

1 Interaction structure constrains  
2 the emergence of conventions in group communication

3 Veronica Boyce<sup>1,\*</sup>, Robert D. Hawkins<sup>2</sup>, Noah D. Goodman<sup>1,3</sup>, Michael C. Frank<sup>1</sup>

4 <sup>1</sup>Psychology Department, Stanford University, Stanford, CA 94305

5 <sup>2</sup>Psychology Department, University of Wisconsin – Madison, Madison, WI, 53715

6 <sup>3</sup>Computer Science Department, Stanford University, Stanford, CA 94305

7  
8 Corresponding author: Veronica Boyce Email: [vboyce@stanford.edu](mailto:vboyce@stanford.edu) Address: Building 420. 450  
9 Jane Stanford Way, Stanford, CA, 94305. Phone: 650-933-0841

10 Pre-print status: this work is on the pre-print server PsyArxiv at [https://osf.io/preprints/psyarxiv/](https://osf.io/preprints/psyarxiv/a3wfy)  
11 [a3wfy](https://osf.io/preprints/psyarxiv/a3wfy) under a CC-BY license.

12 Classification: Social Sciences -> Psychological and Cognitive Sciences

13 Keywords: Communication, Psycholinguistics, Reference games, Convention formation, large-scale  
14 online experiment

---

\*Corresponding author. Email: [vboyce@stanford.edu](mailto:vboyce@stanford.edu).

## Abstract

Real-world communication frequently requires language producers to address more than one comprehender at once, yet most psycholinguistic research focuses on one-on-one communication. As the audience size grows, interlocutors face new challenges that do not arise in dyads. They must consider multiple perspectives and weigh multiple sources of feedback to build shared understanding. Here, we ask which properties of the group’s *interaction structure* facilitate successful communication. We used a repeated reference game paradigm in which directors instructed between one and five matchers to choose specific targets out of a set of abstract figures. Across 313 games ( $N = 1,319$  participants), we manipulated several key constraints on the group’s interaction, including the amount of feedback that matchers could give to directors and the availability of peer interaction between matchers. Across groups of different sizes and interaction constraints, describers produced increasingly efficient utterances and matchers made increasingly accurate selections. Critically, however, we found that smaller groups and groups with less-constrained interaction structures (“thick channels”) showed stronger convergence to group-specific conventions than large groups with constrained interaction structures (“thin channels”), which struggled with convention formation. Overall, these results shed new light on the core structural factors that enable communication to thrive in larger groups.

**Significance Statement:** Human linguistic communication is remarkably versatile, from one-on-one conversations to large group chats on platforms like Slack. Yet existing empirical work has overwhelmingly focused on one-on-one contexts, overlooking the factors that affect real-world communication in larger groups. We address this gap using a large-scale reference game task, examining how groups of varying sizes and interaction structures coordinate on linguistic conventions for novel objects. We found effective and robust communication across multiple settings, but larger (6-person) groups were more sensitive to differences in group structure and communication channels. Our results suggest that “thicker” communication channels - where group members provide more feedback to one another - can result in better communication in larger groups.

## 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 (Traum 2004, Ginzburg & Fernandez 2005, Branigan 2006). 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 (Horton & Gerrig 2002, Weber & Camerer 2003, Horton & Gerrig 2005, Fox Tree & Clark 2013, Yoon & Brown-Schmidt 2014, 2018), while comprehenders must account for the fact that utterances are not necessarily tailored to them (Carletta et al. 1998, Fay et al. 2000, Metzing & Brennan 2003, Rogers et al. 2013, Tolins & Fox Tree 2016, Cohn-Gordon et al. 2019, 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 & Weinheimer 1966, Krauss & Bricker 1967, 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 (Seaman & Basili 1997, MacMillan et al. 2004), 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 (Wittgenstein 1953, Lewis 1969) 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 (Krauss & Weinheimer 1964, Clark & Wilkes-Gibbs 1986). 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 (Haber et al. 2019, Almaatouq et al. 2021, 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. 2020). 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.

