TODO title

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

[1]Stanford University

Abstract

This is an abstract in italics.
This is the second paragraph not in italics.

Keywords
One keyword; Yet another keyword

1 Introduction

Verbal communication is an integral part of our daily lives. We coordinate schedules with partners, socialize with friends over board games, learn and teach in seminar classes, and listen to podcasts. Communicative environments range in size from one-on-one dialogue to broadcast communication to large groups, but the goal of efficient communication is shared across these (Traum 2004, **branigan2006?**, **ginzburg2005?**). Shared referring expressions are a necessity for efficient communication; a thing or an idea needs some sort of name that the interlocutors will jointly understand. 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. But these dynamics may be different in larger groups, which are less studied. Our current work builds on the dyadic reference game tradition by extending it to larger groups.

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, **krauss-ConcurrentFeedbackConfirmation1966?**). 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 Hawkins et al. (2020), 83 pairs completed a similar iterated reference experiment where they communicated via a chat box. Speakers reduced their utterances, producing fewer words per image in later blocks than in earlier blocks, in line with results from face-to-face, oral paradigms.[1]

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

*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.

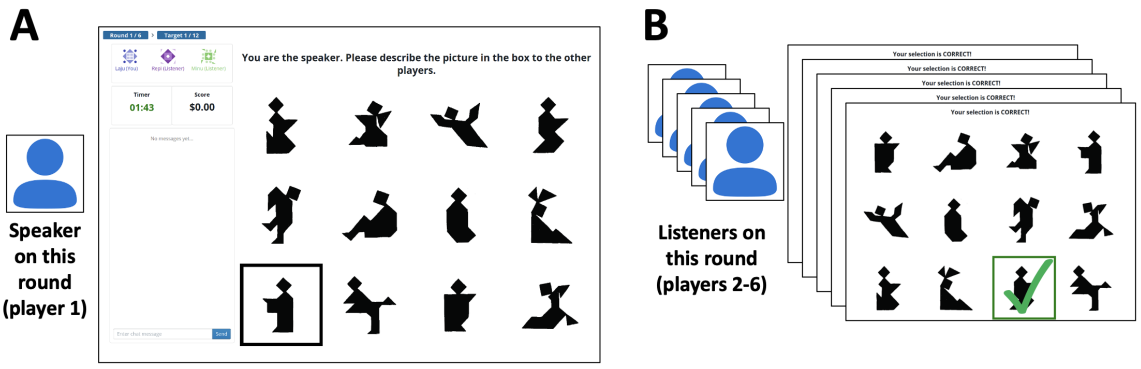

Figure 1: All participants saw all 12 tangram images. (A) Speaker's view during selection phase. (B) During the feedback stage, speakers saw what figure each person chose, but listeners only learned if their selection was correct or incorrect. Listeners were not shown what other listeners chose.

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 (**fox-tree2013?**, **horton2002?**, **horton2005?**). Prior work has focused on manipulating knowledge states by adding new listeners to established groups.

In this context, one approach for speakers is to 'aim low' and produce utterances tailored to the least knowledgeable listener (**yoonAimLowMechanisms2018a?**). For instance, in Yoon & Brown-Schmidt (2014), speakers developed conventions with one listener but then used longer descriptions with a new listener. Another strategy for speakers is to integrate across listeners and balance efficiency with informativeness by 'aiming in the middle'. In Yoon & Brown-Schmidt (2019), speakers communicating to a mixed group of 3 experienced listeners and 1 naive listener used shorter utterances and made fewer accommodations than they did in groups with a greater fraction of naive listeners. Both of these strategies predict that larger groups will be slower to converge than smaller groups.

Disagreements about how to conceptualize referents can also slow groups down. In (**weberCultural-ConflictMerger2003?**), pairs of participants played a reference game with the same image sets before a listener switched groups and joined a different pair, making a group of three. The addition of the new listener slowed both listeners down for multiple rounds. When a listener switched groups, they brought preconceptions about how the pictures should be described which conflicted with how the speaker was used to describing the images. This result predicts that, with more perspectives in play, larger groups may have more difficulty agreeing on common conceptualizations.

In general, listeners expect speakers to maintain conventions and stick to descriptions that were similar to successful descriptions. However, listeners were not surprised to hear different descriptions of a familiar object if it came from a new speaker who had just entered the room (Metzing & Brennan 2003). It's unclear what this finding predicts about new speakers who are present as fellow listeners during prior blocks – will listeners expect them to maintain conventions?

