Interaction structure constrains the emergence of conventions in group communication

Veronica Boyce[1,*], Robert Hawkins[2], Noah D. Goodman[1], Michael C. Frank[1]

[1]Stanford University
[2]University of Wisconsin – Madison

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, interlocuters 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* facilitates 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 interaction between matchers. Across groups of different sizes and interaction constraints, describers used increasingly short utterances to convey their meaning to matchers who selected the targets with increasing accuracy. Smaller groups and groups with less-constrained interaction structures ("thick channels") showed stronger convergence to group-specific conventions, while large groups with limiting interaction structures ("thin channels") struggled with convention formation. Overall, these results shed new light on the core structural factors that enable communication to thrive 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?

[*]Corresponding author. Email: vboyce@stanford.edu

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 group 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), several factors that may slow such deterioration have been identified in qualitative work, each of which relates to the structural "thickness" of the feedback channel (Ahern 1994, Parisi & Brungart 2005).

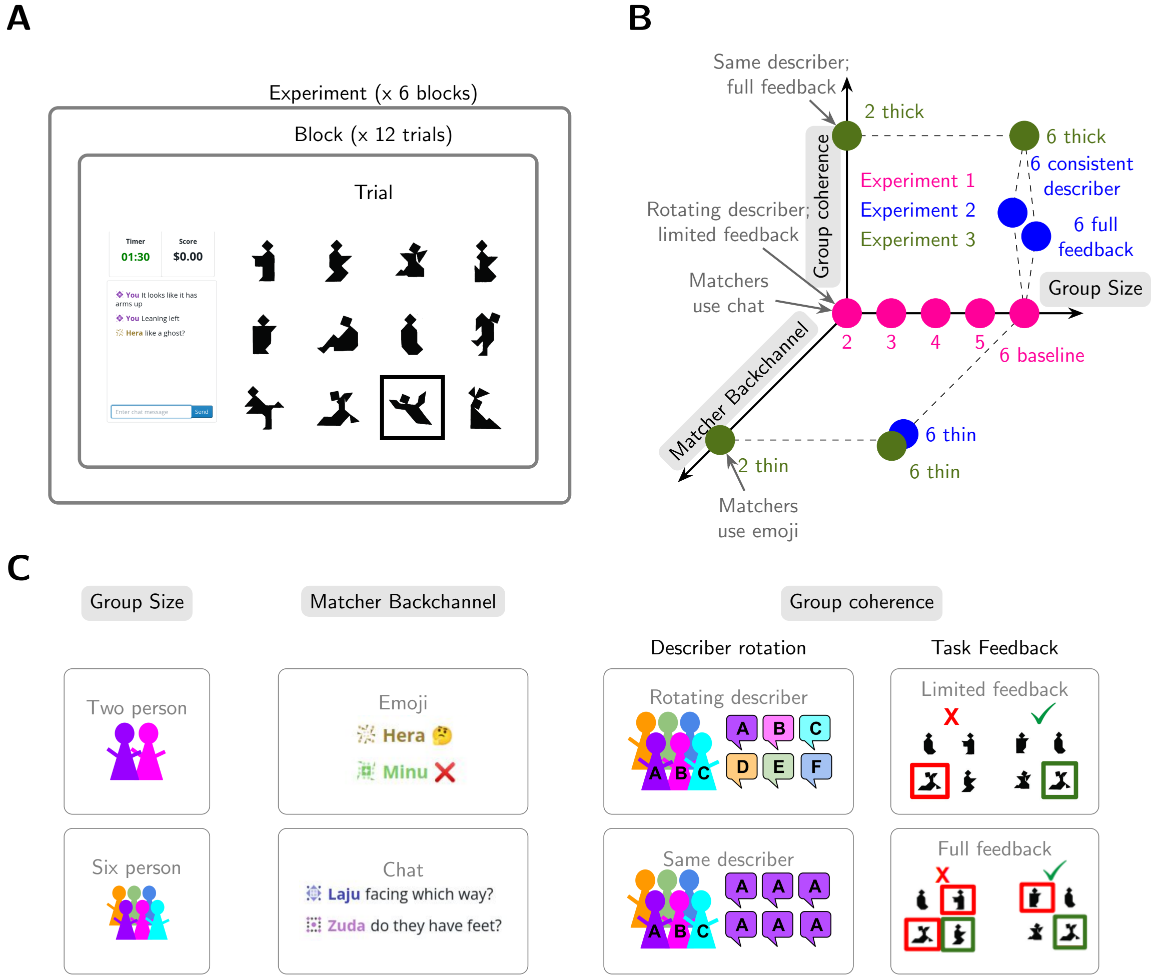
In this paper, we develop an experimental paradigm for evaluating the relative contribution of these factors: a *multi-party repeated reference game.* The ability to distinguish one particular entity from other possible entities, known as *reference*, is one of the most primitive and ubiquitous functions 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. Later utterances require fewer words, but also become more impenetrable to outsiders (Schober & Clark 1989, Wilkes-Gibbs & Clark 1992).

