Interaction structure constrains the emergence of conventions in group communication

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

[1]Stanford University

Abstract

This is an abstract in italics.

Keywords
One keyword; Yet another keyword

0.1 big groups

It's hard to talk with big groups – people aren't on the same page, they try to talk at once, there are disagreeing factions, not everyone has the same common ground. It's awful. And yet, we manage some of the time either to get everyone on the same page or to juggle the different knowledge levels or something.

0.2 review dyadic work

Despite the ubiquity of group communication in everyday life, dyadic communication is the typical paradigm in the lab. One goal of communication across settings is efficient communication which necessitates shared labels to refer to the topics of conversation (Traum 2004, Ginzburg & Fernandez 2005, Branigan 2006). In many cases, there are widely shared conventionalized expressions for objects or ideas, but in other cases, spontaneous ad-hoc expressions must be invented.

The formation of these new reference expressions is well-studied in dyadic contexts and has been a case study for efficient communication more broadly. Clark & Wilkes-Gibbs (1986) established an experimental method for studying the emergence of new referring expressions that has now become standard (building on Krauss & Weinheimer 1964, 1966). Two participants see the same set of tangram figures; the speaker describes each figure in turn so the listener can select the target from the set of figures. The speaker and listener repeat this process with the same images over a series of blocks. Early descriptions are long and make reference to multiple features in the figure, but in later iterations, shorthand conventional names for each figure emerge; this shortening of utterances is called 'reduction'.

Recently, online participant recruitment and web-based experiments have made it possible to study this convergence in larger populations (Haber et al. 2019, Hawkins et al. 2020). In line with results from face-to-face, oral paradigms, speakers reduced their utterances, producing fewer words per image in later blocks than in earlier blocks.[1]

*Corresponding author. Email: vboyce@stanford.edu

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

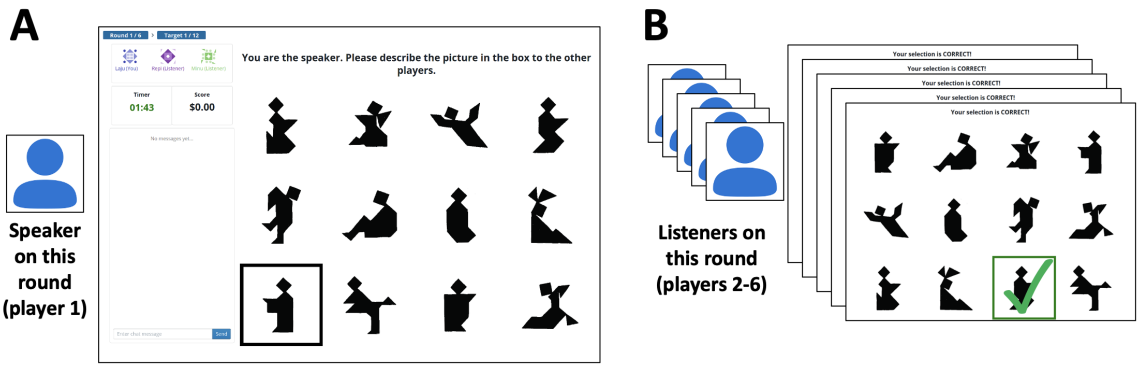

Figure 1: All participants saw all 12 tangram images. (A) Speaker's view during selection phase. (B) During the feedback stage, speakers saw what figure each person chose, but listeners only learned if their selection was correct or incorrect. Listeners were not shown what other listeners chose.

0.3 Current work

How does this process proceed in multi-party communication? 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).

In our current work, we test how the phenomena extend to larger games and games that vary in how coherent the group is and what communication channels are available to the listener. Using online recruitment and testing, we ran 3 experiments comprising 1319 participants across 313 groups of 2–6 participants each, who collectively produced 325579 words of utterances. We analyse the results using traditional metrics of accuracy and number of words, and use new NLP tools to get at how the semantic content of utterances shifts over the course of games.

TODO TOPLINE RESULTS??

Across three experiments, we extend the repeated dyadic reference game paradigm of Hawkins et al. (2020) along a few dimensions. All the experiments can be compared as they share the same core methods, with variations. All of the reference games use 12 target images [CITE CLARK HAWKINS] arranged in a grid, with positions randomized per player. The speaker saw one image picked out as the target and described it to their groupmates (listeners) over a chat interface so each listener could make a selection. After all listeners had selected, players recieved feedback on the selections. This repeated for all the images with the same speaker to comprise a 12 trial block. Then the whole process repeated for a total of 6 blocks. [TODO possibly can reduce this depending on what's in intro].

We parameterized the experiments along a few dimensions [see figure whatever]. One dimension was game size which varied between 2 and 6 player groups. Another dimension was group coherence, which was made up of two components: speaker rotation and feedback. In low group coherence games, the speaker rotated each block, while in high group coherence games, one player was the speaker for the entire game. In low group coherence games, each listener only recieved feedback on if they were indiviually right or wrong in their selection, while in high group coherence games, listeners (like the speakers in all games) saw who had selected what and what the target had been. The last dimension of variation was listener backchannel: in high backchannel games, the listeners could freely send (text) messages to the shared chat, while in the low backchannel games, listeners could send 4 discrete messages (represented as emojis) to the chat. TODO work in motivation/why for these variations?