Work on multi-party communication has focused on the addition of a new person into a pair or group that had built up some shared representations. Our present work complements this prior work by examining the effect of group size during the process of convention formation. We extend the dyadic repeated reference game paradigm of Hawkins et al. (2020) to games for 2–6 players who rotate between speaker and listener roles. This paradigm allows us to confirm that these findings in dyads extend to larger groups: that accuracy and speed will increase across blocks (question 1) and that speakers will reduce their utterances (produce fewer words) in later blocks (question 2). Additionally, we will be able to test for trends across group size, allowing us to ask whether smaller groups use shorter utterances and reduce faster than larger groups (question 3) and how conventions emerge in larger groups (question 4). In sum, these analyses will fill a gap in the literature by providing a basic characterization of how convention-formation and communication occurs in larger groups.

Table 1: Summary of differences in experiments. Game size refers to the number of players per game. Speaker refers to whether there was one speaker the whole game or whether the speaker role rotated every block. Feedback is whether listeners saw only whether they were right or wrong or whether the additionally saw what other listeners had selected and what the correct answer was. Listener chat refers to whethers listeners could type freely in the chat or only communicate by pressing buttons to send four emojis to the chat. Continue games refers to whether games could continue (or start) with fewer than the requisite number of players; this was not intended to be a consequential manipulation, but was done to prevent games from ending if one player dropped out (an issue that was causing data loss in 6 player games).

Experiment	Game size	Speaker	Feedback	Listener chat	Continue games
1	2,3,4,5,6	rotating	self only	text	no
2a	6	one speaker	self only	text	no
2b	6	rotating	self, others, & correct	text	no
2c	6	rotating	self only	four emojis	no
3 thin	2,6	rotating	self only	four emojis	yes
3 thick	2,6	one speaker	self, others, & correct	text	yes

2 Experiment 1

2.1 Methods

Building on the methods of Hawkins et al. (2020), 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 speaker role rotated to another player and the process repeated with the same images. In total, there were 6 blocks, giving each player at least one chance to be the speaker. We recorded what participants said in the chat, as well as who selected what image and how long they took to make their selections.[2] We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.[3]

2.1.1 Participants

We recruited participants between May and July 2021 using the Prolific platform; participants had all self-reported as fluent native English speakers on Prolific's demographic prescreen. 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 (with the intention of a $10 hourly rate), in addition to up to $2.88 in performance bonuses. A total of 390 people each participated in one game.

2.1.2 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.

2.1.3 Procedure

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. Par-

[2] Code to run the experiment, as well as data and analysis code are available at https://osf.io/qdvbr/?view_only=47aebfde243f405e9c42a45cacb697d2.

[3] Our preregistrations are at https://osf.io/cn9f4/?view_only=7fdacd698b24465cb1a8699050af5bfc and https://osf.io/rpz67?view_only=5284203e2b644fc5ac39cf3e723b9a7e.

Table 2: TODO caption. 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
1	2	15	4
1	3	18	2
1	4	19	2
1	5	17	3
1	6	12	6
2a	6	15	3
2b	6	13	4
2c	6	10	6
3: thin	2	35	3
3: thin	6*	44	0
3: thick	2	39	3
3: thick	6*	38	2

ticipants 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. This feedback regime is different from Hawkins et al. (2020) where listeners were shown what the right answer was during feedback. We made this change to prevent listeners from learning conventions purely as a memorized mapping between utterance and correct answer.

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.

2.1.4 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.

Our intended sample size was 20 complete games in each group size, but we ended up with fewer due to games not filling or participants disconnecting early (Table **??**). We excluded incomplete blocks from analyses, but included complete blocks from partial games.

