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]Princeton University

Abstract

Real-world communication frequently requires speakers to address more than one listener at once. As the audience size grows, speakers face new challenges that do not arise in one-on-one conversation. 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 under these challenging conditions. We began by developing a novel extension of the dyadic repeated reference game paradigm for groups ranging from two to six participants. Across 313 games (1319 participants), we manipulated several key constraints on the group's interaction, including the thickness of the feedback channel and the availability of lateral interaction within the audience of listeners. While larger groups suffered disproportionately under these constraints, we found that they were able to converge on efficient shared conventions as effectively as smaller groups under the most favorable interaction structure. Overall, these results shed new light on the core structural factors that enable communication to thrive in larger groups.

(TODO check ordering of legends on Figures!) TODO fix discussion —

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, speakers need to account for the fact that different listeners 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 listeners 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).[1] What enables speakers and listeners 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 access to a shared workspace (Clark & Krych 2004, Garrod et al. 2007). Yet group settings 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), several factors have been identified in previous qualitative work that may slow such deterioration, each of which relates to the "thickness" of feedback (Ahern 1994, Parisi & Brungart 2005).

*Corresponding author. Email: vboyce@stanford.edu

[1]Throughout this paper, we use "speaker" and "listener" to refer to the roles describing and selecting targets, regardless of communication modality.

In this paper, we systematically manipulated these factors in a multi-party repeated reference game paradigm. Reference, or the ability to distinguish one particular entity from other possible entities, 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 to allow a listener to successfully choose 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: speakers must simultaneously achieve shared reference with multiple listeners. 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 breakthroughs 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. Our findings illuminate the mechanisms of social interaction in larger groups and suggest design features that may ease the communciative overhead associated with larger groups in real-world settings.

Results

We recruited 1319 participants through an online crowd-sourcing platform. Participants were organized into 313 groups for a multi-party repeated reference game. In each session, a set of 12 tangram images (Clark & Wilkes-Gibbs 1986, Hawkins et al. 2020, Ji et al. 2022) were presented repeatedly over a series of blocks, and a speaker was asked to use a chat box interface to describe a privately indicated *target* image for a group of listeners. After listeners chatted and eventually made a selection, all players received task feedback. The experiment consisted of 72 trials structured into 6 repetition blocks, where each image was described exactly 6 times over the course of the game (Figure 1B).

Groups were assigned to one of 11 distinct conditions in 3 distinct experiments (Figure 1A). These conditions systematically sampled points along four dimensions that parameterized the interaction space. We manipulated *group size* (ranging from two to six), *role rotation* (whether or not participants took turns being the speaker), richness of *task feedback* (whether or not listeners were able to see one another's responses), and richness of the *listener backchannel* (whether listeners were able to freely respond through the chatbox). Other factors, such as the set of stimuli and background knowledge about one's partners, were held constant across games.

In experiment 1, we varied the size of the games continuously from *2 to 6 players* while keeping other factors constant to explore what how performance varied across group size. For these conditions, the *speaker role rotated* after each block, so that all players had at least one turn as speaker. We gave *limited feedback* to listeners, showing each listener only whether their individual selection was correct or not, but not revealing others' selections or the right answer. Listeners had access to a *chat box* as their backchannel, so they could freely type questions and offer their own descriptions to the group.

In experiment 2, we explored varying different factors within 6-player games. We tried two different variations that we expected to improve group coherence and lead to better performance: having a *consistent speaker* rather than a rotating speaker, and separately, showing all the listeners *full feedback* on what each person in the group had selected and what the right answer had been. Additionally, we tested the role of listener contributions in establishing mutual understanding with a condition where listeners' backchannel was limited to four *emojis*. Listeners could send 4 discrete messages (green check, thinking face, red x, and laughing-crying face) to the chat. This limited backchannel allowed listeners to convey valence and level of comprehension, but not to contribute any referential content.

