Supplement to "Interaction structure constrains the emergence of conventions in group communication"

Contents

Number of games **1**

More on listener utterances **3**

Additional measure of convergence **5**

Distinctiveness of tangrams **6**

Summaries of model outputs **7**

Accuracy models **7**

Reduction models **8**
Primary reduction model 8
Extra reduction model 9
Listener reduction models 9
Initial utterance reduction model 10

Linguistic content models **11**
Convergence within games: comparison to last round 11
Divergence across games 12
Divergence across tangrams 13
Convergence to next 15
Divergence from first 16

Number of games

In experiment 3, the 6* player games did not all have 6 players, both because games continued as participants dropped out and because if there weren't enough players after 5 minutes of waiting, the game would start with whoever was there. All analyses use "intent to treat" and call these 6 player games.

The number of games goes up in some cases because only complete blocks (where the speaker said something every trial) are analysed. If there was initial confusion and a speaker missed a trial, that block was excluded.

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: consistent 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

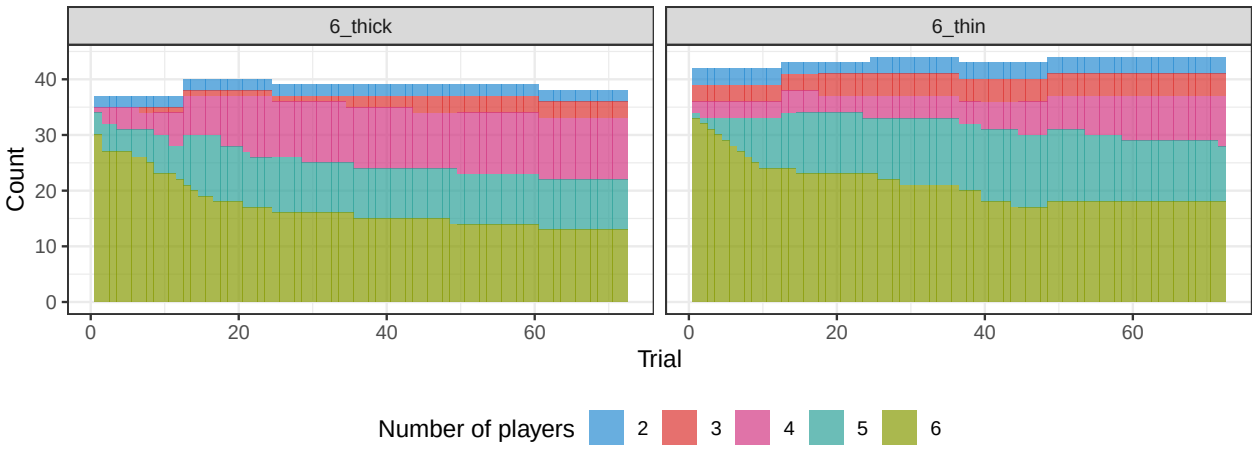

Figure 1: Number of players during 6 thin and 6 thick games in experiment 3. Blocks that were incomplete were excluded, so if a speaker said nothing during a trial, that block was excluded.

More on listener utterances

Listeners' use of backchannel declined over the course of the game. The use of emoji in the thin games is not directly comparable to listener language use in thick games, since some emoji usage (such as the green checkmark) are most likely equivalent to non-referential listener language ("got it" etc.) that was excluded. The higher rate of emoji use versus referential language thus could be due to its non-equivalence, a lower level of accuracy in thin games, or listeners having a lower threshold for sending emojis compared to writing out clarifications.