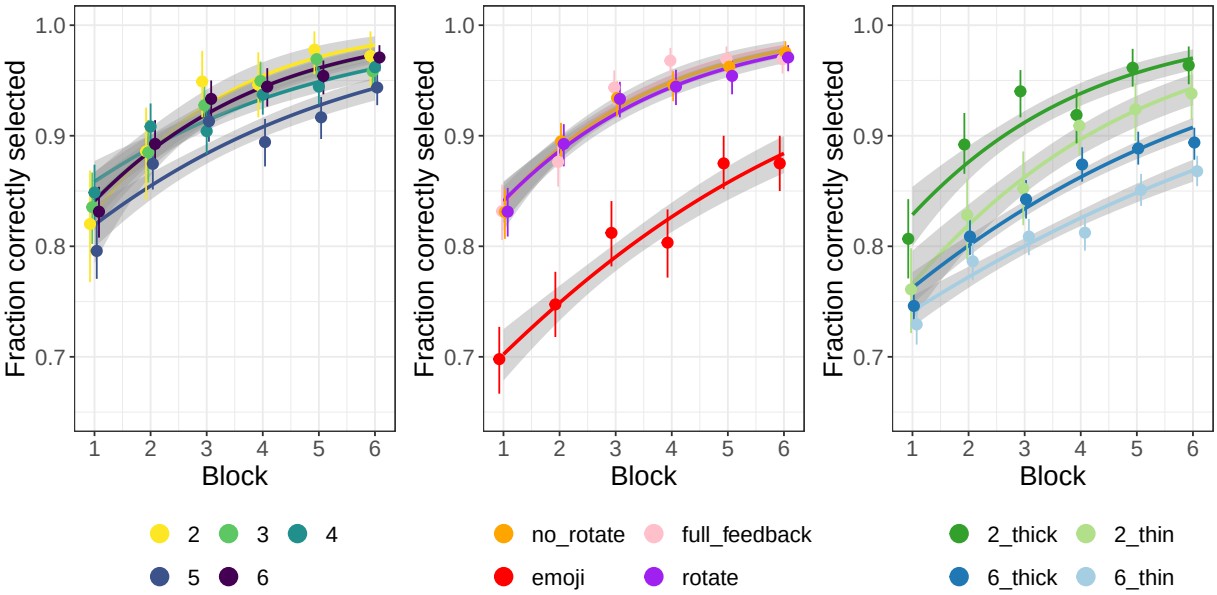

Figure 2: TODO

2.2 Results

2.2.1 Reduction

Our second question was whether speakers reduce their referring expressions in larger groups.

2.2.2 Accuracy

Our first question was whether accuracy and speed increased across groups of different sizes.

2.2.2.1 Accuracy is high and increasing. Most individuals were accurate in their selections, with accuracy rising across blocks (Figure 2.2.2). In a logistic model of accuracy[4], participants are more accurate in later blocks (block: Est=0.38, CrI=[0.25, 0.5]), and there was no strong effect of group size on accuracy (numPlayers: Est=-0.02, CrI=[-0.08, 0.03]) or interaction between block and group size (block:numPlayers: Est=-0.01, CrI=[-0.04, 0.01]).

2.2.2.2 Participants speed up in later blocks. Participants selected images faster in later blocks (Figure **??**), although there was wide variability. In a linear model of selection time[5], participants got faster across blocks (block: Est=-10.03, CrI=[-11.03, -9.03]) and were slightly slower in larger games (numPlayers: Est=1.03, CrI=[0.4, 1.66]). This speed up is consistent with prior work by (**weberCulturalConflictMerger2003?**) which used speed as the dependent measure. Wide variability in selection time meant that especially for larger groups, there was a wide spread in how long it took groups to complete the experiment.

2.2.3 Reduction

2.2.3.1 Speakers' utterances reduce in length. As shown in Figure **??**, the number of words produced by speakers decreases over the course of rounds, both in aggregate and for many individual groups. Nonetheless, in some groups, a later speaker may be more verbose than an earlier speaker. Speakers make longer utterances in early blocks that reduce to shorter utterances in later blocks. From

[4] correct.num∼ block × numPlayers This and all subsequent regression models were run in brms with weakly regularizing priors.