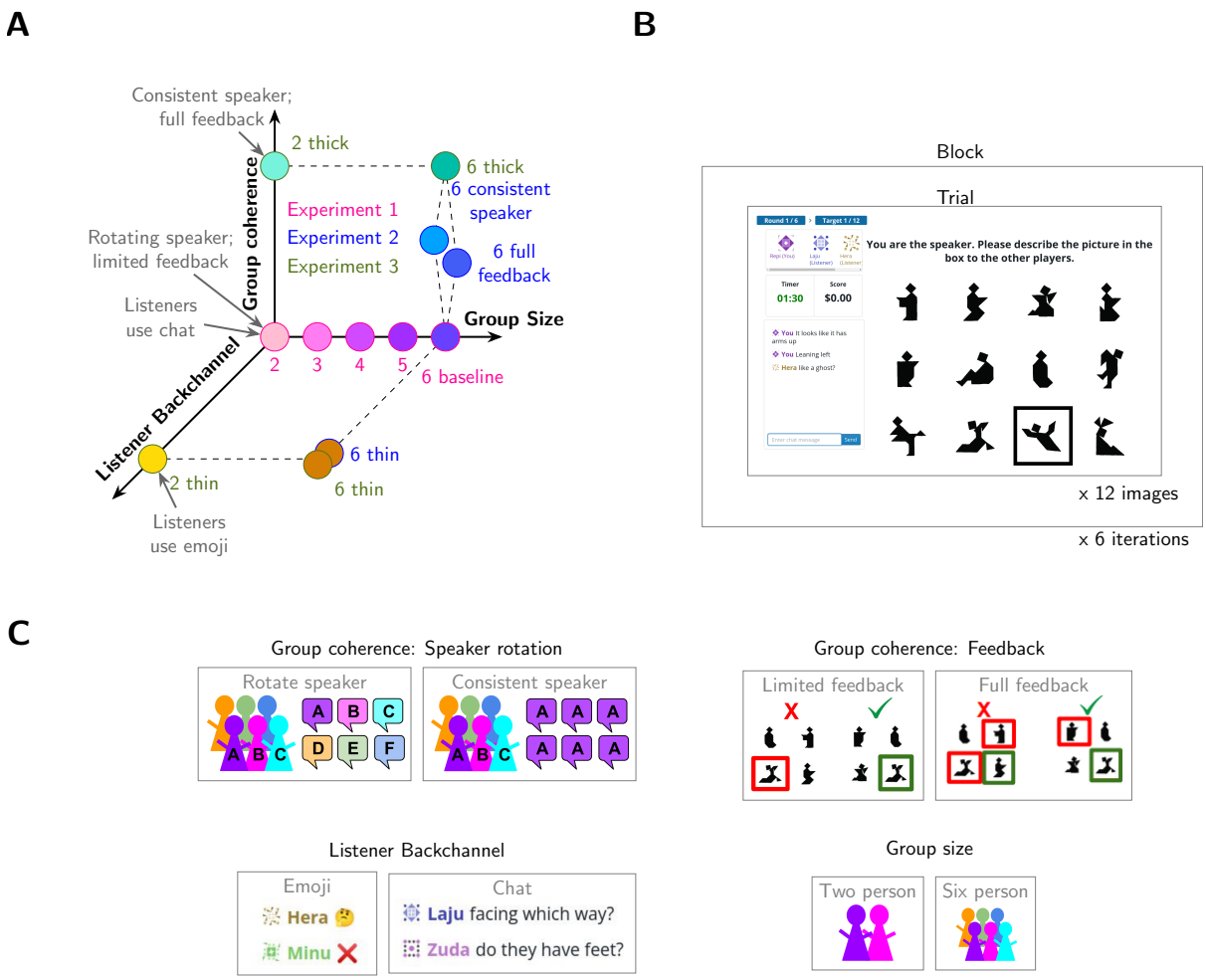

Figure 1: A: Diagram of the experimental space. Experiments varied along 3 dimensions: Group size, group coherence, and listener backchannel. Each condition is shown as a dot. Experiment 1 (pink labels) varied group size from 2-6 players while holding group coherence and backchannel constant. Experiment 2 (blue labels) keep group size constant at 6 and varied the other dimensions. Relative to experiment 1, 6 consistent speaker and 6 full feedback each added one component of group coherence, and 6 thin reduced the backchannel. Experiment 3 (green labels) tested 4 corners of the space, crossing group size (2 or 6 players) with thin games (low coherence, low backchannel) or thick games (high coherence, high backchannel). B: Each trial a speaker described a target image to the listeners, and this process repeated for all 12 images to comprise a block, and the block repeated for a total of 6 iterations. C: Differences between conditions. See text for explanation.

In experiment 3, we crossed the extremes of group size from experiment 1 (2 or 6 people) with extremes of group interactions from experiment 2. In the *thick* condition, we combined a consistent speaker with full feedback to create *high group coherence* and let listeners use the chat freely. In the *thin* condition, we repeated the emoji backchannel condition from experiment 2, which had *low group coherence* and an emoji backchannel. The 2-player thick game was similar to the condition used in Hawkins et al. (2020).

Group Performance

We begin by characterizing group performance along two complementary metrics: (1) listener accuracy and (2) speaker efficiency. Listener accuracy is given by the percent of listeners on each trial who successfully identify the target referent from the speaker's description. Speaker efficiency is given by the number of words produced by the speaker to achieve that degree of listener accuracy in the group. The degree to which speaker efficiency improves without decreasing listener accuracy can be interpreted as indicative of convergence on a more effective shared communication protocol within the group.