## 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 2A). On each trial, everyone in the group was shown a gallery of 12 tangram images (Clark & Wilkes-Gibbs 1986, Hawkins et al. 2020, 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

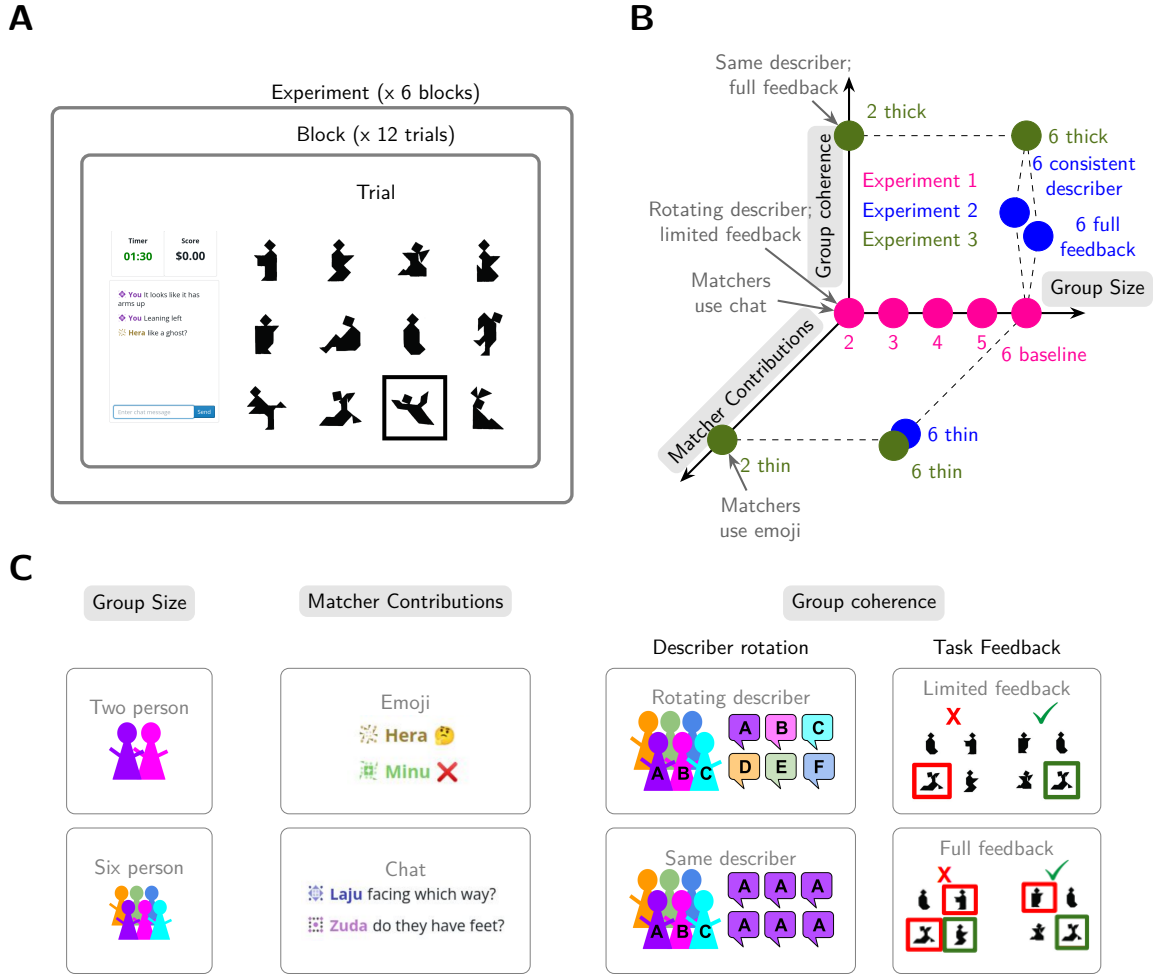


Figure 1: (A) Participants played a repeated reference game in groups of size 2 to 6. On each trial, a describer described the target image to the group of matchers. Each image appeared once per block for six blocks. (B) Experiments varied along 3 dimensions: Group size, group coherence, and matcher 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).

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 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 2C). Other factors, such as the set of stimuli and background knowledge about one’s partners, were held constant across games.

## 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 listener 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).