[5] time∼ block × numPlayers

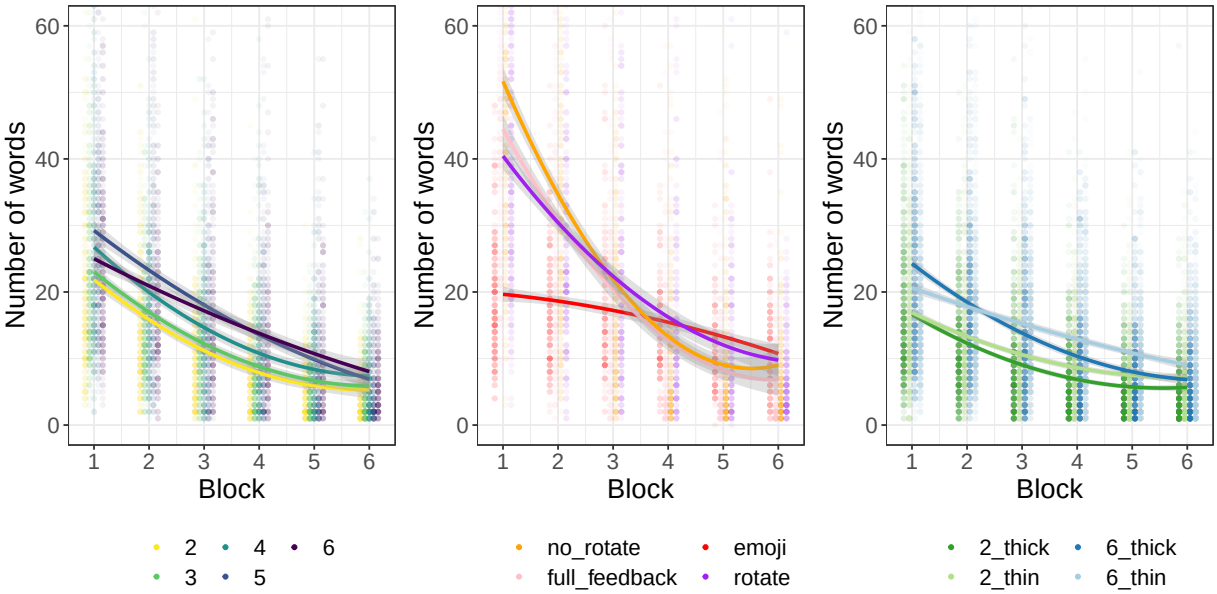

Figure 3: TODO

a linear model[6], the effect of being one block later is -3.35 (CrI=[-4.58, -2.13]) words.

2.2.3.2 Listeners rarely talk. Listeners often don't talk much, but are more likely to ask questions or make clarification in early blocks. In a linear regression for the number of words each listener said [7], there was an effect of block (block: Est=-0.48, CrI=[-0.79, -0.18]), but no clear effect of game size (numPlayers: Est=0.2, CrI=[-0.13, 0.51]).

2.2.4 Effects of group size on conventions

Our third question was whether smaller groups would use fewer words or reduce faster than larger groups.

2.2.4.1 Larger groups say more. The overall effect of having more players in a group is 1.67 (CrI=[0.68, 2.71]) words from the speaker per trial per additional player. There is no clear interaction between block and group size (block:numPlayers: Est=-0.1, CrI=[-0.39, 0.18]). Larger groups saying more is consistent with predictions from audience design that with more listeners to accommodate, the speaker may use multiple conceptualizations, either initially as a hedge or in response to listener clarifications.

2.2.4.2 Speaker experience does not fully explain group size effects. One potential concern is that group size correlates with whether the speaker has had the speaker role before (smaller groups repeat speakers more). To address this confound, we coded for whether the speaker has been speaker in an earlier block[8]. Repeat speakers do use fewer words (speaker.repeat: Est=-8.55, CrI=[-10.41, -6.79]), but there are still effects of group size (numPlayers: Est=1.63, CrI=[0.58, 2.66]) and block (block: Est=-5.26, CrI=[-6.84, -3.69]). The effects of block and repeat speaker are subadditive (block:speaker.repeat: Est=3.2, CrI=[2.65, 3.78]), and there is minimal interaction between block and group size (block:numPlayers: Est=0.08, CrI=[-0.22, 0.41]).

[6] words∼ block × numPlayers + (block|tangram) + (1|playerId) + (1|tangram_group) + (block|gameId)
[7] words∼ block × numPlayers + (block|tangram) + (1|playerId) + (1|tangram_group) + (block|gameId)
[8] words∼ block × numPlayers + block × speaker.repeat + (block|tangram) + (1|playerId) + (1|tangram_group) + (block|gameId)

