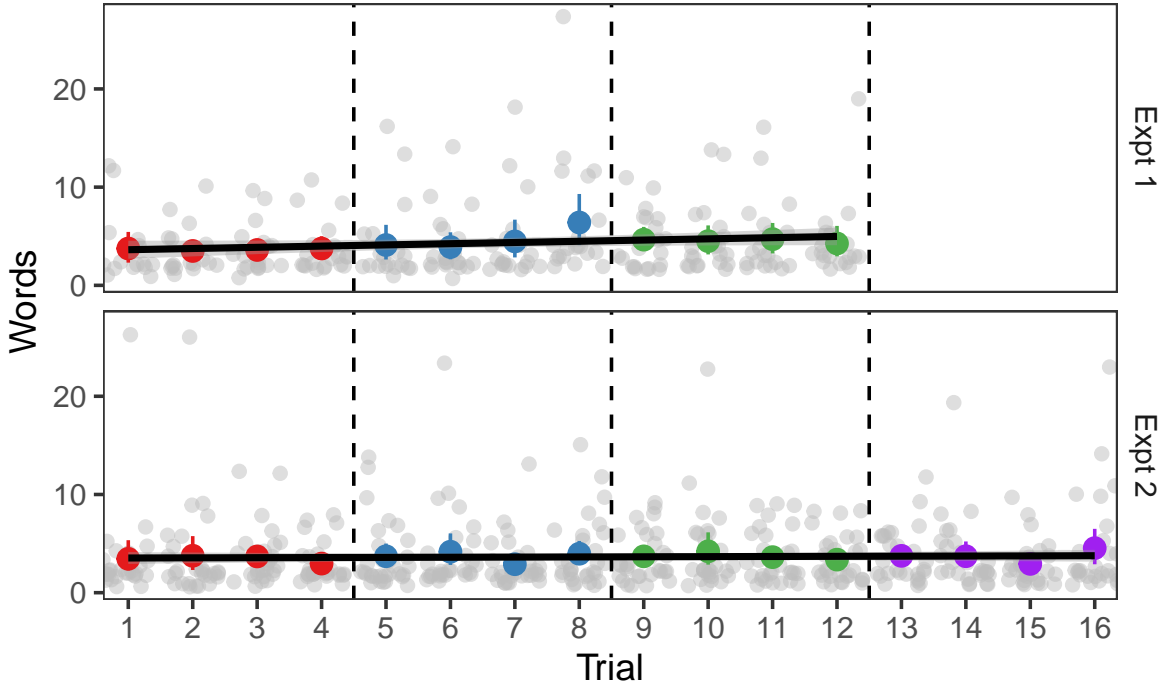


Figure 2.3: Length of description produced by the teller each trial. Grey dots are individual data points, colored dots are per trial means with bootstrapped 95% CIs.

of 0.12 $[-0.04, 0.28]$ words per trial, Figure 2.3). Thus, children’s increasing speed was not from shorter utterances, but instead from some combination of improved task understanding, faster utterance planning, and faster decisions of what to select.

Some examples to illustrate the variety of effective descriptions children employed are shown in Table 2.1 (these specific examples are from Experiment 2, but both experiments had similar distributions of descriptions).

Convergence

While description length is often used as a proxy for measuring convention formation, it does not capture semantic overlap between utterances. Boyce et al. (2024) introduced a more sensitive measure of semantic convergence that compares the content of utterances using word embeddings to trace how similarities within and across games change over time. With only 3 blocks of descriptions, we do not expect semantic similarity for descriptions of a given target to show any meaningful change over time.