As shown in Figure whatever, Experiment 1 had constant low group coherence and high listener backchannel while varying group size. Experiment 2 held group size constant at 6 and each condition deviated from experiment 1 in one aspect: 6 single speaker and 6 full feedback changed components of group coherence and 6 thin had a low backchannel. Experiment 3 tested 4 corners of the experimental space at larger scale, with thin (low backchannel, low coherence) and thick (high backchannel, high coherence) games with either 2 or 6 players.

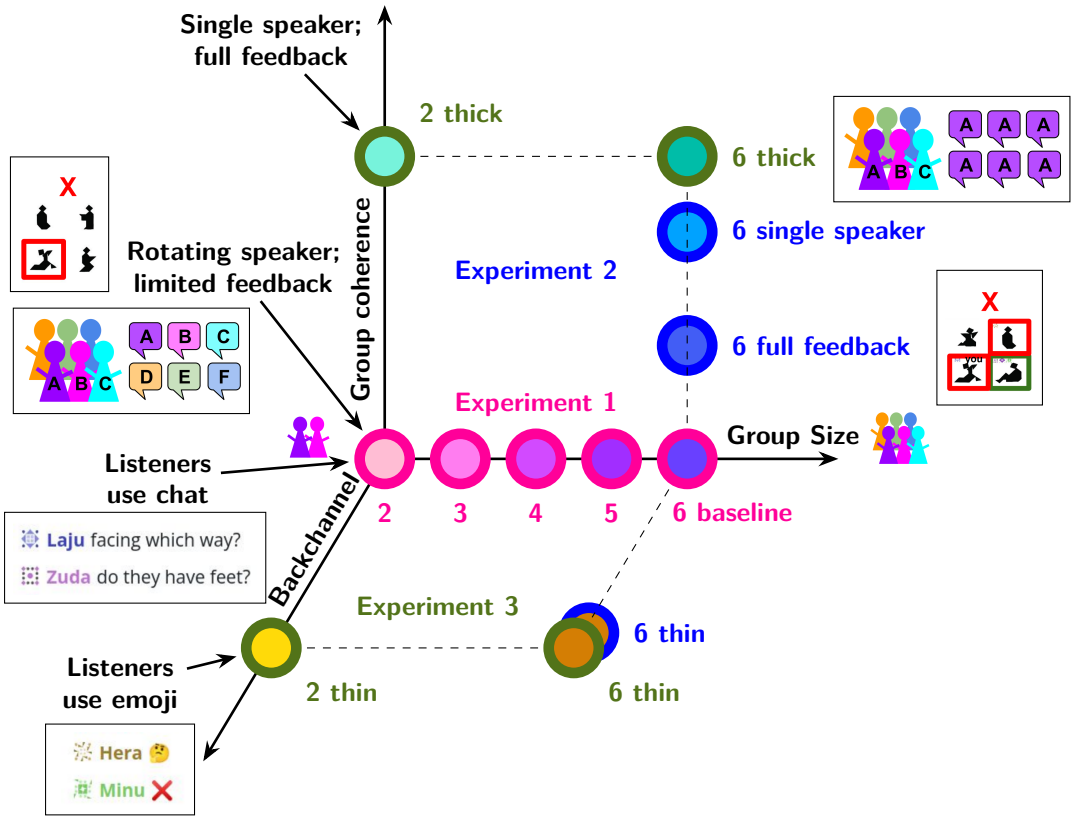

Figure 2: Diagram of the experimental space explored in these experiments. Experiment 1 (pink) has a backchannel where listeners can use the chat and low group coherence from a rotating speaker and limited feedback. Experiment 1's conditions vary the group size from 2 - 6 players. Experiment 2 (blue) games keep group size constant at 6 and vary along the other dimensions. 6 single speaker and 6 full feedback each add one component of group coherence relative to experiment 1. 6 thin varies the backchannel (relative to experiment 1) by having listeners communicate with emoji rather than the full text chat the speaker uses. Experiment 3 (green) games test 4 corners of the space, crossing group size (2 or 6 players) with thin games that have low group coherence and low backchannel or thick games that have high group coherence and high backchannel. TODO FIX ME AND MY ILLUSTRATION!!!