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 an interactive web-based platform (Haber et al. 2019, Hawkins et al. 2023). Repeated reference games in online, chat-based paradigms have closely replicated earlier results from face-to-face studies (Hawkins et al. 2020), and 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 use the multi-party repeated reference game to explore effects of group size and interaction channel thickness through a series of three experiments. **Taken together, our findings illuminate the mechanisms of social interaction in larger groups and suggest that larger groups may be a more sensitive environment for studying communication. TODO WHAT'S OUR CALIBRATED CLAIM OF IMPORTANCE??**

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



You It looks like it has arms up

You Leaning left

Hera like a ghost?

Hera 🤔

Minu ✗

Laju facing which way?

Zuda do they have feet?

Figure 1: (A) Participants played a repeated reference game in groups of size 2 to 6. On each trial, describers described a target image to the matchers. Each image appeared once per block for six blocks. (B) Diagram of the experimental space. Experiments varied along 3 dimensions: Group size, group coherence, and matcher backchannel. (C) Schematic of the axes we manipulated. Experiment 1 (pink) varied group size from 2 to 6 players while holding group coherence and backchannel 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 backchannel).

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 backchannel* (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.

Experiment 1 investigated 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 explored the role of interaction structure purely within the most challenging 6-player groups. We manipulated two factors that we expected to increase group coherence and improve performance. First, 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 which which targets. Second, we gave the group of matchers full feedback about what every other member of the group had selected, and we showed the intended target. We also manipulated a factor that we expected to interfere with the ability to establish mutual understanding and thus impede performance. In the limited backchannel condition, matchers were limited to 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 the backchannel 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, Hawkins et al. 2020).

Group Performance

We first examined how accurately and efficiently groups were able to match images, before moving on to examine how describer's descriptions vary over time and condition.

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.

Smaller and higher coherence groups are more accurate.

We begin by examining the impact of interaction structure on referential success, the ability to correctly transmit the intended target to all matchers.

To test these effects, 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).

Across all conditions, we observed strong positive effects of repetition block, indicating improved performance over time (Figure 2A-C, SI Tables 2-6).

Group size and differences in interaction structure contributed to variation in group performance.