## 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 confidence intervals. See SI Figure

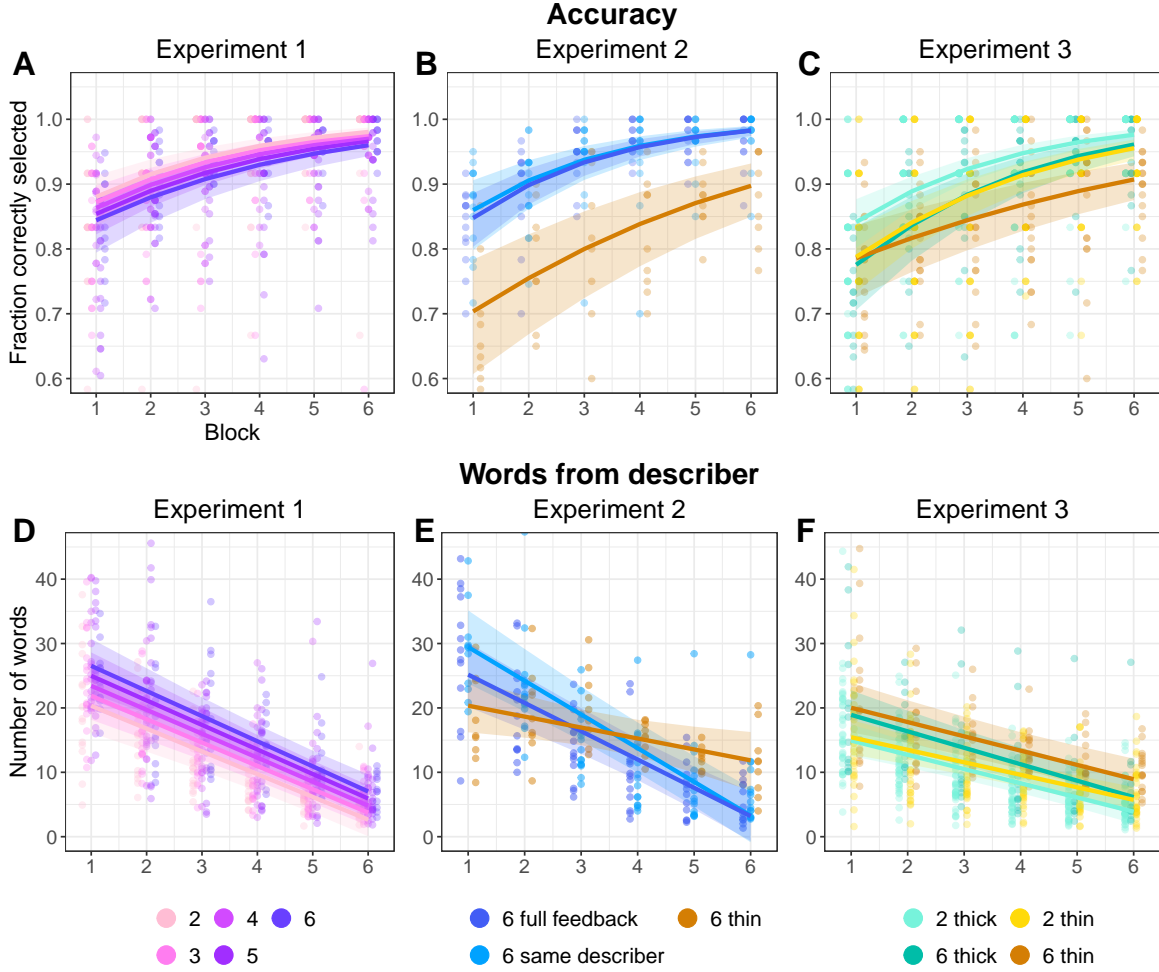


Figure 2: Behavioral results across all three experiments. (A-C). Matcher accuracy at selecting the target image. (D-F). Number of words produced by the descriptor each trial. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a few outliers points are not visible.

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 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 are 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.

### 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 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.

### 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 1. 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

Table 1: Examples from 6-player groups in Experiment 3 of successful descriptions for the same image across repetitions. Describers are indicated with an asterisk. More example descriptions are in SI Tables 1 and 2.

6-person thick game	
<i>Rep 1: 4/5 correct</i>	
A*	sitting down no legs showing
C	no arms?
B	Bunny ears?
A*	legs showing to one side no arms
C	kinda like kneeling?
A*	no feet
A*	yes kneeling
<i>Rep 2: 5/5 correct</i>	
A*	sitting down with no feet showing. legs to one side
A*	no arms
E	Swaddled up like a bb
C	cute bb
D	I think that's the one that looks like a ghost to me, like he's wearing a sheet. (?)
<i>Rep 3: 5/5 correct</i>	
A*	sitting down no arms showing legs to one side
E	the bb
B	Swaddled baby?
F	Baby?
A*	the bb
<i>Rep 4: 5/5 correct</i>	
A*	sitting down with no arms legs to one side
<i>Rep 5: 5/5 correct</i>	
A*	sitting no arms and legs to one side. I will call them Kevin
D	<3 Kevin
C	bb kevin
<i>Rep 6: 5/5 correct</i>	
A*	kevin the baby
E	yess kevin