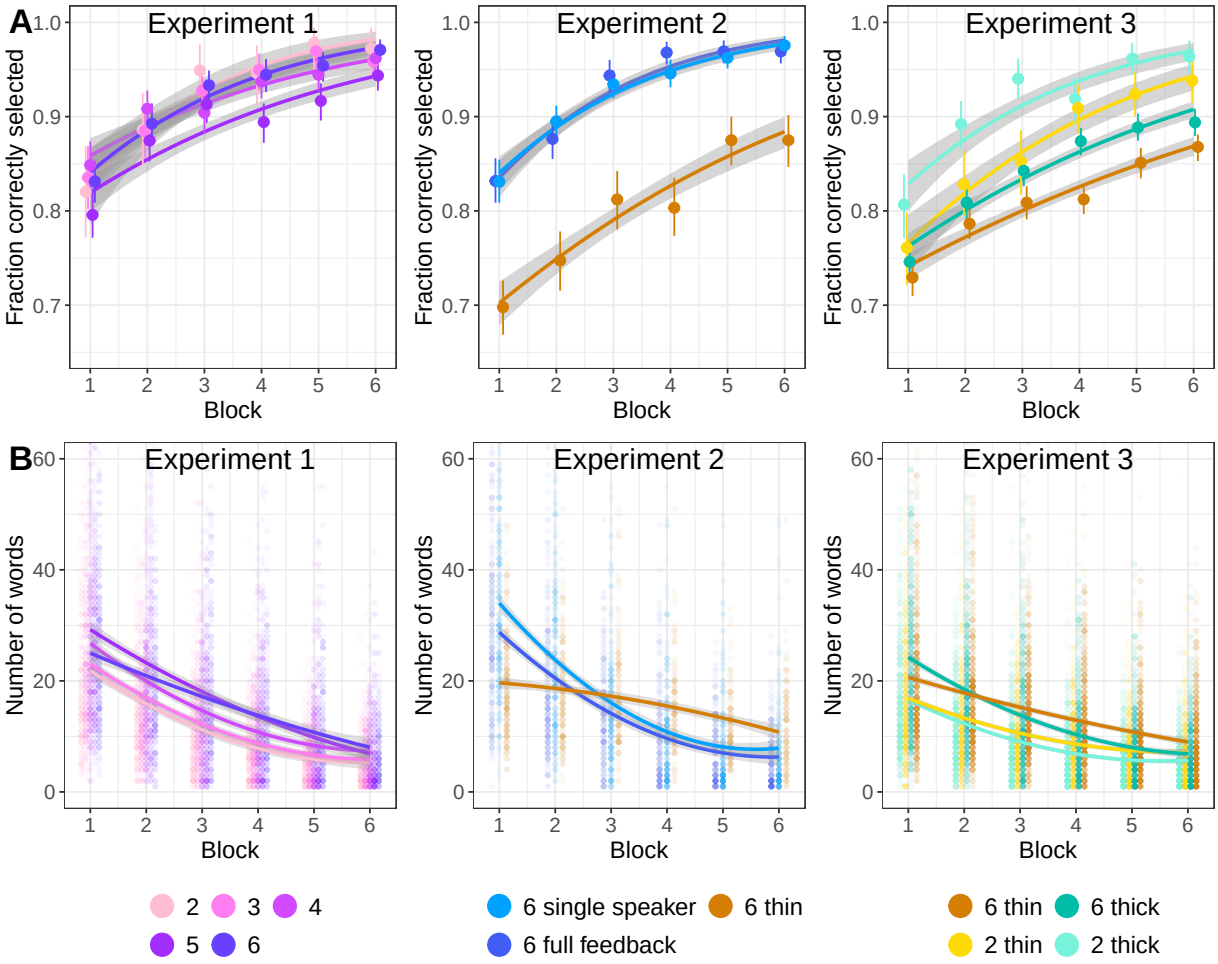

Figure 3: Behavioral results across all three experiments. A. Listener accuracy at selecting the target image. Dots are per condition, per block estimates with 95% bootstrapped CIs. Smooths are binomial fit lines. B. Number of words said by the speaker each trial. Faint dots represent individual trials from individual games. Smooths are quadratic fit lines. Y-axis is truncated, and a few outliers points are not visible.

1 Results

Across the three experiments and multiple conditions, games all had the same structure of 12 images repeated over six blocks, which leads to a high degree of comparability across conditions and experiments. We first compare the results on two behavioral measures of listener accuracy and speaker reduction of words that have been the common markers of reduction phenomena in the literature [CITE]. We then explore semantic patterns of reduction by comparing similarities of utterances to look at how the speaker's language changes within and between games over time.

1.1 Behavioral results

The two key behavioral outcomes were how accurately listeners selected the target images and how many words the speaker produced each trial.

Across all experiments, most individuals were accurate in their selections, with accuracy rising across blocks. Accuracy was noticeably lower, but still far above chance, in the 6 thin games (Figure 3A). In experiment 1, participants were more accurate in later blocks (block: 0.44 [0.31, 0.58]), and there was no strong effects of group size on overall accuracy (numPlayers: -0.07 [-0.2, 0.05]) or improvement rate (block:numPlayers: -0.02 [-0.05, 0.01]). In experiment 2, participants were again more accurate in later blocks (6 single speaker: block: 0.45 [0.39, 0.52], 6 full feedback: block: 0.47 [0.39, 0.54], 6 thin: block:

