Supplement

2023-05-05

Contents

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: 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

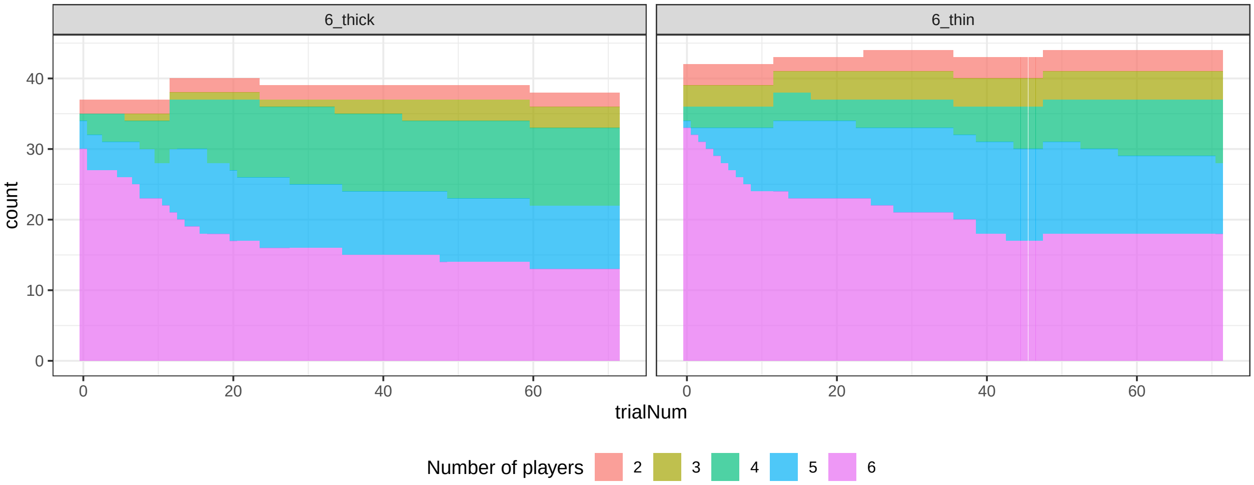

Figure 1: Number of players during 6 thin and 6 thick games in experiment 3.

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 it's non-equivalence, a lower level of accuracy in thin games, or emojis being a lower threshold for sending than written out questions.

When listeners did provide descriptions for references, the amount of language used was greater in early trial than in later trials.

In thin games, the rate at which listeners use specific emojis, or any emoji at all declines over time.

TODO Written about 6thin in experiment 2 and for 2 and 6 thin in 3 Additionally, exclusive to this condition, we will analyse the distribution of emoji's produced as a function of block and its relation to accuracy and speaker utterance length.

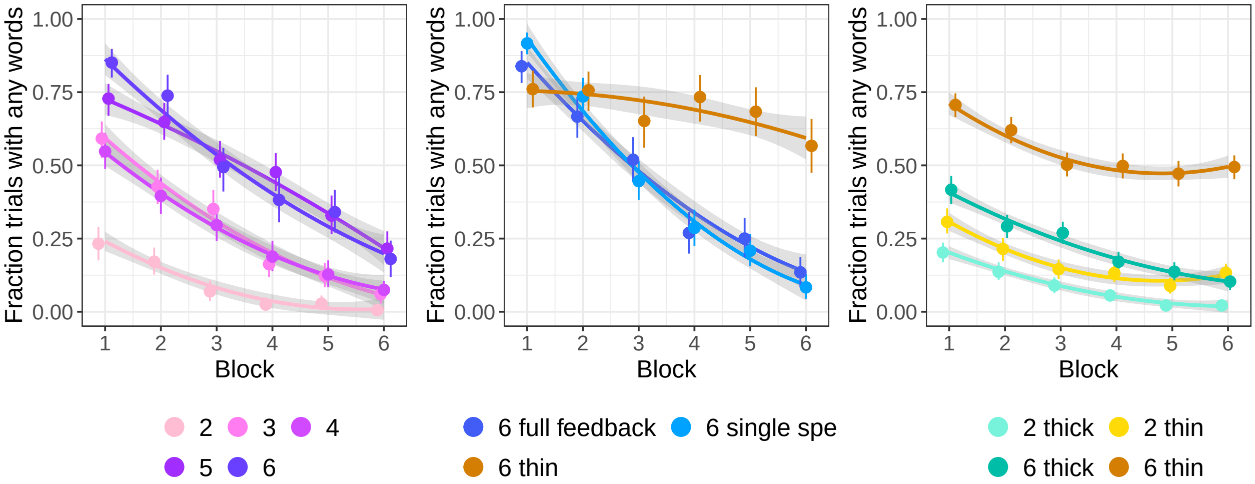

Figure 2: Fraction of trials when any reference language (or emoji) was produced by any listener.