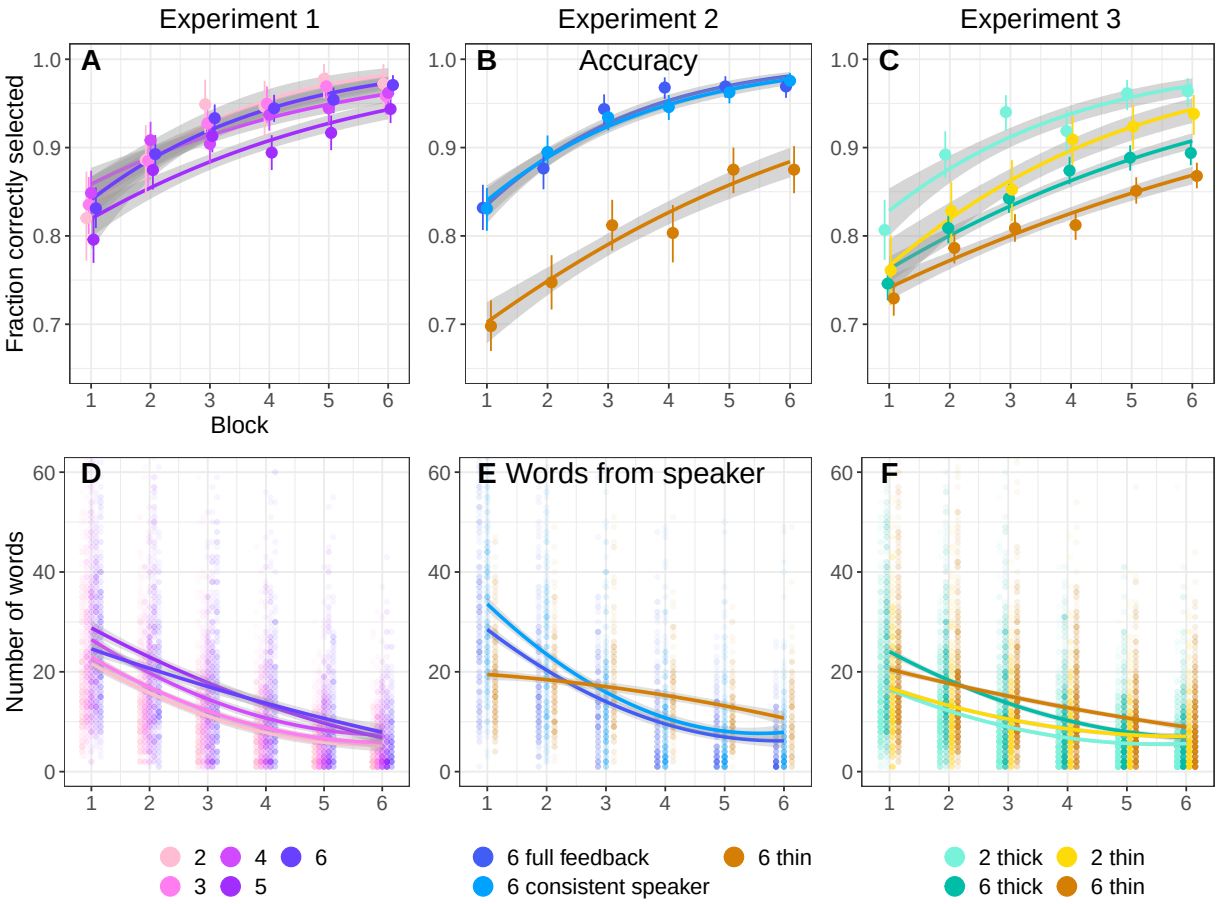

Figure 2: Behavioral results across all three experiments. A-C. Listener accuracy at selecting the target image. Dots are per condition, per block estimates with 95% bootstrapped CIs. Smooths are binomial fit lines. D-F. Number of words said by the speaker each trial. Faint dots represent individual trials from individual games. Smooths are quadratic fit lines. Y-axis is truncated, and a few outliers points are not visible.

Accuracy was high and increased.

In our experiments, listener accuracy rose over repetitions in all conditions and approached ceiling in most conditions; however, 6-player thin games were the least accurate (Figure 2A-C). We constructed logistic models to predict accuracy as a function of the manipulated variables, block, and their interactions. Both initial accuracy and improvement rate showed some variation based on group size and channel width, with smaller and thicker games having better and faster improving accuracy in some experiments. Group size did not have a strong effect on initial accuracy when measured continuously in the baseline condition (Figure 2A, $\beta = -0.07$, 95% CrI $= [-0.2, 0.05]$), but 6-player thick games had lower accuracy than 2-player thick games (Figure 2C, $\beta = -0.64$, 95% CrI $= [-1.05, -0.25]$). Improvement rates were similar, with no effect of group size in the baseline condition ($\beta = -0.02$, 95% CrI $= [-0.05, 0.01]$), but 6-person thick games were slower to improve than 2-person thick games ($\beta = -0.34$, 95% CrI $= [-0.43, -0.25]$).