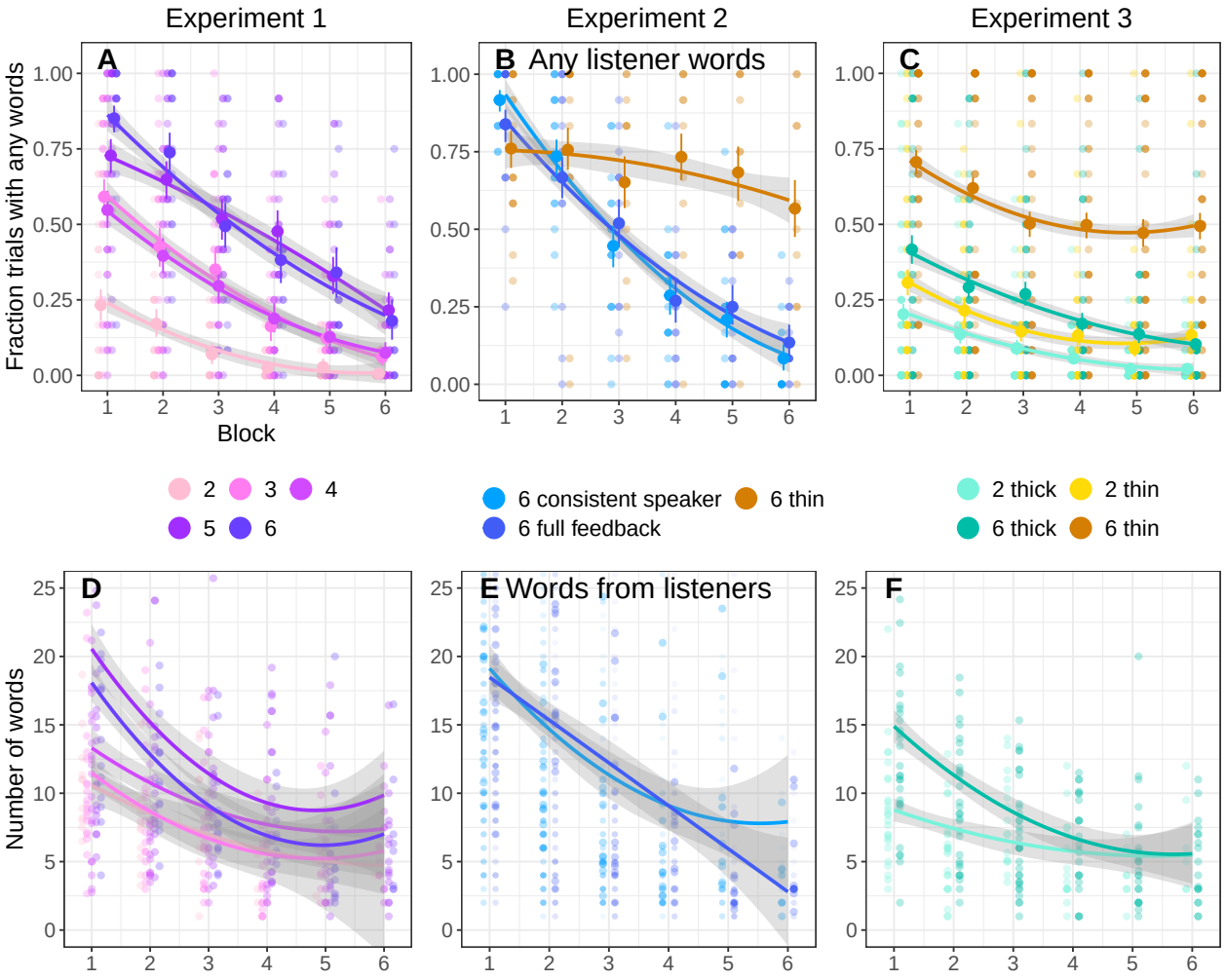

Figure 2: Listener contributions. A-C: Fraction of trials where any listener said anything that was referential. Dots are per game averages. Smooths are binomial fit lines. D-F: On trials where at least one listener contributed, the number of words of referential language produced by listeners. Dots are per game averages. Smooths are quadratic fit lines. Y-axis is truncated, and a few outliers points are not visible.