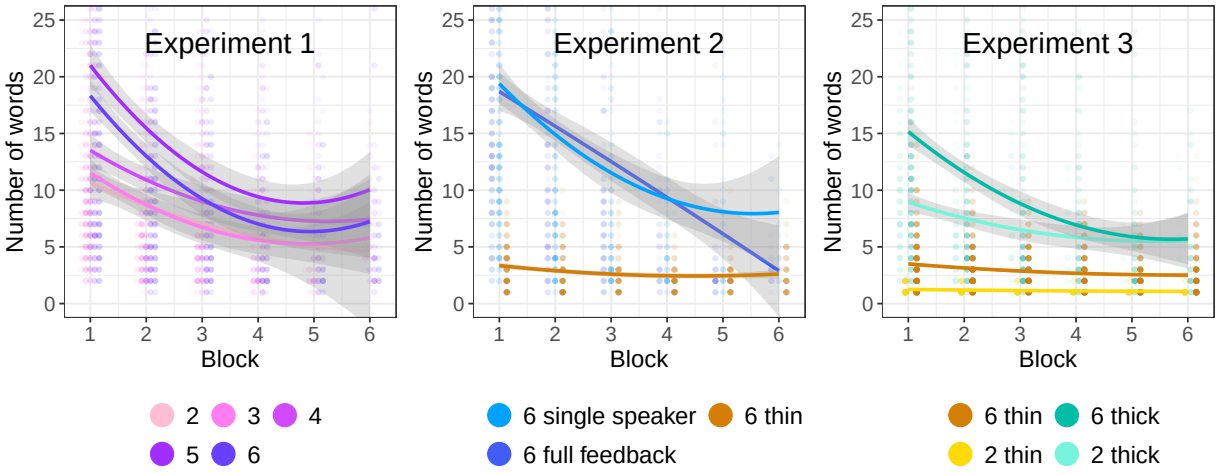

Figure 4: TODO this probably goes in a supplement!

0.23 [0.19, 0.28]). In experiment 3, participants were more accurate in later blocks (block: 0.41 [0.32, 0.5]). Participants in six player games were less accurate (gameSize6: -0.64 [-1.05, -0.25]) and slower to improve (block:gameSize6: -0.34 [-0.43, -0.25]). Thin versus thick games did not have a clear effect on accuracy (block:channelthin: -0.07 [-0.18, 0.04]) or improvement rate (block:channelthin: -0.07 [-0.18, 0.04]).

The high and increasing levels of accuracy indicate that across all of these conditions, participants are able to play the game and succeed in communicating about the images.

The key observation in iterated reference games is that the descriptions the speaker gives of the target images start out long and become shorter over the course of repetitions. This pattern of reduction held across all conditions, with the numbers of words from the speaker decreasing over blocks (Figure 3B). In experiment 1, the overall effect of being one block later was -3.37 [-4.54, -2.24] words. Speakers in larger groups said more; the effect of each additional player was 1.66 [0.66, 2.61] words per trial, with no clear interaction between block and group size (block:numPlayers: -0.1 [-0.36, 0.17]). In experiment 2, the result of being one block later on the number of words the speaker said per trial was block: -5.39 [-6.46, -4.31] for 6 single speaker; block: -4.68 [-5.88, -3.52] for 6 full feedback, and block: -2.15 [-3.44, -1.12] for 6 thin. The rate of reduction was lower in the thin condition than in the other conditions. In experiment 3, reduction occurred overall (block: -2.29 [-2.95, -1.6]). The six player games said more to start with (gameSize6: 7.41 [3.57, 11.18]) and reduced less (block:gameSize6: -1.21 [-2.06, -0.3]) than the two-player games. There were not differences due to channel type (channelthin: 0.63 [-3.18, 4.73]) or channel type over time (block:channelthin: 0.32 [-0.65, 1.24]).

These reduction results confirm and extend what was previously known for 2 player games. Behaviorally, larger games are mostly similar to smaller games, but their speakers tend to say more overall, perhaps related to the increased number of listeners to respond to.

1.1.1 TODO something about how much listeners talk & game anecdotes

Conclusions: should do total lang/game (including zeros) on chat conditions –> listeners don't use that much contentful language ever and it declines quickly

not sure what to say about non-contentful language use? such as yup / got it and chitchat maybe helpful to speaker (or not) and overlaps other channels.

Not sure how to talk about emoji since they do get used – maybe say the do get used, generally used on each trial, but not a direct comparison to contentful language?

notes: this is a slightly unfair comparison, since it's the filtered chat for the chat ones, but raw emojis for the others which might actually map to the "got it" chitchat. Resolve later.

While listener backchannel is implicated as an important way for listeners to get clarification and reach agreement, listeners don't talk that much. The average number of words of reference language per trial

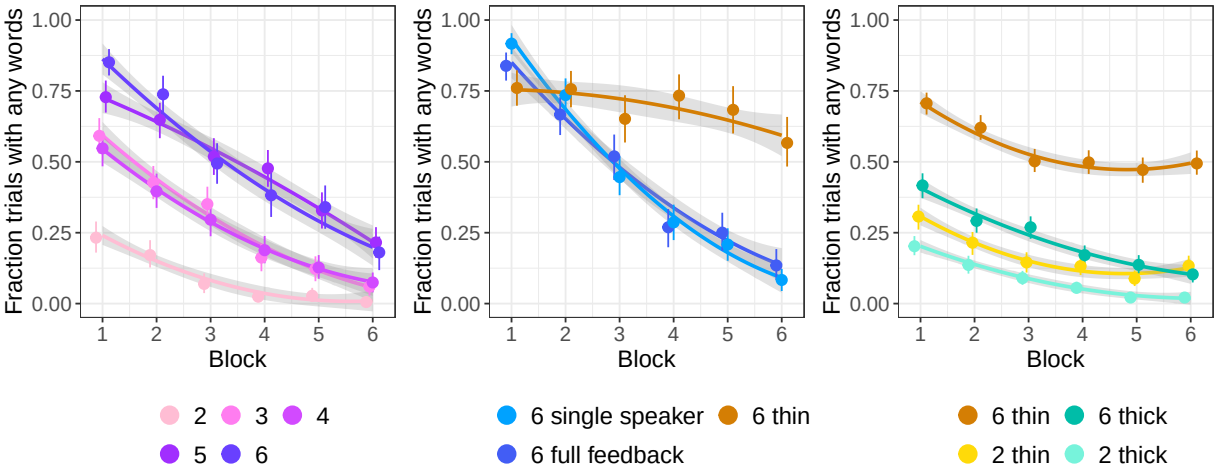

Figure 5: TODO this probably goes in a supplement!