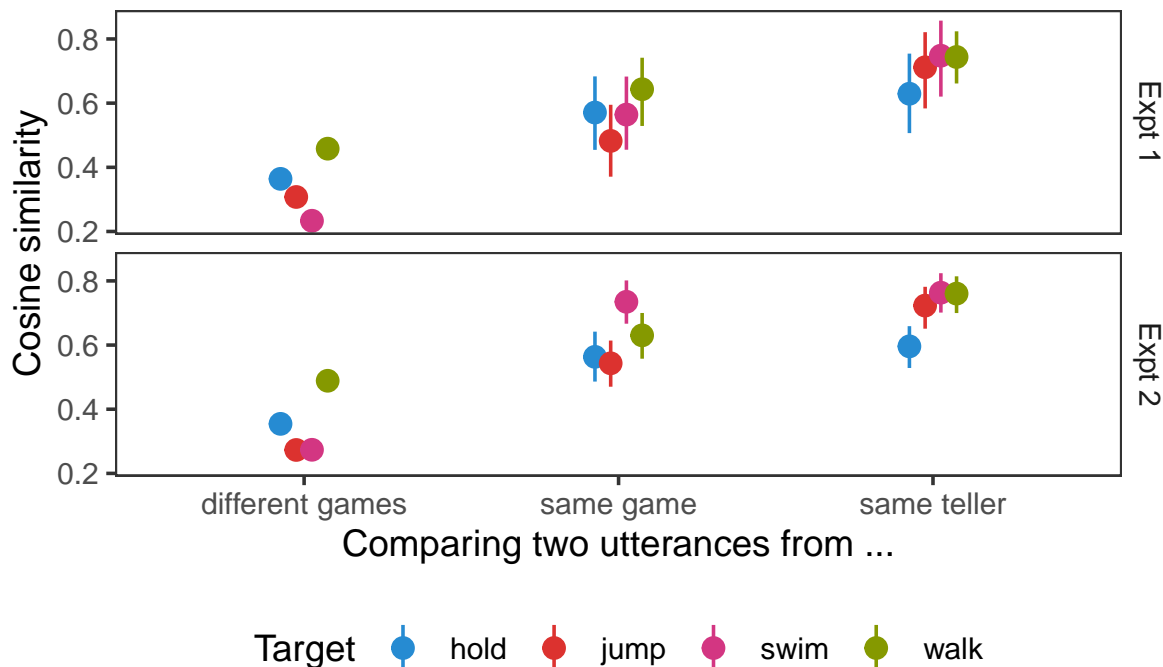


Figure 2.4: Semantic similarity between pairs of descriptions from different sources. Dots are means and lines are bootstrapped 95% CIs.

However, we can test for a more coarse measure of sensitivity to partner: whether children’s utterances are more like their partner’s than children in other games?. If children are fully ego-centric (as suggested by Glucksberg et al., 1966), their choices of descriptions would be independent from their partners.

Following the methods of Boyce et al. (2024), we embedded each description in a semantic vector space using S-BERT (Reimers & Gurevych, 2019), and then used the cosine between embeddings as a measure of semantic similarity.

We compared the semantic similarities between descriptions of the same target based on who produced the description (Figure 2.4). We used a Bayesian mixed effects linear regression to predict similarity.⁴ Utterances were more similar if they came from the same partnership (0.222 [0.184, 0.260]). Utterances were slightly more similar still if they came from the same person (0.135 [0.077, 0.193]), which is expected since children are likely to be fairly consistent with themselves. However, children used

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

descriptions that were much more similar to their partner’s than to other children’s (Figure 2.4), indicating sensitivity to their partner’s expressions.

2.2.3 Discussion

In Experiment 1, 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 scaffolding 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.

2.3 Experiment 2

2.3.1 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 [anonymized link](#).

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 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 10 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.

Table 2.1: Example descriptions children successfully used to identify different target 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 	

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 2.1D).

Procedure

The procedure was much the same as Experiment 1 (Figure 2.1D). We added an initial “bubble popping” exercise to give children practice tapping 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, 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.

2.3.2 Results

Accuracy and speed

In Experiment 2, children’s accuracy was above chance (Odds Ratio: 5.95 [3.07, 11.89]) and relatively stable over time (OR of one trial later: 1.01 [0.94, 1.09], Figure 2.2). 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. Taken together, we find more evidence that children can successfully communicate with each other about these abstract shapes, and do so with increasing efficiency.

Description length

The average length of descriptions on the first trial was 3.44 [2.23, 4.75] words and description length was relatively stable over time (change of 0.02 [-0.05, 0.09] words / trial, Figure 2.3). This finding is comparable to Experiment 1, again finding that children produce short utterances without much change in length over time.

Convergence

As a coarse measure of partner-sensitivity, we repeated the semantic 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 if they came from the same person (0.097 [0.059, 0.132]).

As Experiment 2 had 4 blocks, we examined whether descriptions were converging semantically toward the final description. 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 if they came from the same child, which is expected as a sign of internal consistency (0.067 [0.006, 0.127]). Although over time descriptions did get more similar to the last block utterance, the semantic distance between adjacent block utterances was relatively constant (0.009 [-0.026, 0.044]).

Partnerships often diverged from one another as groups focused on distinct aspects of the image. We tested whether descriptions in different games diverged over time using a Bayesian mixed effects linear regression.⁶ As the games progressed, descriptions to the same target from different games became slightly less similar (-0.013 [-0.018, -0.008]), which indicates that games are converging to different conventions.