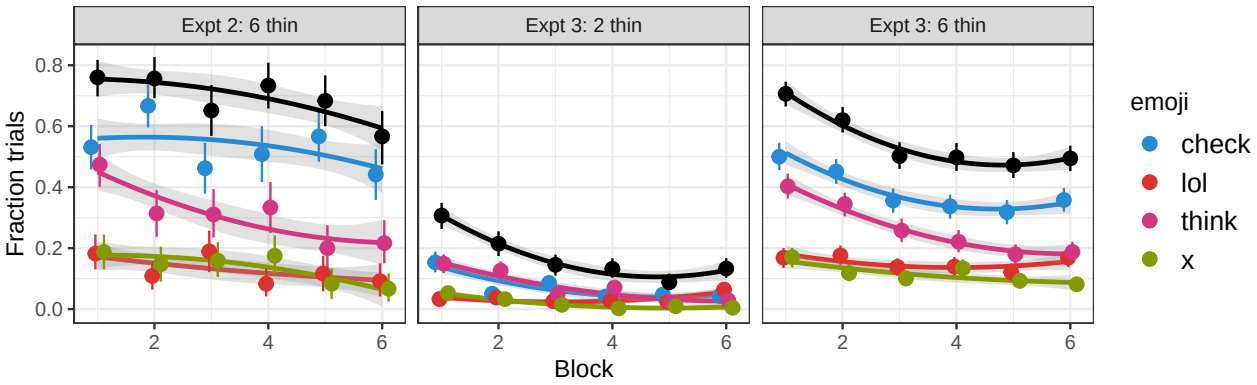

Figure 3: Fraction of trials on which at least one listener produced the labelled emoji. Fraction of trials when any emoji was produced are shown in black. Dots are per condition, per block estimates with 95% bootstrapped CIs. Smooths are binomial fit lines.

We note a deviation from the pre-registration here in the analysis of the emojis. In the pre-registration we said we would "analyse the distribution of emoji's produced as a function of block and its relation to accuracy and speaker utterance length." We did not do this beyond the visualization shown here.

Additional measure of convergence

The main text included the graph for convergence comparing utterances from blocks 1-5 to the utterance from block 6. Here we show two other measures of semantic shifts for descriptions for the same tangram in the same game: similarity to the first utterance and similarity to the next utterance.

Similarity to the first utterance is not very informative (but we pre-registered it). Similarity to the next utterance is what actually drives the convergence phenomena: pairs of utterances from adjacent blocks become closer together over time.