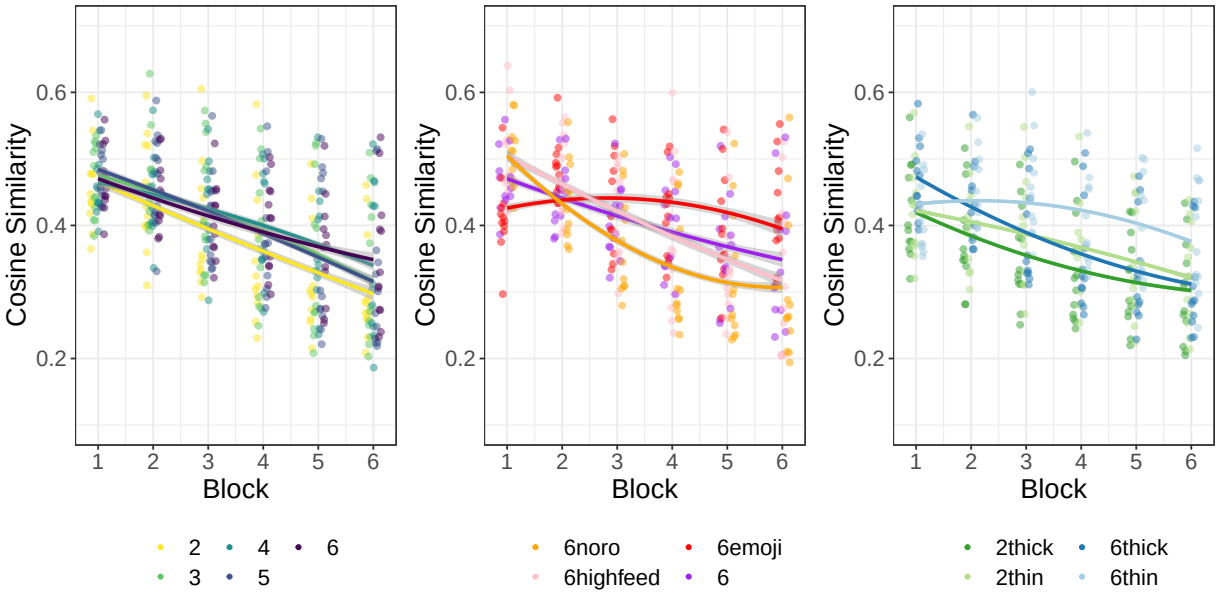

Figure 4: TODO

3 reduction

3.0.0.1 Speakers who don't know the convention reduce less. In our games (which had limited feedback), listeners who got a tangram wrong didn't have a way of knowing what the right answer was unless they asked for clarification in the chat. If a speaker got a tangram wrong as a listener in the previous block, they may not have known the conventional description that went with it, and thus were unlikely to follow the convention. If we assume that reduction is a sign of convention development, then speakers should say more words when they got the tangram wrong the previous block. We added prior errors as an additional predictor to our regression predicting number of words and found that speakers said more words for tangrams after they were incorrect (numPlayers: Est=2.18, CrI=[0.93, 3.44]).

3.0.1 Divergence

3.0.2 Convergence

3.0.2.1 Groups varied in their strategies and reduction. While most groups did form conventions for most tangrams, it's illustrative to look at a case where a group did not. Table 3 shows the transcript of a 4-person group for a specific figure where they described it geometrically every round, leading to long and not very informative descriptions. Nearly all the figures have diamond heads, so this isn't a distinguishing feature, yet it is described. This illustrates the variability between groups, but also why conventions might be useful.

Table 3: Excerpt from a group that did not reduce very much. The speaker for each round is marked with (S). Figure under discussion is row 3, column 3 in Figure 1A.

Block	Person	Text
1	A(S)	Diamond on top. Body with no real arms or legs. The body is shaped like a boot with the diamond on top.
	C	Is the boot pointed left or right?
2	B(S)	diamond on top, large body beneath it. Left is a straight line all the way down, small variations on the right to the main body
3	C(S)	Diamond in center on top. Left side straight, right side carved out like a vase.
4	D(S)	Diamond head, flat topped body, straight on the left side with two triangles pointing out on the left
	D(S)	*on the right
5	A(S)	Diamond on top. Left side is straight, right side is obstructed, looks like a boot
	B	what do you mean by obstructed?
	A(S)	The left side of the body is right, right side has bents in it
6	B(S)	Diamond on top of a long large body/rectangle. Left side is complete, right side has bits missing

A different 4-person group had a member who during the first block shared the idea that the task would be easier if they explicitly gave "codenames" to the figures. The transcript for this group and one of the tangrams is shown in Table **??**. Of note, multiple speakers forget the assigned codename, demonstrating that meta-knowledge doesn't always help. This group also describes the figure in relation to another already-named figured. Nonetheless, the group successfully conventionalizes on a couple reduced names for this figure: "zigzag" and "beggar". This dual-naming of figures from multiple conceptual angles contributed by different speakers also occurs in other games.