Initial accuracy was somewhat higher for consistent speaker games ($\beta = 1.78$, 95% CrI $= [1.4, 2.19]$) and full feedback games ($\beta = 1.35$, 95% CrI $= [0.59, 2.06]$) than for 6 thin games (Figure 2B, $\beta = 0.88$, 95% CrI $= [0.64, 1.12]$), but the difference between thick and thin 2-player games was not reliable (Figure 2C, $\beta = -0.36$, 95% CrI $= [-0.78, 0.05]$). Similarly, improvement rates were higher for consistent speaker games ($\beta = 0.45$, 95% CrI $= [0.39, 0.52]$) and full feedback games ($\beta = 0.47$, 95% CrI $= [0.39, 0.54]$) than for 6 thin games ($\beta = 0.23$, 95% CrI $= [0.19, 0.28]$), but the difference in improvement rates between thick and thin 2-player games was not reliable ($\beta = -0.07$, 95% CrI $= [-0.18, 0.04]$). In sum, the high and increasing levels of accuracy indicate that across all of these conditions, participants are able to succeed in communicating about the images.

Speakers said more in larger games.

Speakers in larger games were more verbose than speakers in smaller games, and in some cases, these speakers showed sharper reduction from initial wordiness to eventual concision (Figure 2D-F). We constructed linear models of the number of words a speaker said each trial as a function of condition and block.

Speakers in larger groups said more to start with than speakers in smaller groups in both baseline conditions (Figure 2D, $\beta = 1.6, 95\%\text{CrI} = [0.62, 2.6]$) and thick conditions (Figure 2F, $\beta = 7.51, 95\%\text{CrI} = [3.63, 11.3]$). The numbers of words said decreased at a similar rate across group size in baseline conditions ($\beta = -0.09$, 95% CrI $= [-0.37, 0.18]$), but 6-player thick games decreased faster than 2-player thick games ($\beta = -1.22$, 95% CrI $= [-2.06, -0.29]$).

Thin 2-player games were similar to thick 2-player games in initial verbosity (Figure 2F, $\beta = 0.8, 95\%\text{CrI} = [-2.85, 4.26]$) and reduction rate ($\beta = 0.29$, 95% CrI $= [-0.56, 1.23]$). However, 6-player thin games reduced more slowly than 6-player thick games ($\beta = 0.93$, 95% CrI $= [0.02, 1.9]$).

The main effect of being one block later was -3.36, (95% CrI $= [-4.56, -2.18]$) words per trial in experiment 1; -5.31, (95% CrI $= [-6.35, -4.3]$) words for 6 consistent speaker; -4.64, (95% CrI $= [-5.81, -3.53]$) words for 6 full feedback, -2.1, (95% CrI $= [-3.37, -1.12]$) words for 6 thin. In experiment 3, the effect of being one block later was -2.24, (95% CrI $= [-2.92, -1.57]$) for 2 thick condition, $\beta = -1.96$, (95% CrI $= [-2.64, -1.3]$) for the 2 thin condition, $\beta = -2.52$, (95% CrI $= [-3.16, -1.9]$) for the 6 thin condition, and $\beta = -3.46$, (95% CrI $= [-4.12, -2.8]$) for the 6 thick condition.

Overall, all conditions exhibited reduction, with variations in the speed at which speaker utterances got shorter which partially covaried with how initially verbose speakers were.

Larger groups lead to more listener backchannels.

Listeners talk much less than speakers, and listener contributions are concentrated in early trials. The more listeners, the more likely it was that some listener talked, and the more listeners said if they talked (Supplement Figure 2). We constructed a logistic regression predicting the presence of any listener utterances in experiment 1 as a function of group size and block, and a linear regression predicting the number of words said by listeners in experiment 1 as a function of group size and block.

The number of trials where any listener said anything related to the image was higher in larger groups (Supplement Figure 2A, $\beta = 0.79$, 95% CrI $= [0.58, 0.98]$) and declined across blocks ($\beta = -0.8$, 95% CrI $= [-0.97, -0.62]$). When referential language was produced, larger groups produced more language (Supplement Figure 2D, $\beta = 2.07$, 95% CrI $= [1, 3.12]$), but the difference in group size closed in later blocks ($\beta = -0.41$, 95% CrI $= [-0.72, -0.11]$).

This pattern is consistent with early listener involvement in establishing a common conceptualization by asking questions and offering alternative descriptions. Once a shared reference description was in place, listener descriptions are rarer and more perfunctory.