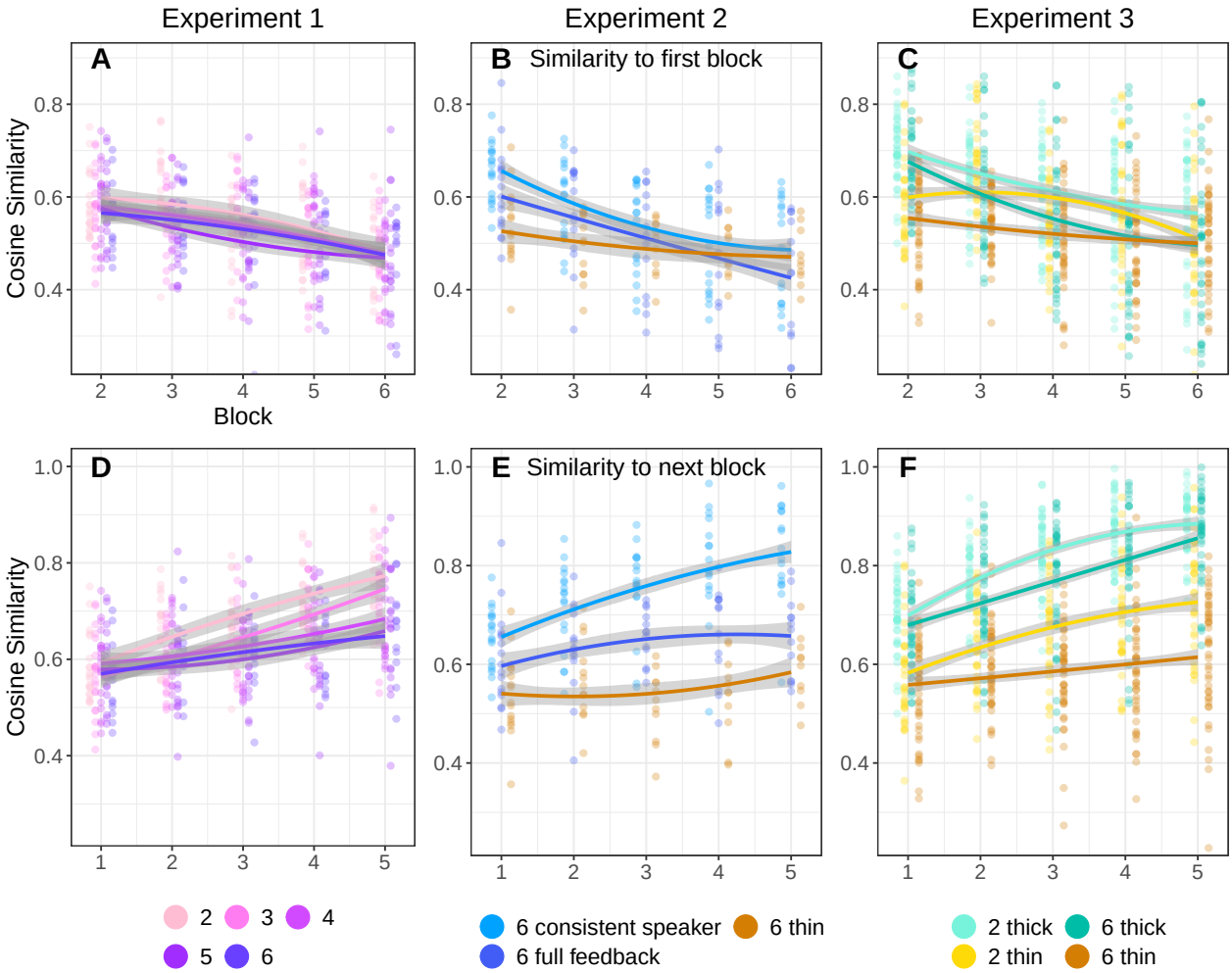

Figure 4: Additional measures of convergence and divergence. A-C is the similarity between utterances on a given block to the first block utterance for the same image, in the same game. Dots are per-game averages, smooths are quadratic. D-F is the similarity between utterances on a given block to the corresponding utterances in the next block. Dots are per-game averages, smooths are quadratic.

Distinctiveness of tangrams

An additional measure of convergence/divergence patterns is how different tangrams get described in the same game – as nicknames evolve, different tangrams get more different descriptions.