⁵ $\text{sim} \sim \text{earlier.block.num} + \text{same.speaker} + (1|\text{game1}) + (1|\text{target})$

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

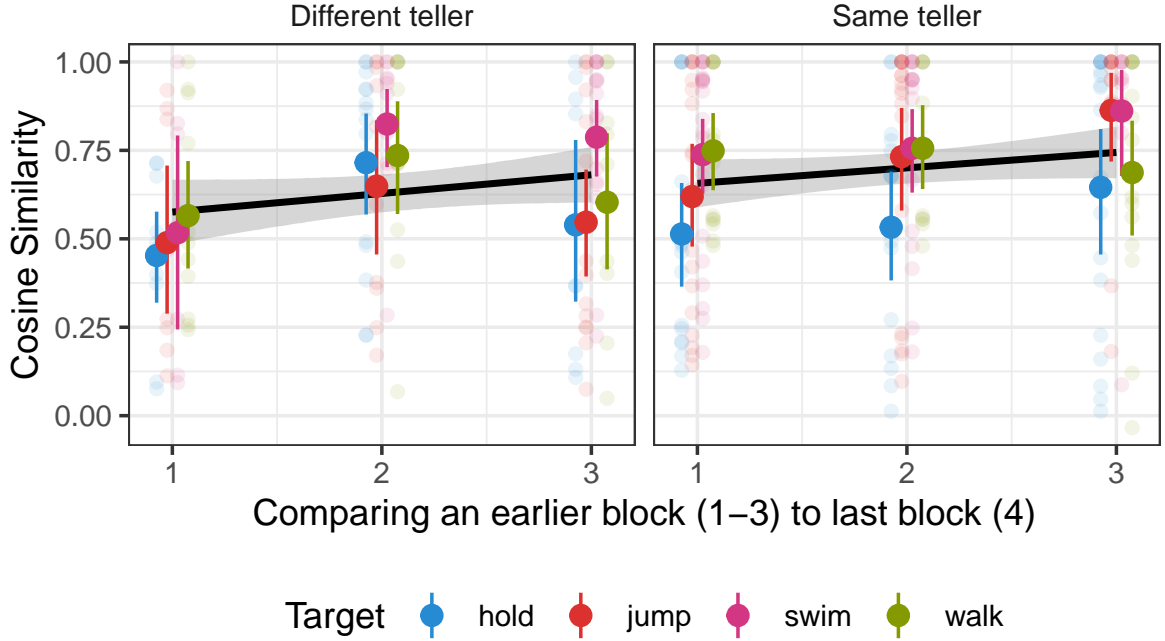


Figure 2.5: Semantic similarity between descriptions from earlier blocks (1-3) and the last block in Experiment 2. Heavy dots are means with bootstrapped 95% CIs; light dots are individual values.

These patterns of increasing similarity within games and increasing divergence between games qualitatively match the patterns found for adults (Boyce et al., 2024).

2.4 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 (OR: 4.64 [1.42, 12.16]) and accuracy numerically increased over the course of the game, although the credible interval included 0 (OR of one trial later: 1.04 [0.98, 1.12]). Descriptions produced by tellers averaged 3.62 [0.09, 7.38] words on the first trial, and description length was relatively stable over time (change of 0.04 [-0.02, 0.11] words per trial). Utterances were more similar if they came from the same partnership (increase in cosine similarity: 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]).

We might expect that whether a description was successful influences whether the same description, or a variant of it, is employed in future rounds. Intuitively, successful descriptions can be copied and built upon, while unsuccessful descriptions should be replaced by a fresh attempt. To test whether accuracy is predictive of similarity to future descriptions, we ran a post-hoc Bayesian linear model predicting similarity to the next block description in terms of accuracy.⁷ Descriptions that elicited a correct response were more similar to the next block description (0.145 [0.055, 0.236]) with no substantial interaction with block number or whether both descriptions came from the same teller. This pattern of results is consistent with the expectation that children are more likely to stick to their own description or repeat the other child’s description if it was previously successful.

2.5 General discussion

Prominent early studies claimed that young children cannot overcome their egocentrism to coordinate with each other in reference games (Glucksberg et al., 1966). However, more recent developmental work has found that young children show emerging communicative and pragmatic sensitivity, especially when cognitive demands are low. Here, we revisited the question of preschoolers’ performance in child-child reference games, using a scaffolded paradigm to reduce extraneous task demands.

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. 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.

Unlike adults, children did not display an increase in accuracy or a shortening of referential expressions over the course of the game. Still, our findings show that

⁷ $\text{sim} \sim \text{earlier_block.num} \times \text{correct} + \text{same_speaker} \times \text{correct} + (1|\text{game1}) + (1|\text{target}) + (1|\text{expt})$

descriptions became increasingly similar to descriptions in the last block and that successful utterances were more similar to future utterances, suggesting that children are adapting their descriptions as the game unfolds. These null findings are likely a result of initially high accuracy in the first block and initially short utterances that leave little room for reduction.