Emoji use is not directly comparable to referential language, but similar trends occurred in the thin games. Emoji use was more common in the 6-player games than the 2-player games and decreased over the course of the game (Supplement Figure 3).

Interim summary

According to these metrics of group performance and efficiency in iterated reference games, larger games are similar to smaller games, except with more talking, especially early in the games. Larger groups seem to generally take the time to elaborate descriptions in order for most listeners to understand, especially when listeners can ask specific clarifying questions. This leads to more communication by both speaker and listeners in early rounds for larger games, sometimes followed by sharp reduction once a shared conceptualization is agreed upon.

Linguistic Content

Partner-specific reduction is characterized by high accuracy and shortening utterances, but the key phenomena of interest is that pairs are forming joint conventions about how to refer to particular images. How well groups are able to do this may vary based on game condition: we might predict that it's easier for

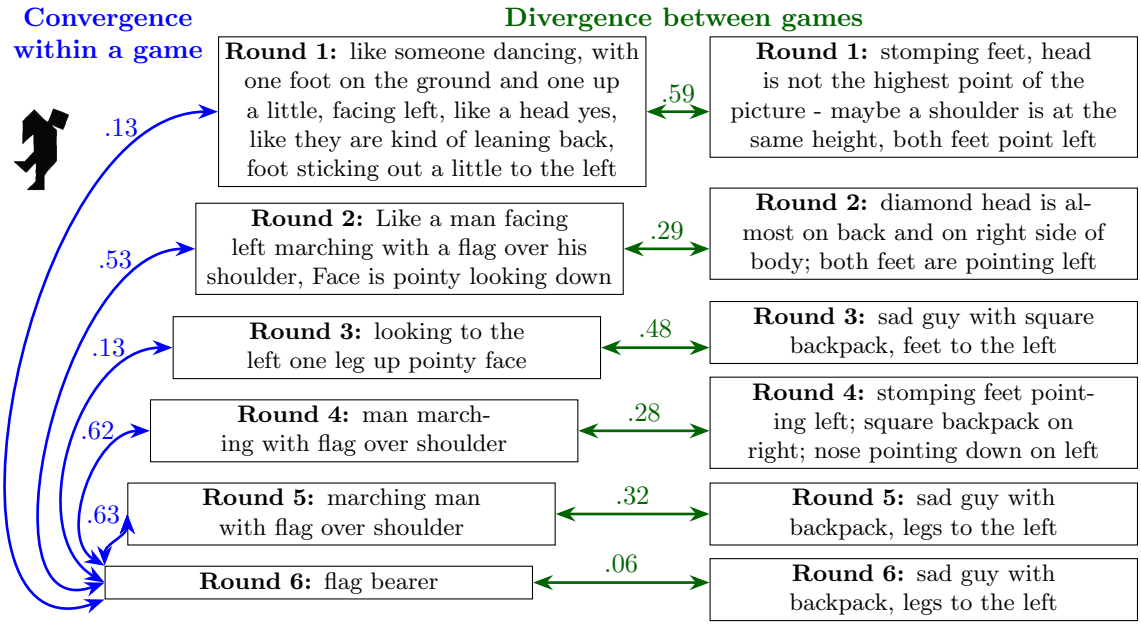

Figure 3: Example utterances describing the shown tangram figure produced by two 3-player games in Experiment 1. To measure convergence within a game (blue), we measured the cosine similarity between SBERT embeddings of descriptions and the embedding of the round 6 utterance (taken to be the convention).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

two people to agree on a label, but harder when the labels are coming from 6 different individuals who may have different conceptualizations. Convergence to a shared name might be faster if only one person provides the labels, since they can match their own labels. Feedback and listener contributions may help a group get on the same page about what descriptions go with what images, letting a speaker reduce to shorthand expressions. While initially many groups may overlap with descriptions that include descriptions of the shapes or body parts in the image, their descriptions are predicted to become increasingly dissimilar as these descriptive portions drop out, leaving just the conventionalized nicknames that are group-dependent.

To assess the linguistic patterns of speaker descriptions, we examine the *semantic similarity* of descriptions within and across games. We quantified description similarity by concatenating speaker 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 near each other in embedding space, which enables making semantically meaningful comparisons between sentences by taking pairwise cosine similarities between the embeddings (Reimers & Gurevych 2019). Thus, we measure the similarity between two utterances by the cosine similarity between their embeddings.

To measure convergence within games to shared nicknames, we compared utterances from blocks 1-5 to the corresponding final (block 6) description for the same image from the same game. As games develop different nicknames, descriptions from different groups will diverge, which we measured by comparing utterances across games for the same image in the same block. Figure 3 illustrates these two measures with example concatenated utterances and their within-game and between-game cosine similarities.