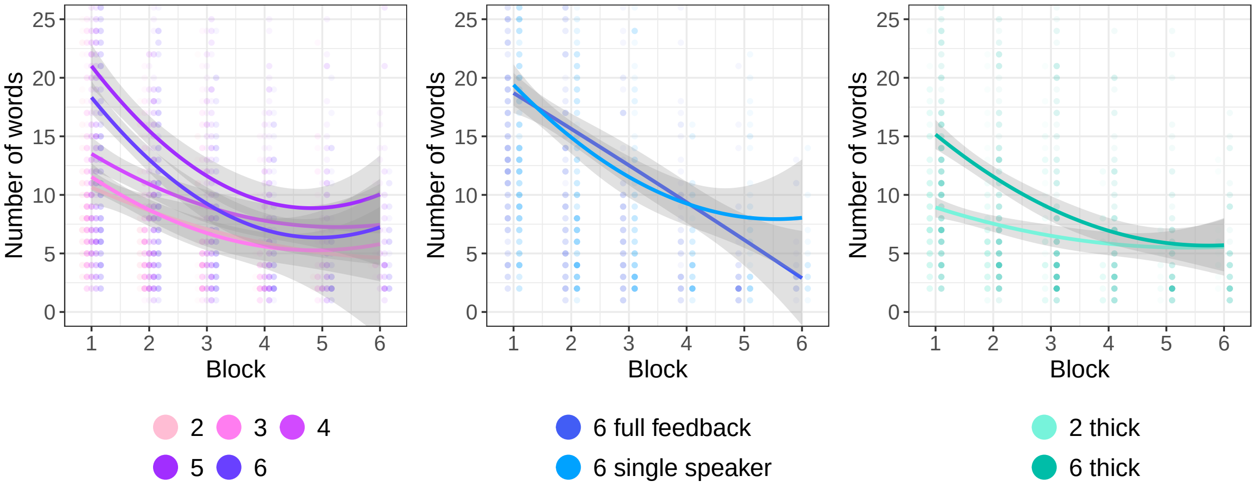

Figure 3: Number of words of referential language produced by listeners over time. Excludes trials where no listeners contributed descriptive language.

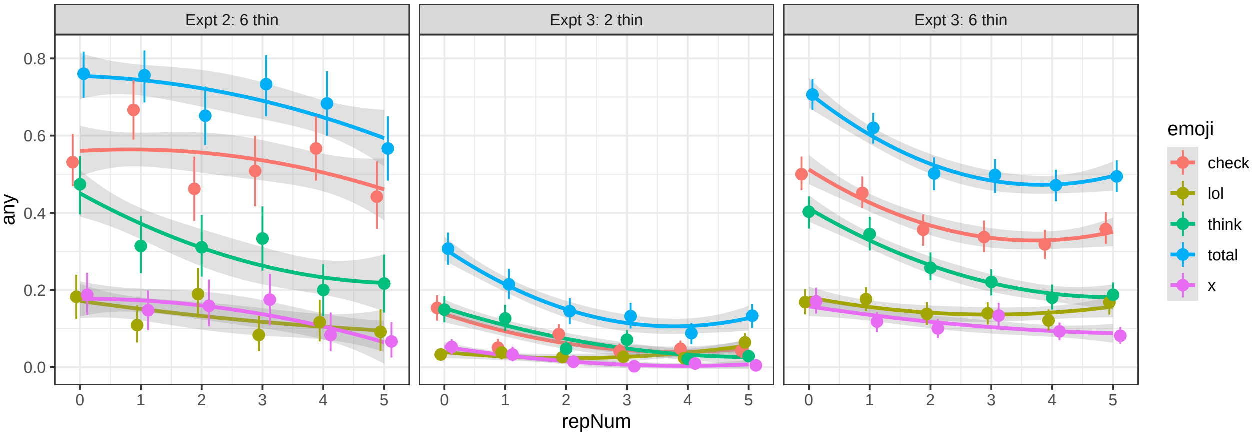

Figure 4: Fraction of trials on which at least one listener produced the labelled emoji (or any emoji.

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.