It is unclear to what extent the uniformly short descriptions we observed are a product of the simplified task or children’s behavioral differences from adults. In this case, the low number of options and relatively easy-to-describe shapes may have obviated the need for long initial descriptions. Indeed, adult controls in Leung et al. (2024) used shorter initial descriptions than adults in studies with larger arrays of harder to distinguish images (Boyce et al., 2024; Hawkins et al., 2020a). However, young children may also struggle to produce longer descriptions, and young children may be more willing to take guesses when adults would seek additional clarification. 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.

The generalizability of our results is limited by the target population, the target images, and the task structure. We sampled a convenience population of children at a university nursery school. The set of tangram images may be easier to distinguish and have higher codability than other target images used in adult reference games. We specifically targeted children’s abilities to construct referring expressions that can be jointly understood, so children were provided scaffolding around taking turns and talking to their partner. Thus, children’s performance should be taken as a proof-of-concept about ability, rather than a claim about how generally children spontaneously demonstrate these abilities.

In the broader picture of language acquisition, there is debate over the timing of the emergence of communicative and pragmatic abilities relative to the acquisition of grammar and meaning. On one side, children seem to learn literal semantics far before they display an understanding of some pragmatic implicatures (Huang & Snedeker, 2009; Noveck, 2001); on the other, 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). Our current

findings 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 already able to establish novel referential conventions with one another as part of their broader ability to communicate and coordinate.

Chapter 3

Processing stuff

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

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>
- Ahern, T. C. (1994). The effect of interface on the structure of interaction in computer-mediated small-group discussion. *Journal of Educational Computing Research*, 11(3), 235–250.
- 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>
- Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, 53(5), 2158–2171.
- Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences. *Behavioral and Brain Sciences*, 1–55. <https://doi.org/10.1017/S0140525X22002874>
- 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*. <https://doi.org/10.31234/osf.io/v8e56>
- 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. (2006). Perspectives on multi-party dialogue. *Research on Language and Computation*, 4(2), 153–177.
- 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>
- Brennan, S. E., & Clark, H. H. (1996). *Conceptual Pacts and Lexical Choice in Conversation*. 12.
- Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the r package brms. *The R Journal*, 10(1), 395–411.
- Caplow, T. (1957). Organizational size. *Administrative Science Quarterly*, 484–505.
- Carletta, J., Garrod, S., & Fraser-Krauss, H. (1998). Placement of Authority and Communication Patterns in Workplace Groups: The Consequences for Innovation. *Small Group Research*, 29(5), 531–559. <https://doi.org/10.1177/1046496498295001>
- Carruthers, P. (2013). Mindreading in Infancy. *Mind & Language*, 28(2), 141–172. <https://doi.org/10.1111/mila.12014>
- Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning*. ERIC.
- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81.
- Clark, H. H., & Wilkes-Gibbs, D. (1986). *Referring as a collaborative process*.
- Cohn-Gordon, Reuben, Levy, R., & Bergen, L. (2019). *The pragmatics of multiparty communication*.
- Dewhirst, H. D. (1971). Influence of perceived information-sharing norms on communication channel utilization. *Academy of Management Journal*, 14(3), 305–315.
- Fay, N., Garrod, S., & Carletta, J. (2000). Group Discussion as Interactive Dialogue or as Serial Monologue: The Influence of Group Size. *Psychol Sci*, 11(6), 481–486. <https://doi.org/10.1111/1467-9280.00292>
- Fox Tree, J. E., & Clark, N. B. (2013). Communicative Effectiveness of Written

- Versus Spoken Feedback. *Discourse Processes*, 50(5), 339–359. <https://doi.org/10.1080/0163853X.2013.797241>
- Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31(6), 961–987.
- Ginzburg, J., & Fernandez, R. (2005). Action at a distance: The difference between dialogue and multilogue. *Proceedings of DIALOR*, 9.
- 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>
- Guilbeault, D., Baronchelli, A., & Centola, D. (2021). Experimental evidence for scale-induced category convergence across populations. *Nature Communications*, 12(1, 1), 327. <https://doi.org/10.1038/s41467-020-20037-y>
- Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., & Fernández, R. (2019). The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, 1895–1910. <https://doi.org/10.18653/v1/P19-1184>
- Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020a). Characterizing the dynamics of learning in repeated reference games. *arXiv:1912.07199 [Cs]*. <https://arxiv.org/abs/1912.07199>
- Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020b). Characterizing the dynamics of learning in repeated reference games. *ArXiv191207199 Cs*. <http://arxiv.org/abs/1912.07199>

- Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2023). From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*, 130(4), 977.
- Hiltz, S. R., Johnson, K., & Turoff, M. (1986). Experiments in group decision making: Communication process and outcome in face-to-face versus computerized conferences. *Human Communication Research*, 13(2), 225–252.
- Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, 18.
- Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, 96(2), 127–142. <https://doi.org/10.1016/j.cognition.2004.07.001>
- 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>
- Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 582–601.
- 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., & Bricker, P. D. (1967). Effects of transmission delay and access delay on the efficiency of verbal communication. *The Journal of the Acoustical Society of America*, 41(2), 286–292.
- Krauss, R. M., Garlock, C. M., Bricker, P. D., & McMahon, L. E. (1977). The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology*, 35(7), 523.

- 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>
- Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3), 343–346. <https://doi.org/10.1037/h0023705>
- Kraut, R. E., Lewis, S. H., & Swezey, L. W. (1982). Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, 43(4), 718.
- 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>
- Lewis, D. (1969). *Convention: A philosophical study*. John Wiley & Sons.
- MacMillan, J., Entin, E. E., & Serfaty, D. (2004). Communication overhead: The hidden cost of team cognition. In *Team cognition: Understanding the factors that drive process and performance*. (pp. 61–82). American Psychological Association. <https://doi.org/10.1037/10690-004>
- 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>

- Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2), 201–213. [https://doi.org/10.1016/S0749-596X\(03\)00028-7](https://doi.org/10.1016/S0749-596X(03)00028-7)
- 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)
- Parisi, J. A., & Brungart, D. S. (2005). Evaluating communication effectiveness in team collaboration. *Ninth European Conference on Speech Communication and Technology (INTERSPEECH)*.
- 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>
- Rogers, S. L., Fay, N., & Maybery, M. (2013). Audience Design through Social Interaction during Group Discussion. *PLOS ONE*, 8(2), e57211. <https://doi.org/10.1371/journal.pone.0057211>
- San Juan, V., Khu, M., & Graham, S. A. (2015). A New Perspective on Children’s Communicative Perspective Taking: When and How Do Children Use Perspective Inferences to Inform Their Comprehension of Spoken Language? *Child Development Perspectives*, 9(4), 245–249. <https://doi.org/10.1111/cdep.12141>

- Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2), 211–232. [https://doi.org/10.1016/0010-0285\(89\)90008-X](https://doi.org/10.1016/0010-0285(89)90008-X)
- Seaman, C. B., & Basili, V. R. (1997). Communication and organization in software development: An empirical study. *IBM Systems Journal*, 36(4), 550–563.
- Swaab, R. I., Galinsky, A. D., Medvec, V., & Diermeier, D. A. (2012). The communication orientation model: Explaining the diverse effects of sight, sound, and synchronicity on negotiation and group decision-making outcomes. *Personality and Social Psychology Review*, 16(1), 25–53.
- Tannen, D. (2005). *Conversational style: Analyzing talk among friends*. Oxford University Press.
- Tolins, J., & Fox Tree, J. E. (2016). Overhearers Use Addressee Backchannels in Dialog Comprehension. *Cogn Sci*, 40(6), 1412–1434. <https://doi.org/10.1111/cogs.12278>
- Tomasello, M. (2008). *Origins of human communication*. MIT press.
- Traum, D. (2004). Issues in Multiparty Dialogues. In F. Dignum (Ed.), *Advances in Agent Communication* (Vol. 2922, pp. 201–211). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24608-4_12
- 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>
- Weber, R. A., & Camerer, C. F. (2003). Cultural Conflict and Merger Failure: An Experimental Approach. *Manag. Sci.*, 49(4), 16.
- Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2), 183–194.
- Wittgenstein, L. (1953). *Philosophical investigations*. Wiley-Blackwell.
- Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(4), 919–937. <https://doi.org/10.1037/a0036161>
- Yoon, S. O., & Brown-Schmidt, S. (2018). Aim Low: Mechanisms of Audience Design in Multiparty Conversation. *Discourse Processes*, 55(7), 566–592. <https://doi.org/10.1080/01650308.2018.1511111>

org/10.1080/0163853X.2017.1286225

Yoon, S. O., & Brown-Schmidt, S. (2019). Audience Design in Multiparty Conversation. *Cogn. Sci.*, 43(8), e12774. <https://doi.org/10.1111/cogs.12774>

Zack, M. H. (1993). Interactivity and communication mode choice in ongoing management groups. *Information Systems Research*, 4(3), 207–239.