3.1 Interim Discussion

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 &

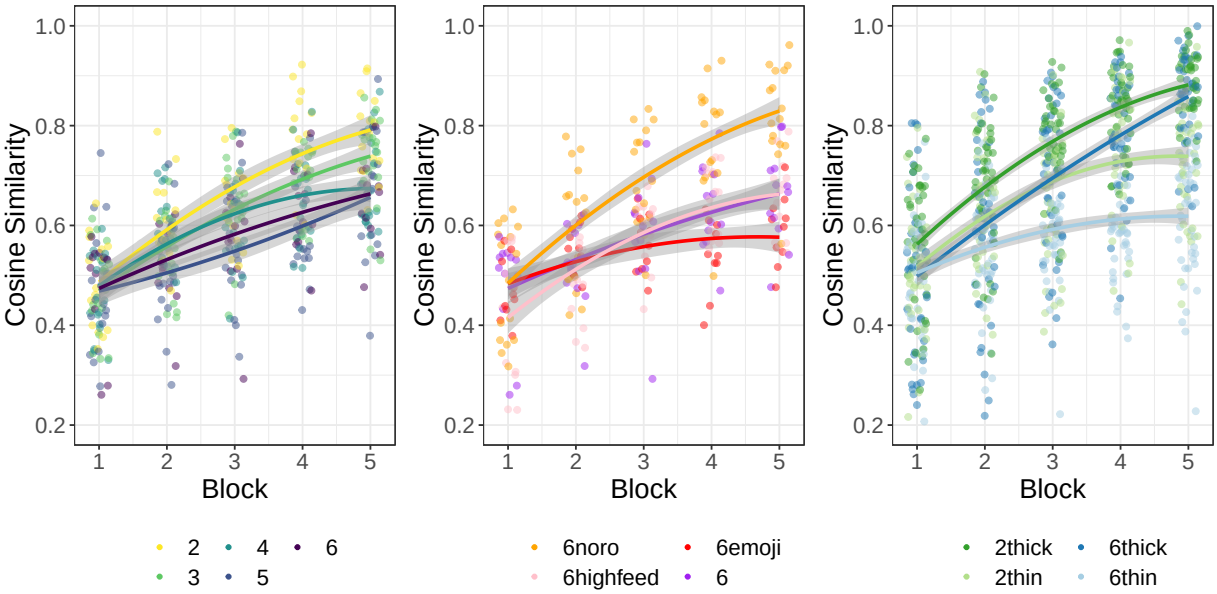

Figure 5: TODO

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 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.

4 Experiment 2

Motivation

5

5.1 Methods

5.1.1 Participants

All participants were self-reported fluent native English speakers who had not participated in experiment 1. All experiment 2 games were 6-player games. Experiment 2a was run in MONTH; participants were paid MONEY for 6-player games, with the speaker getting a bonus MONEY, in addition to up to $2.88 in performance bonuses for each player. Experiment 2b was run in MONTH; participants were paid MONEY for 6-player games.

In all games, participants could additionally make up to We recruited participants between May and July 2021 using the Prolific platform; participants had all self-reported as fluent native English speakers on Prolific's demographic prescreen. 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 (with the intention of a $10 hourly rate), in addition to up to $2.88 in performance bonuses. A total of 390 people each participated in one game.

5.1.2 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.

5.1.3 Procedure

Experiment 2 consisted of three different variations on Experiment 1, so we describe the differences from the Experiment 1 procedure.

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. This feedback regime is different from Hawkins et al. (2020) where listeners were shown what the right answer was during feedback. We made this change to prevent listeners from learning conventions purely as a memorized mapping between utterance and correct answer.

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.

5.2 Results

5.2.1 Accuracy

5.2.2 Reduction

5.2.3 Divergence

5.2.4 Convergence

5.2.5 Emoji usage ?

5.3 Interim discussion

6 Experiment 3

Motivation

6.1 Methods

6.2 Results

7 General Discussion

7.1 Limitations

8 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*

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

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*

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

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

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