per listener (counting those who say nothing) is less than 5 words TODO real numbers for the first block and declines in later blocks. Listeners don't talk very much

Key points about listener talking: They don't do it very much, but they do it a bit. Especially early in the game. Over the course of the game the amount of time any listener talks at all declines, as does the amount that is said. ## Comparisons of language between and within games

In addition to behavioral measures, we can also look at how the descriptions change over time within and between games. We focus on the speaker's description, concatentating it all together into one utterance (this includes any add ons or answers to questions). Then we use SBERT to embed the utterance in a high-dimensional vector space. This turns each speaker description into a long vector, where the vectors represent some of the semantic content of the utterance. Vectors that are close together represent utterances that are more similar to one another, so by looking at cosine similarity (a metric of how close pairs of vectors are) we can see how similar pairs of utterances are.

As a measure of convention formation, we can track how utterances describing the same tangram become increasingly similar over the course of a game. If conventions are forming we expect the similarity to the last block utterance to increase over the course of the game.

If different games go in different directions with their descriptions, we'd expect this similarity between descriptions of the same image in different games to decrease over repetitions.

TODO examples of this in a figure would be great!!!!

As a measure of convention formation, we look at the similarity of descriptions within a game for a particular tangram in different blocks. We take the last round descriptions as the established convention and measure the similarity between earlier speaker utterances and this convention.

We first look at convergence, comparing utterances from the first 5 rounds of a game to the "convention" or round 6 utterance for the same figure. In experiment 1, later utterances are more similar to the last utterance than earlier utterances are (0.09 [0.08, 0.1]). The similarity of the first utterance to last utterances is invariant across group size (-0.01 [-0.02, 0]), but smaller groups converge faster (-0.01 [-0.01, -0.01]). Experiment 2 shows similar patterns of utterances become more similar to the last utterance, particularly in the single speaker condition (0.09 [0.08, 0.09]) where all the utterances come from the same person, but also in the full feedback condition (0.06 [0.05, 0.07]) and to a smaller extent, in the thin condition (0.02 [0.01, 0.03]). In experiment 3, convergence to the last round utterances occurs overall(0.08 [0.07, 0.09]), but the convergence is slower in thin games (-0.02 [-0.03, -0.02]) and especially thin 6 player games (-0.04 [-0.05, -0.02]).

Not only do groups reduce the lengths of their utterances, but each group is converging towards a semantic description for the figures. This is more prominent in smaller games and games with greater group coherence.

The complement to convergence within groups is divergence between groups, as different groups develop their own ways of identifying the different figures. In experiment 1, descriptions become less similar to