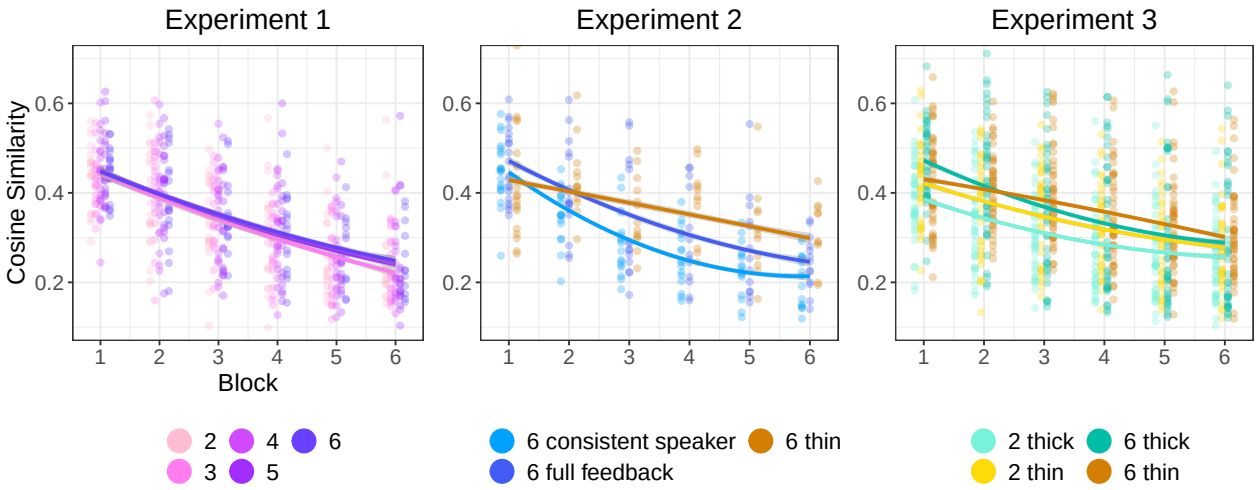

Figure 5: Divergence in descriptions of different tangrams. Cosine similarity between the descriptions of two different tangrams in the same block and group are shown. Dots are per-game averages, smooths are quadratic.

Summaries of model outputs

The following sections contain model outputs. All models were run using BRMS. We report the priors and pre-registration status for each group of models. Tables provide the individual model formulae and the point estimates and 95% credible intervals for the fixed effects.

Note that for all models, block was 0 indexed, so intercepts are what happened during the first block.

Accuracy models

Accuracy models were all run as logistic models with normal(0,1) priors for both betas and sd. This model was not explicitly included in the experiment 1 and 2 pre-registrations; it was included with more ambitious mixed effects (which did not run in a timely manner) in the experiment 3 pre-registration.

Table 2: Experiment 1 logistic model of listener accuracy: correct.num ∼ block × numPlayers + (1|gameId)

Term	Est.	95% CrI
block	0.44	[0.31, 0.58]
block:numPlayers	-0.02	[-0.05, 0.01]
Intercept	2.10	[1.57, 2.65]
numPlayers	-0.07	[-0.2, 0.05]

Table 3: Experiment 2: 6 consistent speaker logistic model of listener accuracy: correct.num ∼ block + (1|gameId)

Term	Est.	95% CrI
block	0.45	[0.39, 0.52]
Intercept	1.78	[1.4, 2.19]

Table 4: Experiment 2: 6 full feedback logistic model of listener accuracy: correct.num ∼ block + (1|gameId)

Term	Est.	95% CrI
block	0.47	[0.39, 0.54]
Intercept	1.35	[0.59, 2.06]

Table 5: Experiment 2: 6 thin logistic model of listener accuracy: correct.num ∼ block + (1|gameId)

Term	Est.	95% CrI
block	0.23	[0.19, 0.28]
Intercept	0.88	[0.64, 1.12]

Table 6: Experiment 3 logistic model of listener accuracy: correct.num ~ block × gameSize × channel + (1|gameId)

Term	Est.	95% CrI
block	0.41	[0.32, 0.5]
block:channelthin	-0.07	[-0.18, 0.04]
block:gameSize6	-0.34	[-0.43, -0.25]
block:gameSize6:channelthin	0.07	[-0.05, 0.19]
channelthin	-0.36	[-0.78, 0.05]
gameSize6	-0.64	[-1.05, -0.25]
gameSize6:channelthin	0.31	[-0.22, 0.87]
Intercept	1.69	[1.39, 1.99]

Reduction models

Primary reduction model

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). This model was pre-registered for each experiment and run with the mixed effects structure as pre-specified.