Descriptions converge within groups.

Across conditions, speaker descriptions increased in semantic similarity to the final description over repetition; convergence was fastest in smaller and higher coherence groups, and was least strong in the 6-player thin condition (Figure 4A-C). We modeled semantic convergence by looking at the similarity between a block 1-5 utterance and the corresponding block 6 utterance as a function of the earlier block number and condition.

The similarity of the first utterance to the last utterance was invariant across group size

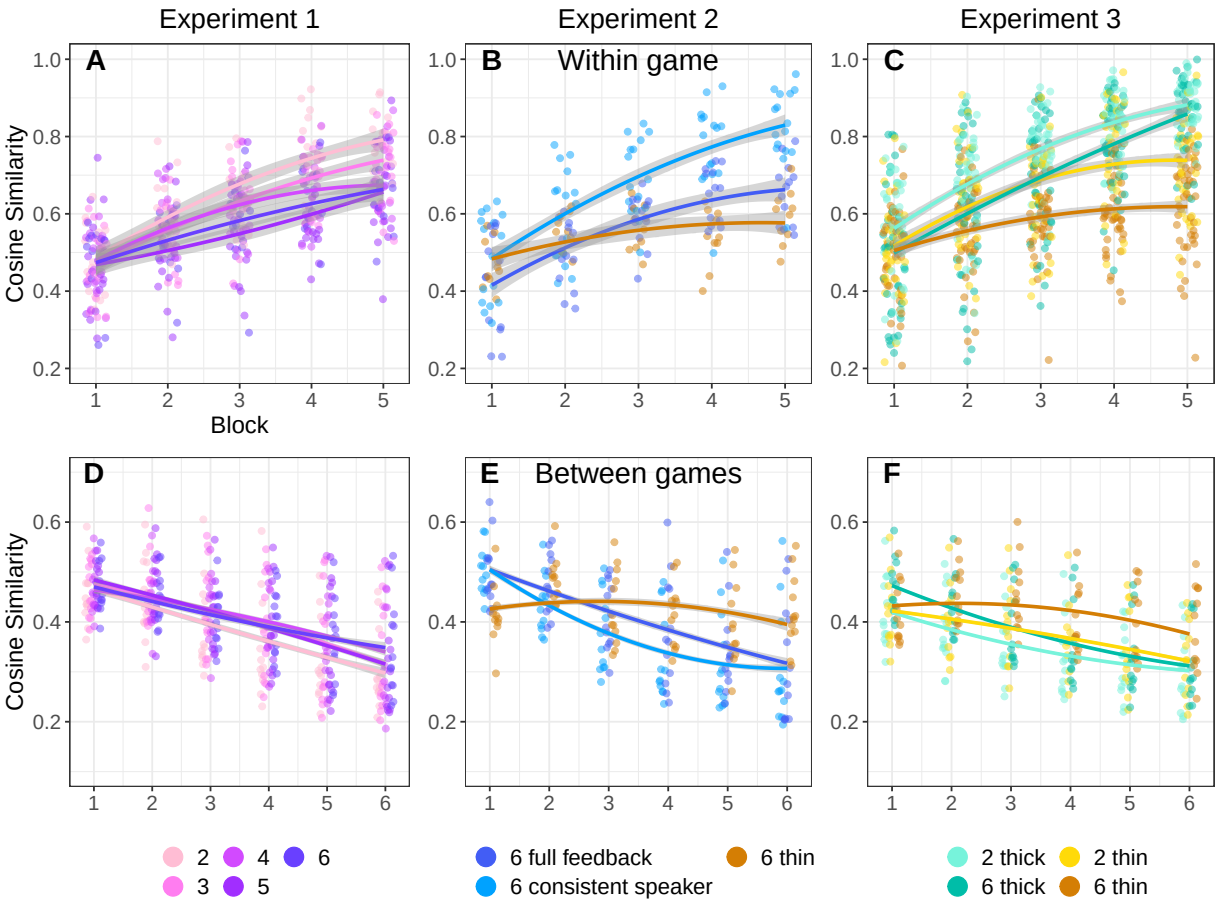

Figure 4: Language similarity results measured with pairwise cosine similarity between embeddings of two utterances. A-C. Convergence of descriptions within games as measured by similarity between an utterance from block 1-5 to the block 6 utterance in the same game for the same image. Dots are per-game averages, smooths are quadratic. 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. Dots are per-image averages, smooths are quadratic.