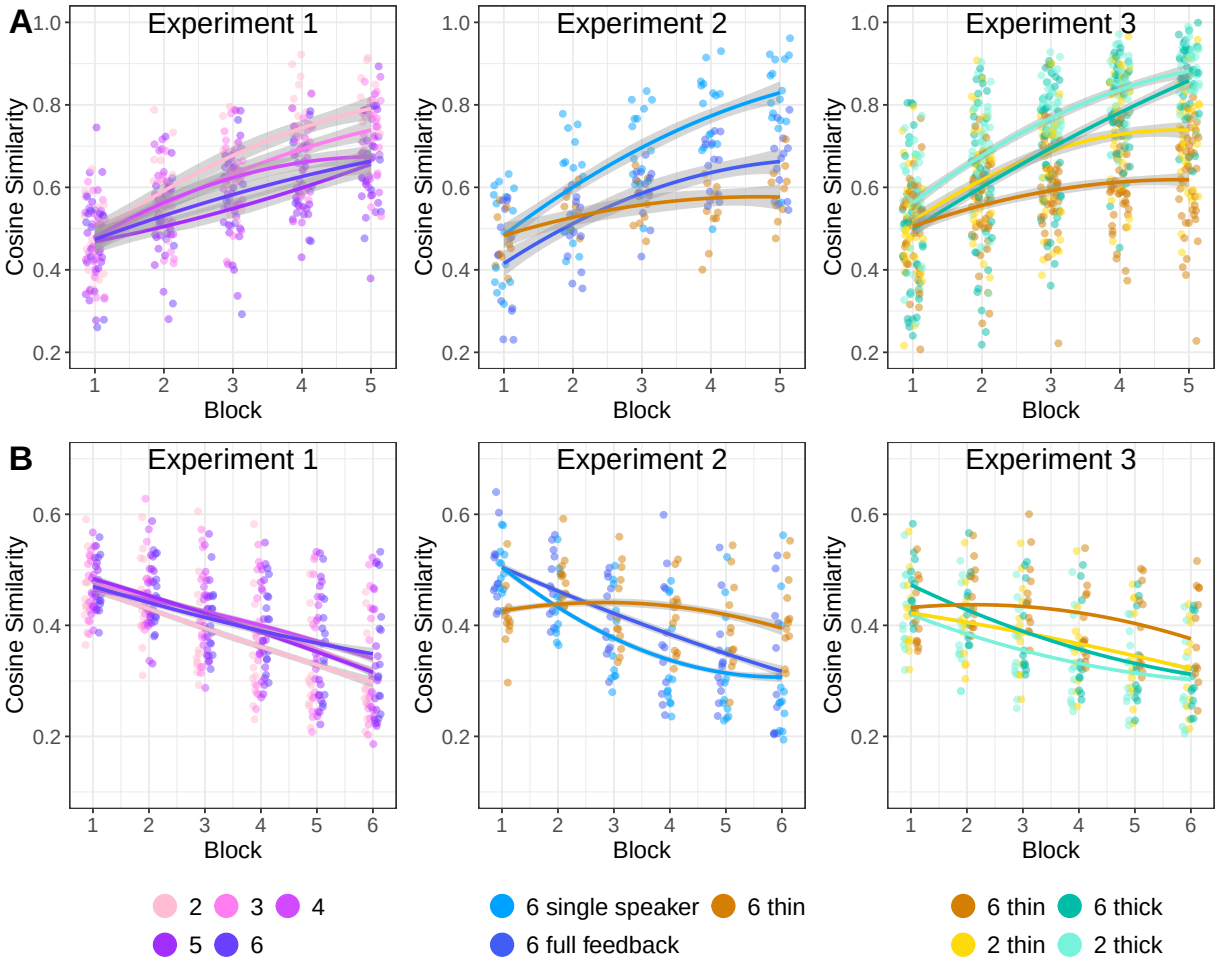

Figure 6: Language similarity results measured with pairwise cosine similarity between embeddings of two utterances. A. Convergence of utterances within games as measured by similarity between an utterance from block 1-5 to the block 6 utterance in the same game for the same image. Dots are per-game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

those used to describe the same figure in other games (-0.04 [-0.04, -0.03]). Group size does not affect the cross-groups similarities in the first block (0 [0, 0]), but smaller groups diverge from each other faster than larger groups (0 [0, 0]). In experiment 2, divergence is stronger is the single speaker (-0.04 [-0.04, -0.04]) and full feedback conditions (-0.04 [-0.04, -0.04]) than in the 6 thin condition (0 [-0.01, 0]). In experiment 3, descriptions from different games get less similar over time (-0.02 [-0.02, -0.02]). There are slight differences in the initial starting points across the different conditions, as well as slight condition differences in how fast the games diverge. In particular, 6 player thin games diverge more slowly (0.02 [0.02, 0.02]).

1.1.2 Distinctiveness of tangrams

Another way of looking at how language changes over the course of the game is looking at how games start to refer to different tangrams more differently. This could reflect initial overlap in descriping many figures as sitting or standing or by leg and arm and head position.

Over the course of the game, descriptions for each tangram become more distinctive (-0.04 [-0.05, -0.04]). In all three subexperiments, the descriptions of tangrams become more distinctive within games across time. (2a -0.05 [-0.05, -0.04], 2b -0.03 [-0.03, -0.02], 2c -0.03 [-0.03, -0.02]).

Tangram distinctiveness within games increased over time (-0.03 [-0.03, -0.03]). There might be more

to say about other effects, but it's mostly a starting places being different in larger games and then the slopes also differ a bit?

1.1.3 TODO there's a note saying to rerun these models for longer with more extensive mixed effects!! (at least of reduction model)

1.1.4 play with more diagrams

Comparing utterances between adjacent rounds reveals similar patterns. Thin games have lower similarity between adjacent blocks (-0.12 [-0.16, -0.09]) as do larger games (-0.03 [-0.07, 0]). Later in the game adjacent blocks are more similar than earlier adjacent blocks (0.05 [0.04, 0.05]), painting an overall nonlinear convergent pattern (as seen in Figure 7).

The last measure of how utterances change within games is how they compare to the first utterance; this is less good because the first utterance has more fluffy language so is less diagnostic, but later utterances are further from the first round utterance than earlier utterances (-0.03 [-0.04, -0.03]). (TODO it's in the pre-reg, but we could dump it in a supplement?)

2 General Discussion

this isn't the only group dynamic; could imagine situations where listeners can see each others work collaborate (point to each other what htey think, perhaps see feedback from speaker to one listener) which might make things reduce much faster

The emergence of conventions has been a key case study for communication more broadly. Yet this issue has – for the most part – been studied only in dyadic communication. While some studies have examined aspects of convention formation in larger groups (e.g., Yoon & Brown-Schmidt 2014, Yoon & Brown-Schmidt 2019), basic descriptive work has not yet investigated how group size changes the dynamics of interaction in a standard referential communication task, in part because such tasks can be difficult to administer to larger groups. Taking advantage of a new online multi-player experiment platform, we ran repeated reference games with groups of 2–6 players and characterized the nature of group performance.