Table 7: Experiment 1: words ~ block × numPlayers + (block|tangram) + (1|playerId) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
block	-3.36	[-4.56, -2.18]
block:numPlayers	-0.09	[-0.37, 0.18]
Intercept	16.87	[11.63, 21.89]
numPlayers	1.60	[0.62, 2.6]

Table 8: Experiment 2: 6 consistent speaker: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
block	-5.31	[-6.35, -4.3]
Intercept	29.65	[24.82, 34.49]

Table 9: Experiment 2: 6 full feedback: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
block	-4.64	[-5.81, -3.53]
Intercept	25.79	[20.97, 30.29]

Table 10: Experiment 2: 6 thin: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
block	-2.1	[-3.37, -1.12]
Intercept	20.3	[17.37, 23.53]

Table 11: Experiment 3: words ~ block × channel × gameSize + (block × channel × gameSize|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
block	-2.24	[-2.92, -1.57]
block:channelthin	0.29	[-0.56, 1.23]
block:channelthin:gameSize6	0.64	[-0.59, 1.81]
block:gameSize6	-1.22	[-2.06, -0.29]
channelthin	0.80	[-2.85, 4.26]
channelthin:gameSize6	-2.21	[-7.16, 3.08]
gameSize6	7.51	[3.63, 11.3]
Intercept	14.74	[11.68, 17.72]

Extra reduction model

For experiment 1, we also pre-specified a model about whether the speaker's correctness on the prior block (when they were a listener) had an effect on how many words of description they produced. Priors were the same as for primary reduction model.

Table 12: Experiment 1: words ~ block × numPlayers + block × wasINcorrect + (block|tangram) + (1|playerId) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
block	-2.17	[-3.39, -1]
block:numPlayers	-0.22	[-0.5, 0.06]
block:wasINcorrect	0.24	[-0.24, 0.72]
Intercept	12.16	[6.48, 18.07]
numPlayers	2.09	[0.88, 3.3]
wasINcorrect	3.07	[1.67, 4.45]

Listener reduction models

These models were not pre-registered.

For the model of how often any listener talked, the priors were normal(0,1) for both beta and sd.

For the model of how much was said on trials when listeners talked, the priors were the same as for the primary (speaker) reduction model.

Table 13: Experiment 1: words ~ block × numPlayers + (block|gameId)

Term	Est.	95% CrI
block	-0.17	[-1.53, 1.3]
block:numPlayers	-0.41	[-0.72, -0.11]
Intercept	4.72	[0.09, 9.44]
numPlayers	2.07	[1, 3.12]

Table 14: Experiment 1: is.words ~ block × numPlayers + (1|gameId)

Term	Est.	95% CrI
block	-0.80	[-0.97, -0.62]
block:numPlayers	0.03	[-0.01, 0.07]
Intercept	-2.67	[-3.54, -1.79]
numPlayers	0.79	[0.58, 0.98]

Initial utterance reduction model

These models were not pre-registered. They looked at speaker reduction only on words that were produced prior to the first listener message each trial. These models were only run on experimental conditions where listeners could contribute textual responses.

Reduction models were run as linear models with the same priors as the primary reduction model.

Table 15: Experiment 1: words ~ block × numPlayers + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	18.66	[14.58, 22.71]
block	-3.56	[-4.54, -2.55]
block:numPlayers	0.27	[0.03, 0.5]
numPlayers	-0.33	[-1.14, 0.53]

Table 16: Experiment 2: 6 consistent speaker: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	18.06	[14.76, 21.44]
block	-2.49	[-3.19, -1.79]

Table 17: Experiment 2: 6 full feedback: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	16.69	[13.41, 20.02]
block	-2.49	[-3.34, -1.62]

Table 18: Experiment 3: words ~ block × gameSize + (block × gameSize|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	13.88	[11.62, 16.2]
block	-2.08	[-2.66, -1.47]
block:gameSize6	-0.65	[-1.43, 0.14]
gameSize6	5.13	[2, 7.95]

Linguistic content models

We ran a number of models predicting the cosine similarity between pairs of S-BERT embeddings of utterances. For all of these models, we used linear models with the priors normal(.5,.2) for intercept, normal(0,.1) for beta, and normal(0,.05) for sd.