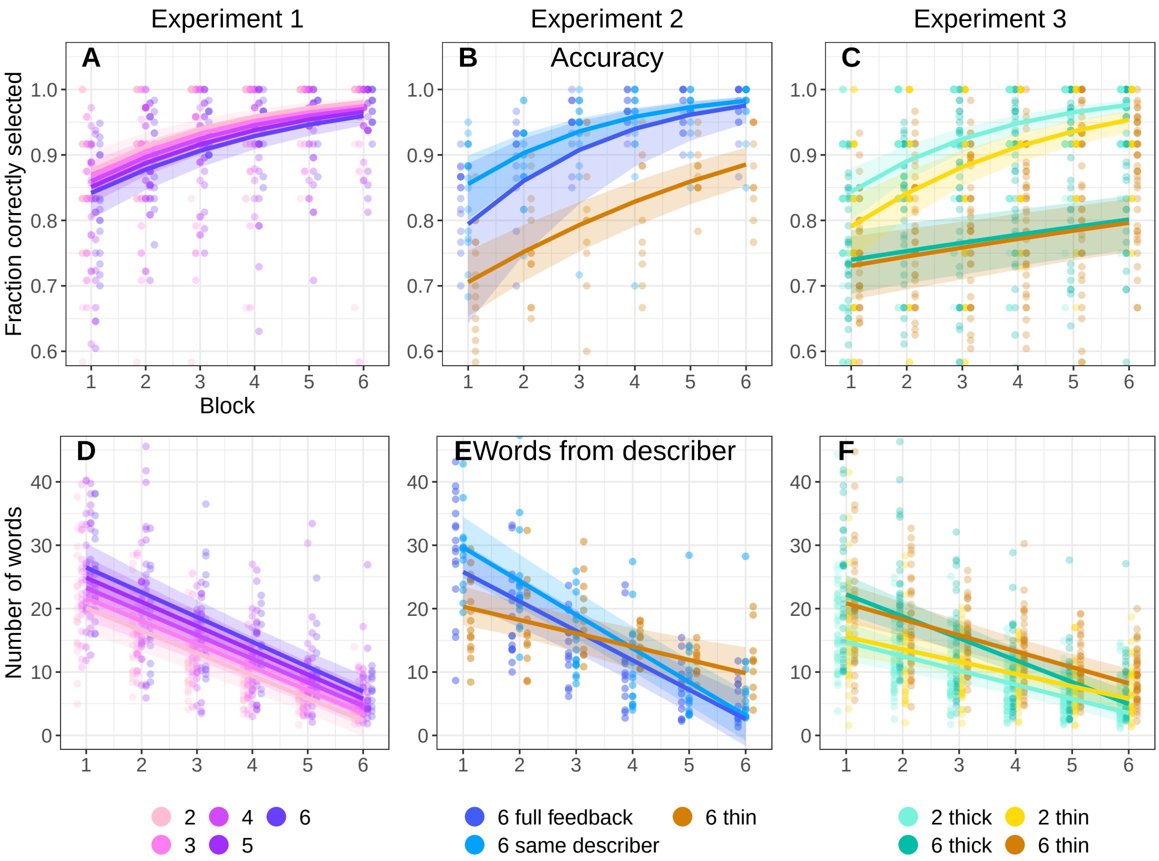


Figure 2: Behavioral results across all three experiments. (A-C). Matcher accuracy at selecting the target image. (D-F). Number of words produced by the describer each trial. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible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

Larger games had lower initial accuracy ($\beta = -0.36$, 95% CrI $= [-0.78, 0.05]$) and improved more slowly ($\beta = -0.36$, 95% CrI $= [-0.78, 0.05]$) than smaller games in Experiment 3, although group size differences were not reliable in Experiment 1 (SI Table 2). Among large groups in experiment 2, accuracy was higher in the thicker conditions than in the condition with thin interaction structure (SI Tables 3-5); however, in experiment 3, no effects of game thickness were reliable (SI Table 6).

We aggregated data from all experiments into a post-hoc mega-analytic model predicting accuracy as a function of game thickness (thin v. not-thin) and game size. Overall, accuracy increases over time ($\beta = 0.46$, 95% CrI $= [0.4, 0.52]$) but the rate of increase is 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, thicker games.

Regardless of group size and interaction structure, groups were far above chance and improved over the course of the game. Smaller groups and groups with higher coherence tended to be more accurate, though the magnitude and reliability of these effects varied across experiments.

Smaller and higher coherence groups use shorter descriptions.

After establishing that groups were able to communicate the targets successfully, 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 (for similar models predicting the length of describer's utterances before any matcher contributions, see SI Tables 15-18).

First, as predicted, describers in larger groups used longer descriptions at the outset than describers in smaller groups (Figure 2D-F). This effect held for the continuous measure of group size for Experiment 1 ($\beta = 1.6$, 95% CrI = $[0.62, 2.6]$) and the 2-person versus 6-person groups in Experiment 3 ($\beta = -2.21$, 95% CrI = $[-7.16, 3.08]$).

Samller groups tend to continue to use shorter descriptions than larger groups over the course of the game. In experiment 1, the rate of reduction was similar across different size groups ($\beta = -3.36$, 95% CrI = $[-4.56, -2.18]$). In Experiment 3, larger groups reduced faster than smaller ones ($\beta = 0.64$, 95% CrI = $[-0.59, 1.81]$), but the faster reduction did not fully make up for the longer initial starting point.

Interaction structure did not show a consistent effect on utterance length. While thin 6 player games showed a flatter reduction trajectory than thicker 6 player games in Experiment 2 (SI Tables 8-10), there was not a reliable effect of game thickness on reduction in Experiment 3 for either smaller or larger groups (SI Table 11).

Aggregating across experiments with a mega-analytic model revealed an overall pattern of reduction ($\beta = -2.8$, 95% CrI = $[-3.24, -2.36]$). 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, describers on average decreased their descriptions by a few words each repetition block. 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 use more backchannels.

As a final measure of group performance, we examine the back-and-forth interactions between the describer and the group of matchers. The backchannel allows matchers to actively provide feedback and seek clarification about the describer's referring expressions. An example transcript from a game where matchers contributed in various ways is in Table 1. Overall, larger groups displayed a higher proportion of trials where at least one matcher produced utterances (Supplement Figure 2A, $\beta = 0.79$, 95% CrI = $[0.58, 0.98]$), which declined across blocks ($\beta = -2.67$, 95% CrI = $[-3.54, -1.79]$). Over time, the length of matcher interjections decreases (Supplement Figure 2B, $\beta = 4.72$, 95% CrI = $[0.09, 9.44]$). This pattern is consistent with early matcher involvement in establishing a common conceptualization by asking questions and offering alternative descriptions. We found that emoji use in Experiment 3 followed similar trends (Supplement Figure 3). The amount of text produced by matchers is much less than that produced by describers, but matchers contributions also reduce over time, in both frequency and length.

Table 1: Example transcript of a 6-player group for Experiment 1 describing the same image each repetition. Matchers sometimes asked questions or offered clarifications, including in reference to prior descriptions.