Consistent with dyadic games, listeners' selection accuracy increased over blocks at the same time as listeners sped up their selections (question 1). Crucially, speakers reduced the length of their descriptive utterances as they conventionalized on concepts for each image (question 2). Because speakers rotated, this reduction finding is robust: not only did speakers say less in later repetitions than they themselves said earlier, speakers later in the order said less than speakers earlier in the rotation. This reduction varied with group size; smaller groups used shorter utterances, but group size did not significantly interact with block (question 3). The trajectory of reduction also depended on whether the current speaker correctly identified the tangram in the prior block and whether the current speaker was new to being speaker. This pattern is consistent with both the 'aim low' and 'aim middle' hypotheses from previous work (Yoon & Brown-Schmidt 2014, Yoon & Brown-Schmidt 2019).

What was specifically different across group sizes? Smaller groups showed more agreement in how each tangram was identified across blocks (question 4), coming to consensus earlier: Their overlap between descriptions in the first 5 blocks to the final block was higher, and words in the final block tended to originate earlier. The greater diversity in how tangrams were described in larger groups could be explained by slower convergence to a convention or parallel competing conceptualizations favored by different speakers. Larger groups have more people for the speaker to communicate to, but also more people who might interrupt with questions, and more people who have opinions about what each image looks like. Bigger groups differ from smaller groups in a number of ways, however, and disentangling these differences is an area for future work.

Group interactions are rich, and this experiment is necessarily a schematic simplification with a number of limitations. Real-life situations vary widely in who the interlocuters are, their relationships, their goals, and their environment (Carletta et al. 1998, Fay et al. 2000). Our participants were a convenience sample of Prolific workers who were strangers to each other; thus we miss richness that could come from prior relationships or shared community. Reference is only one goal out of many possible communicative goals, and the tangram images are artificial. We provided less feedback than previous studies such as Hawkins

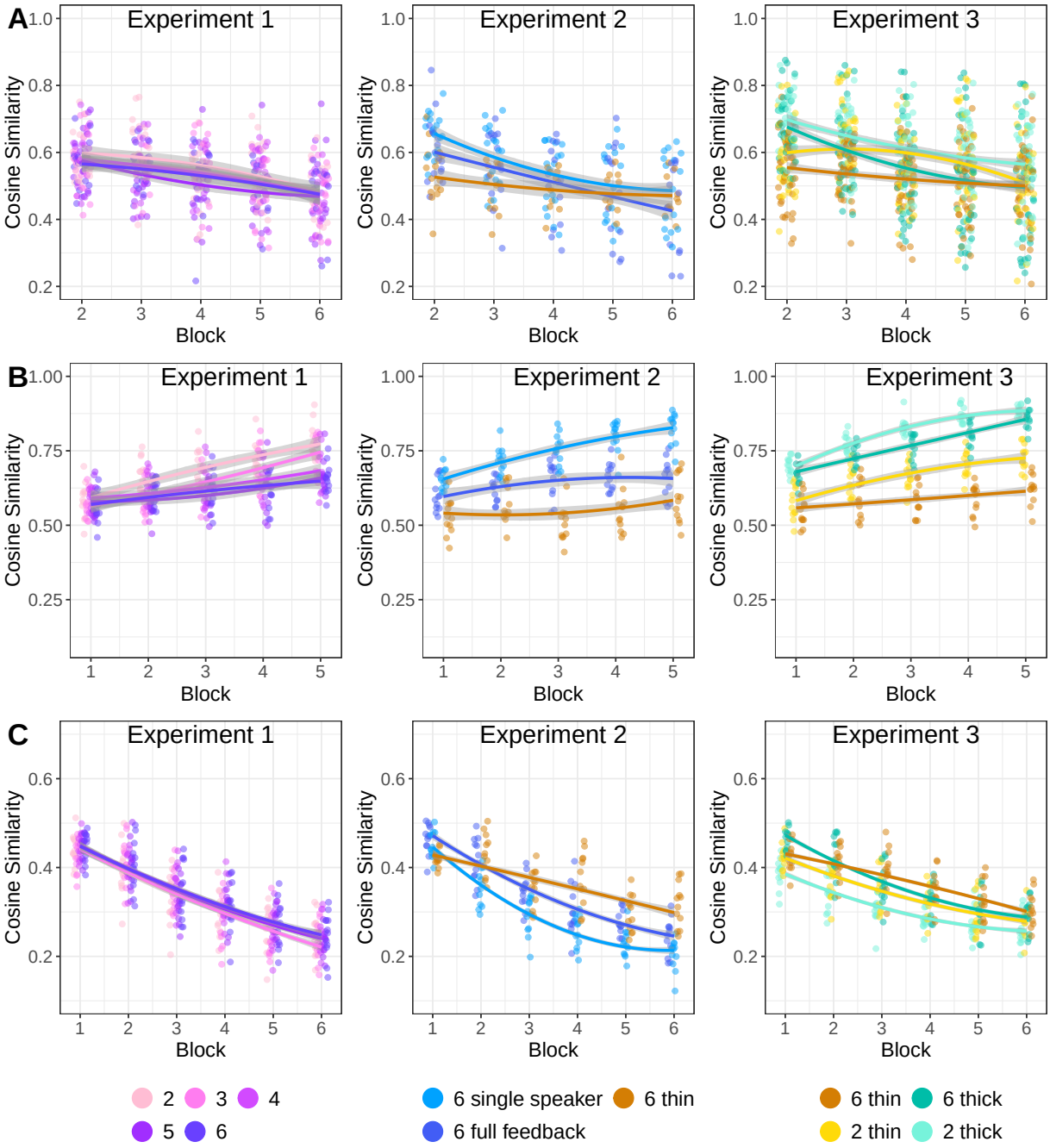

Figure 7: Stuff probably not to include. A is similarity to first utterance. B is similarity between utterances from adjacent blocks. C is divergence in descriptions of different tangrams within a group