6-person thin game	
<i>Rep 1: 5/5 correct</i>	
L*	man with no arms or legs
O	😏
P	😏
L*	slight protuberance on the right
<i>Rep 2: 5/5 correct</i>	
N*	Can't see any arms. Imagine wrapped in a blanket completely.
N*	Armless and legless
N*	Burrito with a head
M	😄
O	😄
P	😏
<i>Rep 3: 5/5 correct</i>	
O*	burrito
<i>Rep 4: 5/5 correct</i>	
Q*	burrito
<i>Rep 5: 5/5 correct</i>	
M*	burrito
<i>Rep 6: 5/5 correct</i>	
P*	And our last one! Anyone else hungry after this? I quite fancy a burrito!
L	😄
Q	✅
N	😄😄😄
Q	😄✅
M	✅
Q	😄
N	✅✅✅



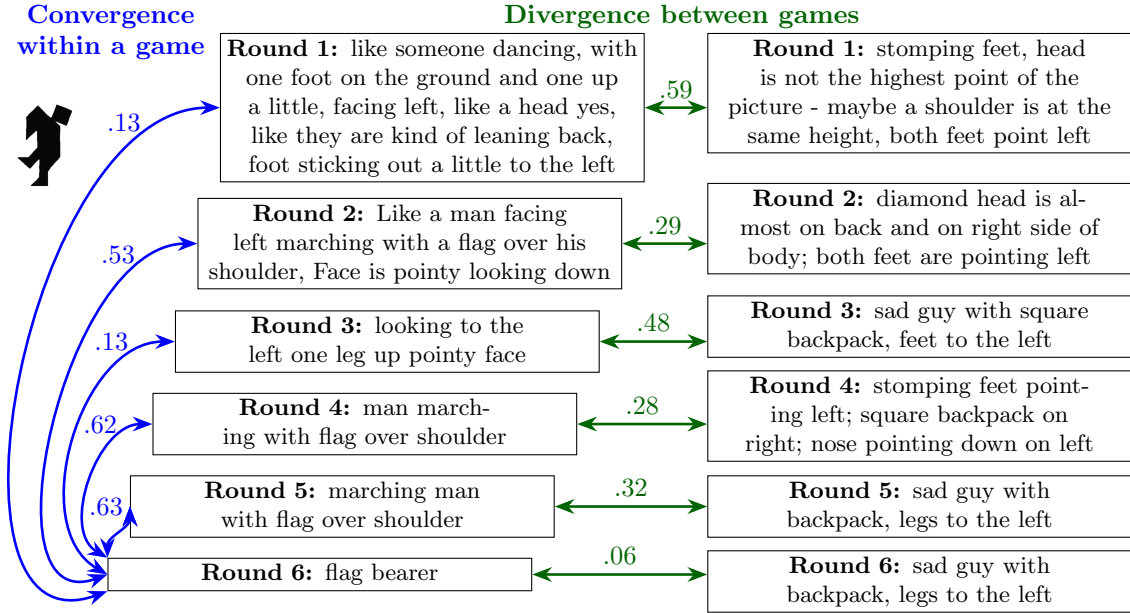


Figure 3: Example utterances describing the shown tangram figure produced by two 3-player games in Experiment 1. To measure convergence within a game (blue), we measured the cosine similarity between SBERT embeddings of descriptions and the embedding of the round 6 utterance (taken to be the convention).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

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.

### 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 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 4A-C; SI Tables 25-29). All conditions showed some convergence toward a conventional “shorthand” 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 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 4C,  $\beta = 0.009$ , 95% CrI =  $[0.001, 0.017]$ , this effect is not robust to sensitivity analysis, SI Figure 5C and SI Table 68). Second, thicker games tended to converge faster than thin games (Figure 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.