Block 1	
A (describer)	classic ghost or man with his hands up
A (describer)	slanted
A (describer)	slanted to the left
D	slanted left? Right arm hanging out?
A (describer)	both arms up
A (describer)	and out
F	no legs
A (describer)	no legs ⌢
Block 2	
C (describer)	whole body tilted to the left, as if falling over
C (describer)	arms in air
D	Like a 1 legged exercise lunge though? Holding a medicine ball?
F	left side like an open wrench
C (describer)	like a ghost falling over
F	no legs
D	Okayyyy Ghost. Got it
Block 3	
D (describer)	The classic ghost. Arms are up in the air
D (describer)	Whole body slanted left
F	both arms up not legs
D (describer)	The right arm is further from the head than left
D (describer)	No legs visible
Block 4	
B (describer)	whole body slanting left with arms in the air
D	Classic ghost?
Block 5	
E (describer)	slanted shadow or ghost, arms up, no feet
Block 6	
F (describer)	classic ghost, both arms up, no legs, slanting back to the left

Interim summary

As an initial check on group performance, we examined three metrics of communicative performance in groups of different sizes and interaction structures. Groups in all conditions were able to communicate about the target images with a high and increasing degree of accuracy even as describers and matchers both decreased the length of their descriptions over time. Smaller groups and groups with thicker interaction structures tended to perform better, although some of the comparisons were not reliably different from one another.

Linguistic Content

In the previous sections, we confirmed that groups across conditions followed the classic patterns of increasing accuracy and decreasing description length. Here, we aim to better understand the mechanisms that allow describers to use shorter descriptions while matcher accuracy increases. In particular, we explore the hypothesis that interaction structure and group size affect performance by

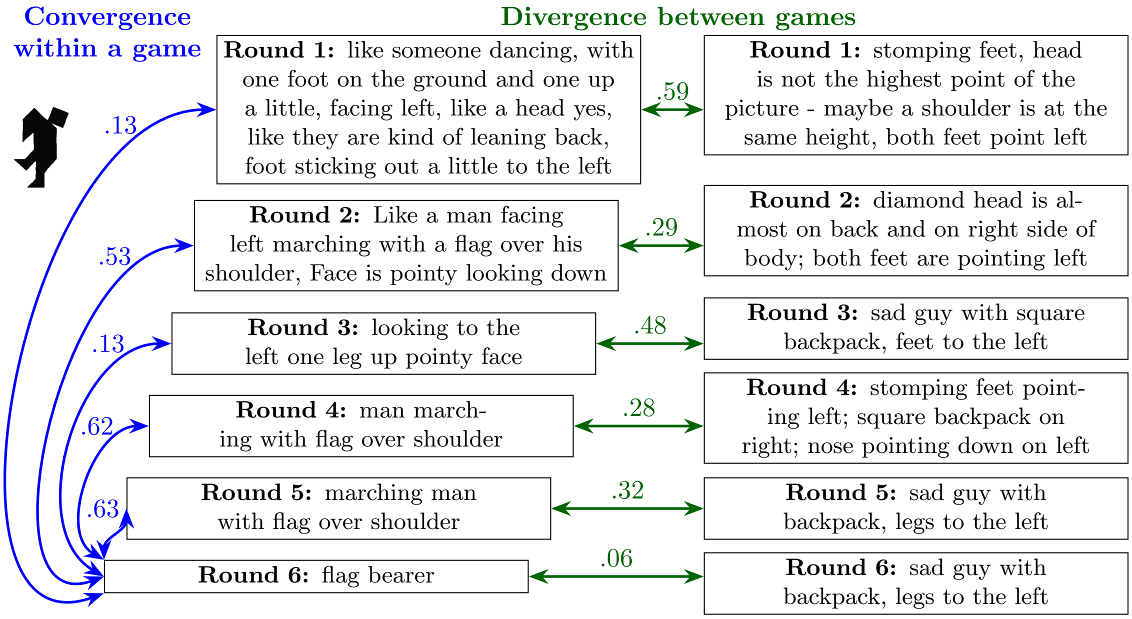

Convergence within a game
Divergence between games
Round 1: like someone dancing, with one foot on the ground and one up a little, facing left, like a head yes, like they are kind of leaning back, foot sticking out a little to the left
.59
Round 1: stomping feet, head is not the highest point of the picture - maybe a shoulder is at the same height, both feet point left
.13
Round 2: Like a man facing left marching with a flag over his shoulder, Face is pointy looking down
.29
Round 2: diamond head is almost on back and on right side of body; both feet are pointing left
.53
Round 3: looking to the left one leg up pointy face
.48
Round 3: sad guy with square backpack, feet to the left
.13
Round 4: man marching with flag over shoulder
.28
Round 4: stomping feet pointing left; square backpack on right; nose pointing down on left
.62
Round 5: marching man with flag over shoulder
.32
Round 5: sad guy with backpack, legs to the left
.63
Round 6: flag bearer
.06
Round 6: sad guy with backpack, legs to the left

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.

acting through a *convention formation* process (Clark & Wilkes-Gibbs 1986). Under a recent models 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