($\beta = -0.008$, 95% CrI $= [-0.021, 0.005]$), but smaller groups converged faster (Figure 4A, $\beta = -0.008$, 95% CrI $= [-0.011, -0.005]$). The 6-player thick games started off with greater distance between first and last utterances than 2-player thick games ($\beta = -0.069$, 95% CrI $= [-0.113, -0.025]$) but closed the gap over time ($\beta = 0.009$, 95% CrI $= [0.001, 0.017]$). Overall, smaller games reach a stable description faster than larger games.

Convergence was especially rapid for the consistent speaker condition where all the utterances come from the same person (Figure 4B, $\beta = 0.086$, 95% CrI $= [0.078, 0.094]$). Convergence was slower in thin games than thick games (Figure 4C, $\beta = -0.025$, 95% CrI $= [-0.033, -0.017]$).

Beyond the slower convergence in thin games, 6-player thin games showed substantially slower convergence even compared to 2-player thin games (expt 3, $\beta = -0.035$, 95% CrI $= [-0.047, -0.025]$).

Across games, convergence towards the last utterance was driven by cumulative increasing similarity between pairs of utterances in adjacent blocks (Supplement Figure 4D-F). 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, descriptions were more similar if provided by the same person, if fewer people were in the game, and if listeners could contribute via a text channel.

Language in thicker games diverges faster from other games.

Over repetitions, speaker descriptions diverged from descriptions used in other groups: divergence was fastest in groups with thick communication channels, while the 6-player thin condition games barely diverged at all (Figure 4D-F). We modeled semantic divergence by looking at the similarity between a pair of utterances for the same image from the same block across different games as a function of the block number and condition.

Initial similarities between groups were the same regardless of group size (Figure 4D, $\beta = 0.002$, 95% CrI $= [0, 0.004]$), but smaller groups diverged from each other slightly faster than larger groups ($\beta = 0.001$, 95% CrI $= [0.001, 0.002]$). The 2-player thick condition diverged at a moderate speed (Figure 4F, $\beta = -0.024$, 95% CrI $= [-0.025, -0.023]$), and the 6-player thick condition had initially higher similarity ($\beta = 0.051$, 95% CrI $= [0.047, 0.055]$ and faster divergence ($\beta = -0.008$, 95% CrI $= [-0.01, -0.007]$ in comparison.

Divergence was stronger in the consistent speaker (Figure 4E, $\beta = -0.041$, 95% CrI $= [-0.043, -0.039]$) and full feedback conditions ($\beta = -0.038$, 95% CrI $= [-0.04, -0.035]$) than in the 6 thin condition ($\beta = -0.004$, 95% CrI $= [-0.006, -0.001]$). Compared to the 2-player thick games, 2-player thin games started with slightly higher similarity (Figure 4F $\beta = 0.014$, 95% CrI $= [0.01, 0.018]$) and diverged slightly more slowly ($\beta = 0.004$, 95% CrI $= [0.002, 0.005]$).

The most noticeable effect in this analysis was that the 6-player thin games diverged much more slowly than the other conditions. The interaction between larger group size and thinner channel was associated with a much lower rate of divergence (Figure 4F, $\beta = 0.017$, 95% CrI $= [0.015, 0.019]$).

Divergence between groups, a sign of increasing group-specificity of descriptions, occurred across all conditions. Smaller games and games with thicker channels diverged more than larger or thinner games. In particular, the 6-player thin games showed a qualitatively different pattern of minimal divergence, diverging even less than would have been expected based on the rates in the other 3 conditions in Experiment 3.

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, as seen in the difference between 6 thin and either 2 thin or 6 thick.

General Discussion

Communication often occurs in multi-party settings, but research on referential communication often does not. In dyadic work, iterated reference games have been used to establish a phenomena of reduction over repeated reference, characterized by speaker-listener pairs creating short nicknames that they mutually understand, but which are not shared by other groups. In this work, we asked how this process of reference formation unfolds under varying interaction structures.

All conditions show hallmarks of reduction, and interaction structure constrains speed of convention formation.

Across 3 online experiments and 11 experimental conditions, we varied game features including group size, form of listener backchannel, and degree of group coherence. All conditions showed the hallmarks of reduction: increasing accuracy, reduction in speaker utterances, semantic convergence within games, and differentiation of descriptions between groups. Even with larger groups and more constrained means of communication, reduction still occurs.

However, while results are directionally the same across conditions, the interaction structure of a group substantially affects how rapidly groups develop partner-specific conventions. Smaller groups and games with thicker communication channels converged faster and more robustly than games that were larger or had thinner communication channels. These factors add together to form the overall group experience. The differences between the 6-player thin condition and both the 2-player thin condition and other 6-player

conditions point to an interaction: 2-player games can cope with limited feedback mechanisms, but 6-player games suffer without access to more feedback. Group dynamics differ depending on group size, and larger groups may be more sensitive to other factors affecting interaction structure. Multi-player groups thus make for a richer and more sensitive environment to study communication phenomena applicable to both pairs and small groups.