### 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

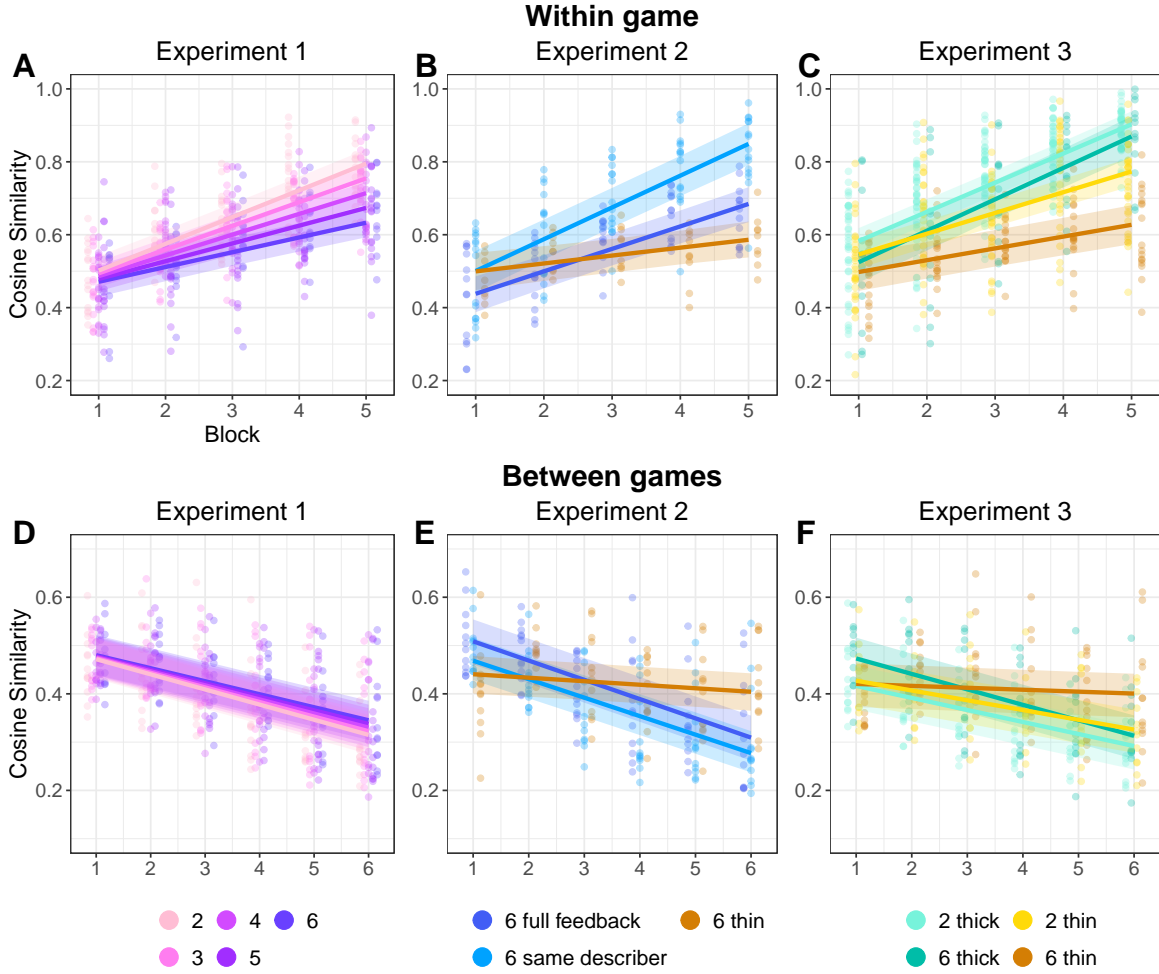


Figure 4: Language similarity results measured with pairwise cosine similarity between embeddings of two utterances. (A-C). Convergence of descriptions within games as measured by similarity between an utterance from block 1-5 to the block 6 utterance in the same game for the same image. (D-F). Divergence 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.

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 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 4F,  $\beta = 0.017$ , 95% CrI = [0.015, 0.019]), and in Experiment 2, 6-person thin games barely diverged at all (Figure 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 thin condition, however, retains a higher level of 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 low level of utterances that do not contain any of these classes of words.

## Discussion

Communication often occurs in multi-party settings, but research on referential communication typically does not focus on such settings – largely due to practical obstacles. Dyadic reference games have been used to measure informational efficiency, characterized by describer-matcher pairs creating conventional (stable but somewhat arbitrary) labels which are not shared by other groups. In the current work, we asked how this process of reference formation unfolds in larger groups and under varying interaction structures. Across 3 online experiments and 11 experimental conditions, we varied game features including group size, modality of matcher contributions, and degree of group coherence. All conditions replicated classic phenomena: increasing accuracy, reduction in describer utterances, semantic convergence within games, and differentiation of descriptions between groups. However, we also found that the interaction structure of a group substantially affects how rapidly groups develop partner-specific conventions. Small groups may be able to 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 has often been taken as an index of group-specific convention formation (Clark & Wilkes-Gibbs 1986, Brennan & Clark 1996, 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 converging on a unified group-specific label. Instead, they may be converging to a multi-label solution based on group 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.