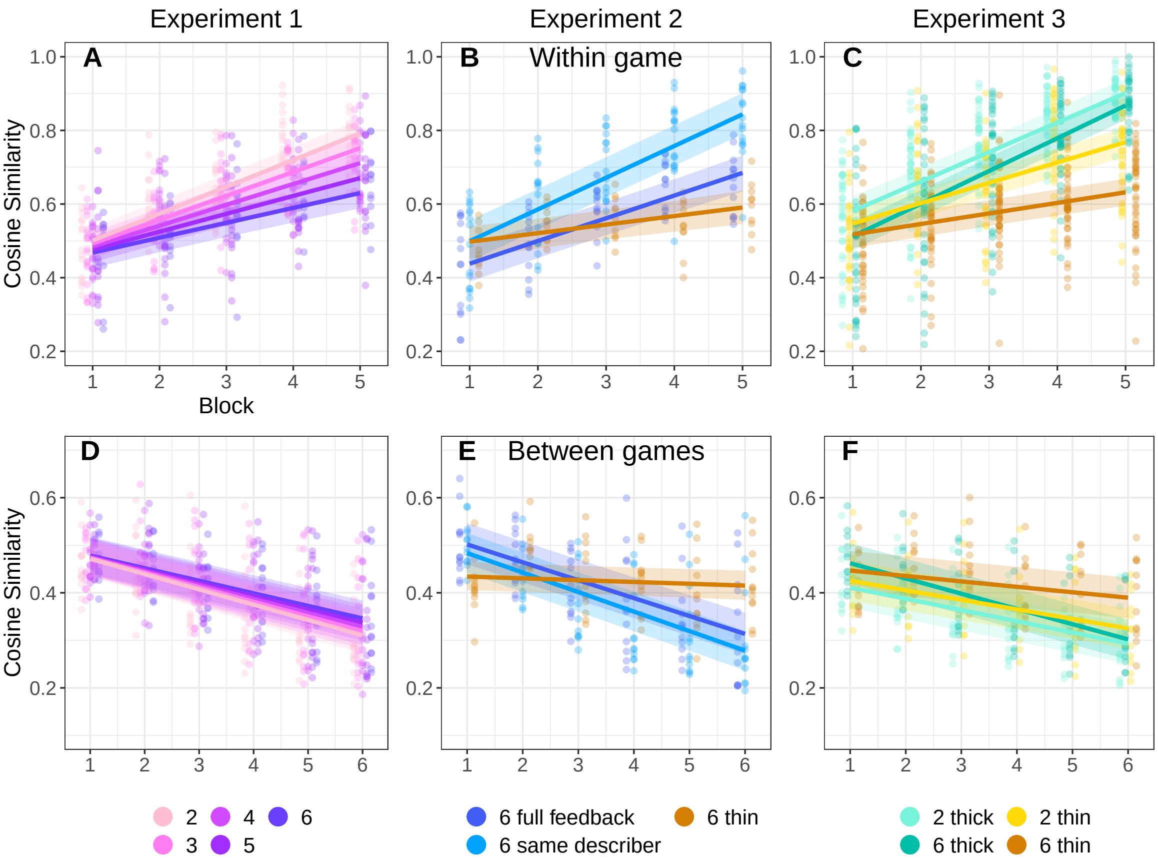


Figure 4: Language similarity results measured with pairwise cosine similarity between embeddings of two utterances. (A-C). Convergence of descriptions within games as measured by similarity between an utterance from block 1-5 to the block 6 utterance in the same game for the same image. (D-F). Divergence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

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 concatenated utterances and their within-game and between-game cosine similarities.

Descriptions converge faster in groups with higher coherence.

Across conditions, describer descriptions increased in semantic similarity to the final description over repetition, indicating convergence toward a stable description; convergence was fastest in smaller and higher coherence groups (Figure 4A-C; SI Tables 19-23).

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.

Smaller groups reached more stable descriptions faster than larger games. In Experiment 1, initial similarity was invariant across group size ($\beta = 0.517,\ 95\%\ \text{CrI} = [0.458, 0.573]$), but smaller groups converged faster (Figure 4A, $\beta = 0.089,\ 95\%\ \text{CrI} = [0.076, 0.102]$). In experiment 3, 6-player thick games started off further from their eventual convention than 2-player thick games ($\beta = 0.009,\ 95\%\ \text{CrI} = [0.001, 0.017]$) but closed the gap over time (Figure 4C, $\beta = -0.035,\ 95\%\ \text{CrI} = [-0.047, -0.025]$).

Thicker games converged 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 ($\beta = 0.08,\ 95\%\ \text{CrI} = [0.074, 0.086]$).

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-player thin games showed substantially slower convergence even compared to 2-player thin games (expt 3, $\beta = -0.025,\ 95\%\ \text{CrI} = [-0.033, -0.017]$).

Aggregating across experiments confirms this pattern: Thin games converge less than thick games ($\beta = -0.016,\ 95\%\ \text{CrI} = [-0.025, -0.008]$), and larger thinner games are especially slow to converge ($\beta = -0.007,\ 95\%\ \text{CrI} = [-0.01, -0.004]$).