These models were verbally described (but not formally specified) in the pre-registrations for experiment 2 in the full feedback and thin conditions and for experiment 3, for looking at divergence between games, convergence within games (compared to first block, next block, and last block utterances), and divergence between tangrams within games.

Convergence within games: comparison to last round

This is the primary convergence metric presented in the main paper.

Table 19: Experiment 1: sim ~ earlier × condition + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
condition	-0.008	[-0.021, 0.005]
earlier	0.089	[0.076, 0.102]
earlier:condition	-0.008	[-0.011, -0.005]
Intercept	0.517	[0.458, 0.573]

Table 20: Experiment 2: 6 consistent speaker: sim ~ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
earlier	0.086	[0.078, 0.094]
Intercept	0.499	[0.444, 0.556]

Table 21: Experiment 2: 6 full feedback: sim ~ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
earlier	0.062	[0.051, 0.072]
Intercept	0.438	[0.389, 0.487]

Table 22: Experiment 2: 6 thin: sim ~ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
earlier	0.023	[0.013, 0.033]
Intercept	0.498	[0.453, 0.54]

Table 23: Experiment 3: sim ~ earlier × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
channelthin	-0.034	[-0.08, 0.011]
channelthin:gameSize6	0.039	[-0.021, 0.097]
earlier	0.080	[0.074, 0.086]
earlier:channelthin	-0.025	[-0.033, -0.017]
earlier:channelthin:gameSize6	-0.035	[-0.047, -0.025]
earlier:gameSize6	0.009	[0.001, 0.017]
gameSize6	-0.069	[-0.113, -0.025]
Intercept	0.581	[0.542, 0.62]

Divergence across games

This is the divergence metric presented in the paper.

Table 24: Experiment 1: sim ~ block × condition + (1|tangram)

Term	Est.	95% CrI
block	-0.035	[-0.038, -0.032]
block:condition	0.001	[0.001, 0.002]
condition	0.002	[0, 0.004]
Intercept	0.468	[0.429, 0.507]

Table 25: Experiment 2: 6 consistent speaker: sim ~ block + (1|tangram)

Term	Est.	95% CrI
block	-0.041	[-0.043, -0.039]
Intercept	0.484	[0.442, 0.526]

Table 26: Experiment 2: 6 full feedback: sim ∼ block + (1|tangram)

Term	Est.	95% CrI
block	-0.038	[-0.04, -0.035]
Intercept	0.502	[0.46, 0.546]

Table 27: Experiment 2: 6 thin: sim ∼ block + (1|tangram)

Term	Est.	95% CrI
block	-0.004	[-0.006, -0.001]
Intercept	0.434	[0.406, 0.465]

Table 28: Experiment 3: sim ∼ block × channel × gameSize + (1|tangram)

Term	Est.	95% CrI
block	-0.024	[-0.025, -0.023]
block:channelthin	0.004	[0.002, 0.005]
block:channelthin:gameSize6	0.017	[0.015, 0.019]
block:gameSize6	-0.008	[-0.01, -0.007]
channelthin	0.014	[0.01, 0.018]
channelthin:gameSize6	-0.030	[-0.035, -0.024]
gameSize6	0.051	[0.047, 0.055]
Intercept	0.411	[0.368, 0.453]

Divergence across tangrams

This is an additional metric comparing the similiarities between descriptions for different tangrams within a game. It measures how distinct the descriptions for different tangram images are.