Even within the general framework of iterated reference, there is a high dimensional feature space of possible experiments. We sampled only a few points along a few dimensions in the space that felt salient. In our experiment 3, we grouped some factors together in order to have more games in each condition: a fully factorial design would have been too expensive to power adequately. We instantiated a “thin” channel by limited matchers to 4 discrete utterances (emojis), but there are other ways to manipulate channel width for describers and matchers, such as rate limiting typing or adding time pressure. Future work could sample other points in the experimental space, including exploring other manipulations on channel thickness, the effects of different target images, or groups of people with real-life prior connections.

351 People communicate using language in a variety of modalities that vary whether the  
352 participants use oral or written language, whether they are co-present in the same space, and whether  
353 they have visual access to each other. All of these situations have a shared communicative and  
354 linguistic base, but different situations have different affordances and different norms for how to  
355 communicate. In this work, we used a web-based chat modality where communication was text  
356 based and there was not co-presence or visual access. Given previous work has found overall similar  
357 trends across different modalities (Hawkins et al. 2020), we suspect that the general pattern of  
358 effects we see here in terms of group size and group coherence are likely to extend to other modalities.  
359 However, different modalities may allow for different approaches that may be more or less sensitive  
360 to group size, describer rotation, or different levels of matcher contributions. For instance, in  
361 face-to-face oral settings, it may be easier for describers to continuously talk until interrupted or  
362 track the comprehension of individual group members, which could lead to overall higher amounts of  
363 language and higher performance. People communicate in a variety of ways in everyday life, and  
364 studying different of these situations is useful both for understanding the specific modalities and for  
365 understanding their shared communicative underpinnings.

366 We cannot make causal claims about specific mechanisms by which manipulations such as group size  
367 resulted in different outcomes: there are many differences between being in a 2-person group versus  
368 a 6-person group that could lead to the different outcomes. In a dyad, producers can tailor their  
369 utterances to the one matcher, but in large groups, producers must balance the competing needs  
370 of different comprehenders (Schober & Clark 1989, Tolins & Fox Tree 2016). These effects likely  
371 vary by both the knowledge state of and communication channels available to the comprehenders  
372 (Horton & Gerrig 2002, Horton & Gerrig 2005, Fox Tree & Clark 2013). Further work digging into  
373 the language used and the interactions between participants might unearth plausible mechanisms for  
374 how differences in group size and interaction structure influence outcomes, and this in turn could  
375 then point towards future experimental conditions.

376 Communication occurs across a broad range of situations, varying on many dimensions, including  
377 group size, medium of interaction, and group structure. A narrow focus on dyads with rich  
378 communication channels can lead to theories that mispredict how interactions play out in multi-party  
379 groups with varying interaction structure. Sampling from a broader range of communicative situations  
380 is thus a critical part of better understanding human communication.

## 381 Materials and Methods

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

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

## 390 Participants

391 This research was covered by the Stanford IRB under protocol 20009 “Online investigations of language  
392 learning”. Participants were recruited using the Prolific platform. All participants self-reported as

---

<sup>1</sup>Experiment 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>

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 or 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.

## Materials

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

## 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. (2020), 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.

## 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.

## 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 normal(0,1) priors for betas and sd. Models of describer efficiency were run as linear models with an intercept prior of normal(12,20), a beta prior of normal(0,10), an sd prior of normal(0,5) and a random-effect correlation prior of lkj(1). For all of the models of SBERT similarity, we used linear models with the priors normal(.5,.2) for the intercept, normal(0,.1) for betas, and 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.

## Author’s Note

Experiment 1 was previously reported in Proceedings of the Annual Meeting of the Cognitive Science Society 44 (2022).

We thank the LangCog Lab, Saxelab, CAMP 5, HSP 2023, and Cogsci 2022 audiences for helpful



530 feedback on this work.

531 This work was supported by a Hoffman-Yee Grant from the Stanford Institute for Human-Centered  
532 AI.

533 We report CRediT taxonomy contributions as follows: All authors did conceptualization and  
534 methodology. VB did data curation, formal analysis, investigation, visualization, and writing –  
535 original draft. VB and RDH did software. RDH, NDG, and MCF did writing – editing. MCF and  
536 NDG did funding. MCF did supervision.

## 537 **References**