Across games, convergence towards the last utterance was driven by cumulative increasing similarity between pairs of utterances in adjacent blocks (Supplement Figure 4D-F, SI Tables 34-38). 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.

All conditions showed some convergence toward a conventional nickname for the picture, but the speed of convergence was affected both by group size and channel width. Overall, stable descriptions emerged earlier if the describer role was consistent, if the group was smaller, and if matchers could contribute via a text channel.

Games diverge from one another faster with thicker channels.

While groups may initially overlap in their descriptions, including details of shapes or body parts, their descriptions are predicted to become increasingly dissimilar as groups increasingly adapt to their shared history.

We modeled semantic divergence using a mixed-effects linear regression model predicting the similarity between a pair of utterances for the same image as a function of the block number and condition.

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 24-28). However, different game sizes and interaction structures revealed very different strengths of divergence.

Smaller games used more group-specific language. In experiment 1, smaller games diverged more quickly than larger games in experiment 1 ($\beta = -0.035,\ 95\%\ \text{CrI} = [-0.038, -0.032]$). In experiment 3, 2-player thick games started off more dissimiliar than 6-player thick games, although 6-player games diverged faster and eventually approached the dissimilarity levels of 2-player thick games (SI Table 28).

Thicker interaction structure was associated with stronger group-specific divergence. In experiment 3, 2-player thin games diverged more slowly than 2-player thick games ($\beta = -0.024,\ 95\%\ \text{CrI} = [-0.025, -0.023]$). As with the convergence patterns, large games with thin interaction structures had the flattest trajectories, as thinness and largeness compounded. 6-player thin games diverged even less than 2-player thin games (Figure 4F, $\beta = 0.004,\ 95\%\ \text{CrI} = [0.002, 0.005]$), and in experiment 2, 6-player thin games barely diverged at all (Figure 4E, $\beta = 0.434,\ 95\%\ \text{CrI} = [0.406, 0.465]$).

A mega-analytic model confirms this pattern: thin games differentiate less between groups ($\beta = 0.005,\ 95\%\ \text{CrI} = [0.004, 0.007]$) and large thin groups differentiate even less ($\beta = 0.004,\ 95\%\ \text{CrI} = [0.004, 0.005]$).

Interim summary

Smaller groups show higher within-group similarities and between-group differences, sometimes showing up in the initial round and sometimes developing as a change over time. The thicker the games the faster and stronger the divergence and convergence patterns. The combination of a large game and a thin communication channel hampers within-game convergence and between-game divergence much more than either game size or thinness independently, leading to little evidence of group adaptation or group specificity in 6-player thin games.

General 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, form of matcher backchannel, 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 thus reveal important factors for communication 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. **TODO is there a better word than "comparable" to mean "not that different" possibly rephrase so it's clear we aren't proving the null** 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 **TODO what does "multi-modal" mean here?** multi-modal solution based on group priors (Guilbeault et al. 2021). Thus, we encourage measures of semantic content (and not just performance) when evaluating convention formation.

Just 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.

We cannot make claims about causal mechanisms between how experimental set-ups such as group size resulted in different outcomes: for instance, there are many differences between being in a 2-person group versus a 6-person group that could lead to the different outcomes. 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). These effects likely vary by both the knowledge state of and communication channels available to the comprehenders (Horton & Gerrig 2002, Horton & Gerrig 2005, Fox Tree & Clark 2013). 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, and this in turn could then point towards future experimental conditions.

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

Methods

For all experiments, we used Empirica (Almaatouq et al. 2020) to create real-time multi-player iterated reference games. In each game, one of the players started as the describer who saw an array of tangrams with one highlighted and communicated which figure to click to the other players (matchers). After the describer had described each of the 12 images in turn, the process repeated with the same images over a total of 6 such blocks (72 trials). We recorded what participants said in the chat, as well as who selected what image and how long they took to make their selections.

These experiments were designed sequentially and pre-registered individually.[1] We followed the analysis plan for each, although accuracy models were not explicitly specified until experiment 3, and linguistic analyses were only verbally described starting with 2b. Results from some pre-registered models are omitted from the main text for brevity but are shown in the Supplement.

Participants

Participants were recruited using the Prolific platform, and all participants self-reported as fluent native English speakers on Prolific's demographic prescreen. Participants each took part in only one experiment. Experiment 1 took place between May and July 2021, experiment 2 between March and August 2022, and experiment 3 in October 2022. As games varied in length depending on the number of participants, we paid participants based on group size, with the goal of a $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 had the describer role for the entirety of a 6-player game, they gained an additional $2 bonus. Across all games, each participant could earn up to $2.88 in performance bonuses. A total of 1319 people participated across the 3 experiments, for roughly 20 games in each condition in experiments 1 and 2 and 40 games per condition in experiment 3. A breakdown of number of games and participants in each condition is shown in SI Table 1.