Table 29: Experiment 1: sim ∼ block × condition + (1|gameId)

Term	Est.	95% CrI
block	-0.043	[-0.046, -0.039]
block:condition	0.000	[-0.001, 0.001]
condition	0.003	[-0.008, 0.014]
Intercept	0.429	[0.382, 0.473]

Table 30: Experiment 2: 6 consistent speaker: sim ∼ block + (1|gameId)

Term	Est.	95% CrI
block	-0.046	[-0.048, -0.044]
Intercept	0.416	[0.389, 0.443]

Table 31: Experiment 2: 6 full feedback: sim ~ block + (1|gameId)

Term	Est.	95% CrI
block	-0.047	[-0.049, -0.044]
Intercept	0.459	[0.422, 0.496]

Table 32: Experiment 2: 6 thin: sim ~ block + (1|gameId)

Term	Est.	95% CrI
block	-0.025	[-0.028, -0.022]
Intercept	0.432	[0.393, 0.471]

Table 33: Experiment 3: sim ~ block × channel × gameSize + (1|gameId)

Term	Est.	95% CrI
block	-0.027	[-0.029, -0.025]
block:channelthin	-0.001	[-0.003, 0.002]
block:channelthin:gameSize6	0.011	[0.008, 0.015]
block:gameSize6	-0.010	[-0.013, -0.008]
channelthin	0.038	[-0.001, 0.082]
channelthin:gameSize6	-0.053	[-0.115, 0]
gameSize6	0.073	[0.035, 0.113]
Intercept	0.378	[0.352, 0.404]

Convergence to next

We also looked at how similar an utterance was to the next block utterance for the same image in the same group: this can be thought of as the derivative of the to-last comparison. (Although cosine similarities are not actually additive in the same way integrals are).

Table 34: Experiment 1: sim ∼ earlier × condition + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
condition	-0.004	[-0.014, 0.006]
earlier	0.063	[0.051, 0.075]
earlier:condition	-0.008	[-0.011, -0.006]
Intercept	0.591	[0.541, 0.641]

Table 35: Experiment 2: 6 consistent speaker: sim ∼ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
earlier	0.043	[0.037, 0.05]
Intercept	0.660	[0.619, 0.702]

Table 36: Experiment 2: 6 full feedback: sim ∼ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
earlier	0.015	[0.006, 0.024]
Intercept	0.605	[0.569, 0.643]

Table 37: Experiment 2: 6 thin: sim ∼ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
earlier	0.010	[0, 0.019]
Intercept	0.533	[0.49, 0.578]

Table 38: Experiment 3: sim ∼ earlier × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
channelthin	-0.124	[-0.159, -0.088]
channelthin:gameSize6	0.000	[-0.051, 0.049]
earlier	0.046	[0.041, 0.052]
earlier:channelthin	-0.010	[-0.018, -0.002]
earlier:channelthin:gameSize6	-0.018	[-0.029, -0.007]
earlier:gameSize6	-0.003	[-0.011, 0.004]
gameSize6	-0.034	[-0.069, 0.003]
Intercept	0.714	[0.682, 0.746]

Divergence from first

We also looked at how similar an utterance was to the first block utterance for the same image. This is not very informative because first round utterances tend to be pretty noisy with lots of hedges and filler words.

Table 39: Experiment 1: sim ~ later × condition + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
condition	-0.010	[-0.022, 0.003]
Intercept	0.647	[0.591, 0.705]
later	-0.030	[-0.041, -0.019]
later:condition	0.001	[-0.002, 0.004]

Table 40: Experiment 2: 6 consistent speaker: sim ~ later + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.680	[0.628, 0.728]
later	-0.042	[-0.049, -0.035]

Table 41: Experiment 2: 6 full feedback: sim ~ later + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.644	[0.584, 0.706]
later	-0.044	[-0.052, -0.037]

Table 42: Experiment 2: 6 thin: sim ~ later + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.537	[0.49, 0.584]
later	-0.014	[-0.023, -0.004]

Table 43: Experiment 3: sim ~ later × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
channelthin	-0.076	[-0.123, -0.026]
channelthin:gameSize6	-0.062	[-0.127, 0.001]
gameSize6	-0.017	[-0.062, 0.03]
Intercept	0.721	[0.681, 0.76]
later	-0.034	[-0.039, -0.028]
later:channelthin	0.011	[0.003, 0.019]
later:channelthin:gameSize6	0.021	[0.01, 0.032]
later:gameSize6	-0.011	[-0.019, -0.004]