Reduction and partner-specific convention formation pattern separately.

Theoretical approaches treat reduction in referring expression length as a consequence of partner-specific convention formation (Clark & Wilkes-Gibbs 1986, Brennan & Clark 1996, Yoon & Brown-Schmidt 2014, 2018), but in our current work, reduction and semantic measures of partner-specificity pattern differently in the 6-player thin condition. The 6-player thin games show much less divergence between games and convergence within games, even compared to the 2 thin and 6 thick conditions, but 6-player thin games showed smaller (and statistically inconconclusive) differences to 6 thick games for accuracy and reduction. This gap between the group performance measures and the semantic measures raises the possibility that it is possible to become more concise (and more accurate) without developing group-specific nicknames, but instead perhaps relying on group priors and reducing the amount of detail (Guilbeault et al. 2021). This gap highlights the need to use measures of the type of language (and not just amount of language) when looking for convention-formation phenomena.

Limitations and future directions.

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 listeners to 4 discrete utterances (emojis), but there are other ways to manipulate channel width for speakers and listeners, 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, speakers can tailor their utterances to the one listener, but in large groups, speakers must balance the competing needs of different listeners (Schober & Clark 1989, Tolins & Fox Tree 2016). These effects likely vary by both the knowledge state of and communication channels available to the listeners (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.

Conclusion

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 communication space is necessary to better characterize the phenomena of interest.

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 speaker who saw an array of tangrams with one highlighted and communicated which figure to click to the other players (listeners). After the speaker had identified each of the 12 images in turn, the process repeated with the same images, but a total of 6 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.[2] We followed the analysis plan, although additional analyses were added to early experiments that were only pre-registered in later experiments. Results from some pre-registered models were omitted from the main text, 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 speaker role for the entirety of a 6-player game, they gained an addition $2 bonus. Across all games, each participant could early 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 the Supplement.

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 such as "top left" were ineffective as the image might be in a different place on the speaker's and listeners' 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 to be able to play the game. They were then directed to a "waiting room" screen until their partner(s) were ready.

Each trial, the speaker described the highlighted tangram image so that the listeners could identify and click it. All participants were free to use the chat box to communicate, but listeners could only click once the speaker had sent a message. Once a listener clicked, they could not change their selection. There was no signal to the speaker or other listeners about who had already made a selection.

Once all listeners had selected (or a 3-minute timer ran out), participants were given feedback. Listeners learned whether they individually had chosen correctly or not; listeners who were incorrect were not told the correct answer. The speaker saw which tangram each listener had selected, but listeners did not. Listeners got 4 points for each correct answer; the speaker got points equal to the average of the listeners' points. These points translated into performance bonus at the end of the experiment.

In each block, each of the 12 tangrams was indicated to the speaker once. The same person was the speaker for an entire block, but participants rotated roles between blocks. Thus, over the course of the 6 blocks, participants were speakers 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 speaker was chosen to keep participants more equally engaged (the speaker role is more work), and to give a more robust test for reduction and convention.

After the game finished, participants were given a survey asking for optional demographic information and feedback on their experience with the game.

[2] 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: consistent speaker 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

Differences in 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 consistent speaker condition differed only in that one person was designated the speaker for the entire game, rather than having the speaker 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; listeners still saw text saying whether they individually were right or wrong. This was similar to some dyadic work, such as Hawkins et al. (2020) where listeners were shown what the right answer was during feedback. For the thin condition, we altered the chatbox interface for listeners. Instead of a textbox, listeners had 4 buttons, each of which sent a different emoji to the chat. Listeners 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, we added notifications that appeared in the chat box saying when a player had made a selection.

Differences in 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 speaker throughout, and the feedback participants received 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 speaker role reassigned if necessary), unless the game would drop below 2 players. The distribution of players in these 6* player games is at Supplement 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 Supplement Table 1).

When skimming transcripts to tag non-referential utterances, we noticed that one game in the 6-player thick game had a speaker who did not give any sort of coherent descriptions, even with substantial listener prompting. We excluded this game from analyses.

Modelling strategy

In experiment 3, some of the 6-player games did not have 6 players for the entire game. We do not model this, as it is unclear at what point in the game group size is most relevant. We note that this is a conservative choice that will underestimate differences between 2-player and (genuine) 6-player games, by labelling some smaller groups as 6-player.

We ran all models in BRMS (Bürkner 2018) with weakly regularizing priors. We were often unable to fit the full mixed effects structure that we had pre-registered in a reasonable amount of time, so we included what heirarchical effects were reasonable. (All model results and priors and formulae are reported in the Supplement). Accuracy models were run as logistic models with normal(0,1) priors for both betas and sd. Reduction models 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.

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