Materials

We used the 12 tangram images used by Hawkins et al. (2020) and Clark & Wilkes-Gibbs (1986). These images were displayed in a grid with order randomized for each participant (thus descriptions

[1] Experiment 1: https://osf.io/cn9f4 for the 2-4 player groups, and https://osf.io/rpz67 for the 5-6 player data

such as “top left” were ineffective as the image might be in a different place on the describer’s and matchers’ screens). The same images were used every block.

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

From Prolific, participants were directed to our website where they navigated through a self-paced series of instruction pages explaining the game. Participants had to pass a quiz on the instructions to be able to play the game. They were then directed to a “waiting room” screen until their partner(s) were ready.

On each trial, the describer described the highlighted tangram image so that the matchers could identify and click it. All participants were free to use the chat box to communicate, but matchers could only click once 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.

Once all matchers had selected (or a 3-minute timer ran out), participants were given feedback. Matchers learned whether they had chosen correctly or not; matchers who were incorrect were not told the correct answer. The describer saw which tangram each matcher had selected, but matchers did not. 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.

In each block, each of the 12 tangrams was indicated to the describer once. 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 hyppotheses 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.

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 one way. The same describer condition differed only in that one person was designated the describer for the entire game, rather than having the describer role rotate. The full feedback condition differed from experiment 1 in that all participants were shown what each person had selected and what the right answer was; matchers still saw text saying whether they individually were right or wrong. This condition was similar to some dyadic work, such as Hawkins et al. (2020), where matchers were shown the right answer during feedback. For 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 instructions. They could send the emojis as often as desired, for instance, initially indicating confusion, and later indicating understanding. In addition,

run later. Experiment 2: same describer at https://osf.io/f9xyd, full feedback at https://osf.io/j5zbm, and thin at https://osf.io/k5f4t. Experiment 3: https://osf.io/untzy we added notifications that appeared in the chat box saying when a player had made a selection.

Experiment 3

The thin channel condition in experiment 3 was the same as the thin condition in experiment 2, above. The thick condition combined the two group coherency enhancing variations from experiment 2: one person was the designated describer throughout, and the feedback to participants included the right answer and what each player had selected. Across both conditions in experiment 3, notifications were sent to the chat to indicate when a participant had made a selection.

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 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 full and continued even after a participant disconnected (with describer role reassigned if necessary), unless the game dropped below 2 players. The distribution of players in these games that were initially recruited to be 6 player games is in SI Figure 1. 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 1).

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 what hierarchical effects were reasonable. Models of accuracy had by-group random intercepts; models of reduction had full mixed effect structure; models of S-BERT similarities had random intercepts per game and image as applicable. (All model results and priors and formulae are reported in the Supplement). Models of matcher accuracy were logistic models with normal(0,1) priors for both 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 correlation prior of lkj(1). For all of the models of SBERT similarity, we used linear models with the priors normal(.5,.2) for intercept, normal(0,.1) for beta, and normal(0,.05) for sd.

As an additional post-hoc analysis, we did mega-analytic models combining data across all experiments. For these models, we contrasted the 3 thin conditions (2c, and the two thin conditions of experiment 3) with all the other conditions lumped together as thick-ish. We coded game size as a continuous measure.

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. We note that this choice will underestimate differences between 2-player and (genuine) 6-player games, by labeling some smaller groups as 6-player groups.

We do not know 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 don't know to what extent groups that start and continue at the full size may 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 1 and SI Figure 1.

References

Ahern TC (1994) The effect of interface on the structure of interaction in computer-mediated small-group discussion. *Journal of Educational Computing Research* **11**:235–250

Almaatouq A, Becker J, Houghton JP, Paton N, Watts DJ, Whiting ME (2020) Empirica: A virtual lab for high-throughput macro-level experiments. *ArXiv200611398 Cs*

Almaatouq A, Griffiths TL, Suchow JW, Whiting ME, Evans J, Watts DJ (2022) Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences. *Behavioral and Brain Sciences*:1–55. doi:10.1017/S0140525X22002874

Branigan H (2006) Perspectives on multi-party dialogue. *Research on Language and Computation* **4**:153–177

Brennan SE, Clark HH (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**: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**:531–559. doi:10.1177/1046496498295001

Cazden CB (1988) Classroom discourse: The language of teaching and learning. ERIC

Clark HH, Krych MA (2004) Speaking while monitoring addressees for understanding. *Journal of memory and language* **50**:62–81

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

Cohn-Gordon Reuben, Levy R, Bergen L (2019) The pragmatics of multiparty communication.