et al. (2020); this regime imitates situations where interlocutors can't show each other examples, but it's not representative of all communicative environments. Further, our text-based online paradigm meant that participants' individual identities were not especially salient. In sum, communication takes place in a plethora of situations; our experiment provides some insights, but also misses many complexities that should be a focus of further experiments.

The experimental paradigm presented here could be a valuable tool to disentangle the mechanisms of group size and determine which design parameters are relevant to reduction. Luckily, with an online implementation, recruiting for and running experiments is feasible, and thus it will be possible to iterate on this experiment to determine how far the patterns generalize. While much is left to be explored, this initial data set provides a rich corpus of how humans adapt language dynamically to communicate.

Table 1: The number of games in each experiment and condition. Complete games finished all 6 blocks; partial games ended early due to disconnections, but contributed at least one complete block of data. 6* indicates that some games started with fewer than 6 players or continued with fewer than 6 players after participants disconnected.

Experiment	Players	Complete	Partial	Total Participants
1: baseline	2	15	4	38
1: baseline	3	18	2	60
1: baseline	4	19	2	84
1: baseline	5	17	3	100
1: baseline	6	12	6	108
2: single speaker	6	15	3	108
2: full feedback	6	13	4	102
2: thin	6	10	6	96
3: thin	2	35	3	76
3: thin	6*	44	0	235
3: thick	2	39	3	84
3: thick	6*	38	2	222

2.1 Limitations

3 Methods gore

We extended the dyadic repeated reference game paradigm of Hawkins et al. (2020) along a few dimensions. As diagrammed in Figure 2, three dimensions of variation we considered were group size, listener backchannel, and group coherence. In experiment 1, we expanded from dyadic reference games to group games with 2–6 players who rotated between speaker and listener roles. In experiment 2, we built on the 6 player games by exploring three variations, two which increased group coherence by increasing feedback to listeners or having a single speaker for the entire game, and one that reduced the listener backchannel. For experiment 3, we considered the extremes of group size and performance, informed by the prior experiments. The thin channel repeated the reduced-backchannel, low-group coherence condition, and we created a thick channel by combining the two sources of group coherence together. We then crossed these thick and thin condition with groups of 2 and 6 players and collected more data in each of these conditions.

For all experiments, we used Empirica (Almaatouq et al. 2020) to create real-time multi-player reference games. In each game, one of the players started as the speaker who saw an array of tangrams with one highlighted (Figure 1A) 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] TODO more comments on pre-reg

3.1 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.

[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: single 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

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. A breakdown of number of games and participants in each condition is shown in Table 1.

3.2 Materials

We used the 12 tangram images used by Hawkins et al. (2020) and Clark & Wilkes-Gibbs (1986) (see Figure 1). 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.

3.3 Procedure

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

3.3.1 Experiment 1

We implemented the experiment using Empirica, a Javascript-based platform for running real-time interactive experiments online (Almaatouq et al. 2020). 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 partners were ready.

Once the game started, participants saw screens like Figure 1A. 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 (Figure 1B). 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.

3.3.2 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 single 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.

3.3.3 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 perspon was the designated speaker throughout, and the feedback participants received included the right answer and what each player had selected. TODO confirm. Across both conditions in experiment 3, notifications were sent to the chat to indicate when a participant had made a selection.

3.4 Data pre-processing and exclusions

Participants could use the chat box freely, which meant that the chat transcript contained some non-referential language. The first author skimmed through the chat transcripts, tagging utterances that did not refer to the current tangram. These were primarily pleasantries ("Hello"), meta-commentary about how well or fast the task was going, and confirmations or denials ("ok", "got it", "yes", "no"). We exclude 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 playes in these 6* player games is at TODO! The realities of online recruitment and disconnection meant that the number of games varied, although we aimed for 20 games in each condition in experiments 1 and 2, and 40 per condition in experiment 3. We excluded incomplete blocks from analyses, but included complete blocks from partial games (See Table 1 for counts).

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.

3.5 Modelling strategy

TODO 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 (CITE) 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 formulae are reported in TODO supplement). Accuracy results used a logistic model, other results use linear models.

4 References

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*

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

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
Clark HH, Wilkes-Gibbs D (1986) Referring as a collaborative process. *Cognition*
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
Ginzburg J, Fernandez R (2005) Action at a distance: The difference between dialogue and multilogue. *Proceedings of DIALOR*:9
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*
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
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
Schober MF, Clark HH (1989) Understanding by addressees and overhearers. *Cognitive Psychology* **21**:211–232. doi:10.1016/0010-0285(89)90008-X
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
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 (2019) Audience Design in Multiparty Conversation. *Cogn Sci* **43**:e12774. doi:10.1111/cogs.12774