Dewhirst HD (1971) Influence of perceived information-sharing norms on communication channel utilization. *Academy of Management Journal* **14**: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**:481–486. doi:10.1111/1467-9280.00292

Fox Tree JE, Clark NB (2013) Communicative Effectiveness of Written Versus Spoken Feedback. *Discourse Processes* **50**:339–359. doi: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**:961–987

Ginzburg J, Fernandez R (2005) Action at a distance: The difference between dialogue and multilogue. *Proceedings of DIALOR*:9

Guilbeault D, Baronchelli A, Centola D (2021) Experimental evidence for scale-induced category convergence across populations. *Nat Commun* **12**:327. doi: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. In: *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.* Association for Computational Linguistics, Florence, Italy, p 1895–1910. Available from: https://www.aclweb.org/anthology/P19-1184 [Last accessed 1 February 2022]. doi:10.18653/v1/P19-1184

Hawkins RD, Frank MC, Goodman ND (2020) Characterizing the dynamics of learning in repeated reference games. *ArXiv191207199 Cs*

Hawkins RD, Franke M, Frank MC, Goldberg AE, Smith K, Griffiths TL, Goodman ND (2023) From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review* **130**:977

Hiltz SR, 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**:225–252

Horton WS, Gerrig RJ (2002) SpeakersÕ experiences and audience design: Knowing when and knowing how to adjust utterances to addresseesq. *Journal of Memory and Language*:18

Horton WS, Gerrig RJ (2005) The impact of memory demands on audience design during language production. *Cognition* **96**:127–142. doi:10.1016/j.cognition.2004.07.001

Ji A, Kojima N, Rush N, Suhr A, Vong WK, Hawkins R, Artzi Y (2022) Abstract visual reasoning with tangram shapes. In: *Proceedings of the 2022 conference on empirical methods in natural language processing.*p 582–601

Krauss RM, Bricker PD (1967) Effects of transmission delay and access delay on the efficiency of verbal communication. *The Journal of the Acoustical Society of America* **41**:286–292

Krauss RM, Weinheimer S (1964) Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychon Sci* **1**:113–114. doi:10.3758/BF03342817

Krauss RM, Weinheimer S (1966) Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology* **4**:343–346. doi:10.1037/h0023705

Krauss RM, Garlock CM, Bricker PD, McMahon LE (1977) The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology* **35**:523

Kraut RE, Lewis SH, Swezey LW (1982) Listener responsiveness and the coordination of conversation. *Journal of personality and social psychology* **43**:718

Lewis D (1969) Convention: A philosophical study. John Wiley & Sons

MacMillan J, Entin EE, Serfaty D (2004) Communication overhead: The hidden cost of team cognition. In: *Team cognition: Understanding the factors that drive process and performance.* American Psychological Association, Washington, DC, US, p 61–82. doi:10.1037/10690-004

Metzing C, Brennan SE (2003) When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language* **49**:201–213. doi:10.1016/S0749-596X(03)00028-7

Parisi JA, Brungart DS (2005) Evaluating communication effectiveness in team collaboration. In: *Ninth european conference on speech communication and technology (INTERSPEECH).*

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. doi:10.48550/arXiv.1908.10084

Rogers SL, Fay N, Maybery M (2013) Audience Design through Social Interaction during Group Discussion. *PLOS ONE* **8**:e57211. doi:10.1371/journal.pone.0057211

Schober MF, Clark HH (1989) Understanding by addressees and overhearers. *Cognitive Psychology* **21**:211–232. doi:10.1016/0010-0285(89)90008-X

Seaman CB, Basili VR (1997) Communication and organization in software development: An empirical study. *IBM Systems Journal* **36**:550–563

Swaab RI, Galinsky AD, Medvec V, Diermeier DA (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**:25–53

Tannen D (2005) Conversational style: Analyzing talk among friends. Oxford University Press

Tolins J, Fox Tree JE (2016) Overhearers Use Addressee Backchannels in Dialog Comprehension. *Cogn Sci* **40**:1412–1434. doi:10.1111/cogs.12278

Traum D (2004) Issues in Multiparty Dialogues. In: Dignum F (ed) *Advances in Agent Communication.* Springer Berlin Heidelberg, Berlin, Heidelberg, p 201–211. Available from: http://link.springer.com/10.1007/978-3-540-24608-4_12 [Last accessed 1 February 2022].

doi:10.1007/978-3-540-24608-4_12

Weber RA, Camerer CF (2003) Cultural Conflict and Merger Failure: An Experimental Approach. *Manag Sci* **49**:16

Wilkes-Gibbs D, Clark HH (1992) Coordinating beliefs in conversation. *Journal of memory and language* **31**:183–194

Wittgenstein L (1953) Philosophical investigations. Wiley-Blackwell, New York, NY, USA

Yoon SO, Brown-Schmidt S (2014) Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **40**:919–937. doi:10.1037/a0036161

Yoon SO, Brown-Schmidt S (2018) Aim Low: Mechanisms of Audience Design in Multiparty Conversation. *Discourse Processes* **55**:566–592. doi:10.1080/0163853X.2017.1286225

Yoon SO, Brown-Schmidt S (2019) Audience Design in Multiparty Conversation. *Cogn Sci* **43**:e12774. doi:10.1111/cogs.12774

Zack MH (1993) Interactivity and communication mode choice in ongoing management groups. *Information Systems Research* **4**:207–239