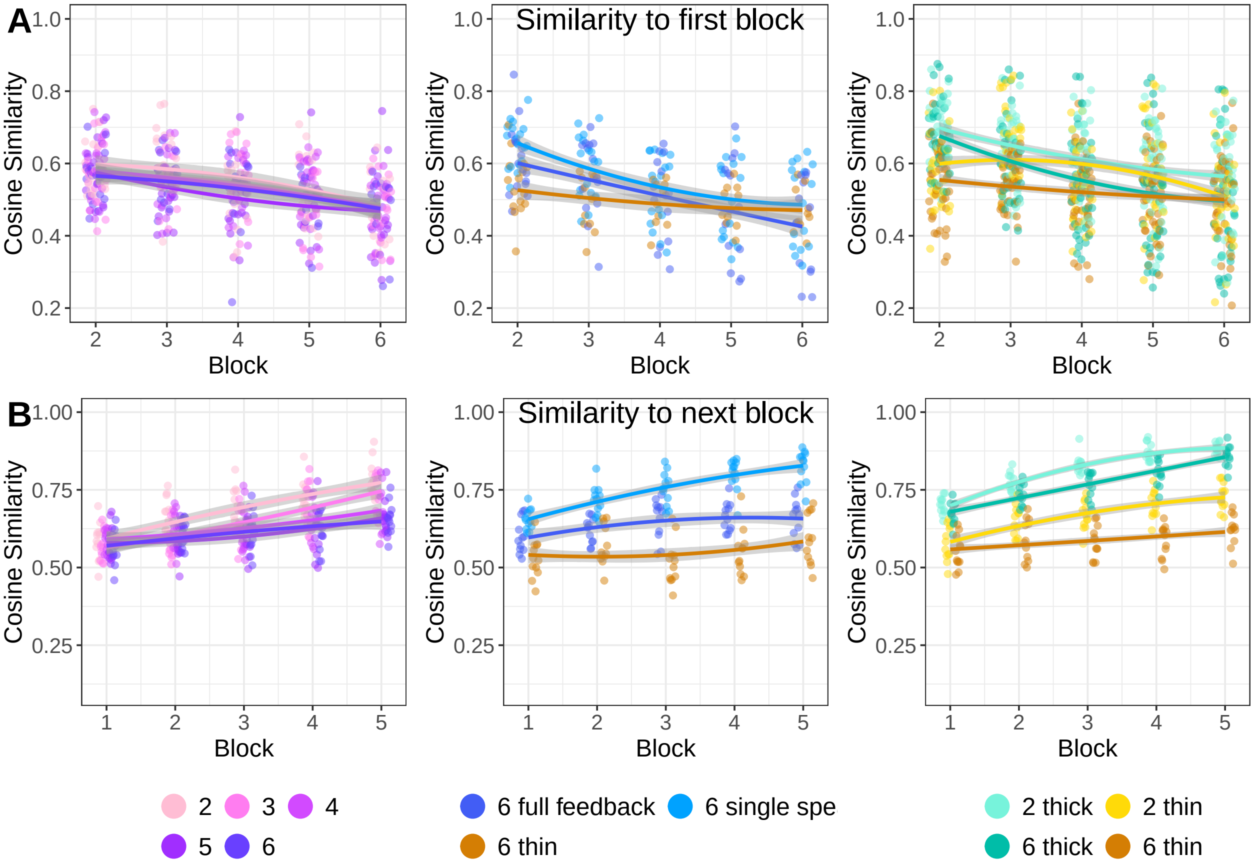

Figure 5: Additional measures of convergence and divergence. A is similarity to first utterance. B is similarity between utterances from adjacent blocks.

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.

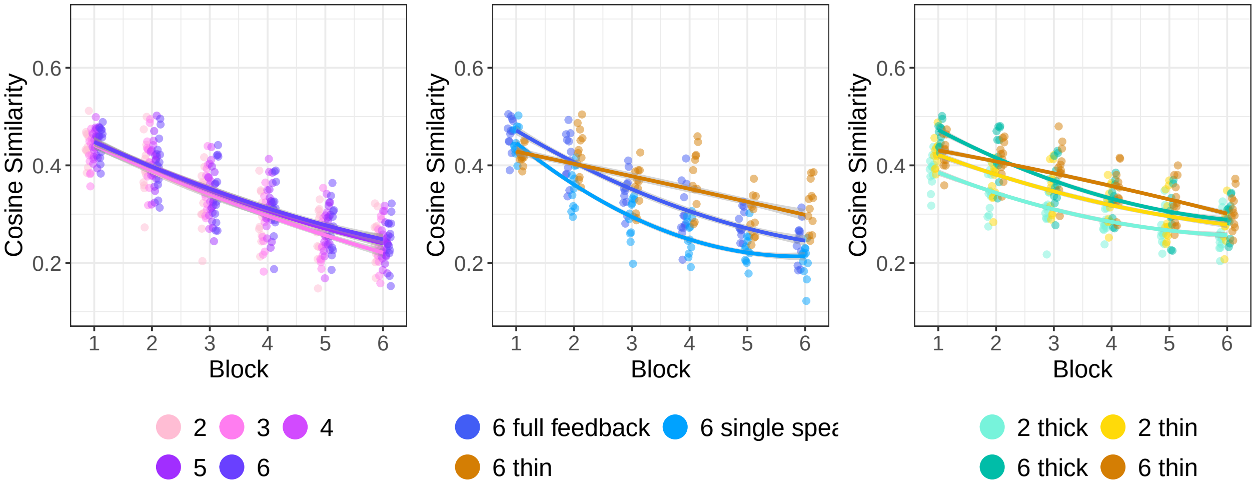

Figure 6: Divergence in descriptions of different tangrams within a group

Summary of model reporting

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-reg.

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

Term	Est.	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 single speaker logistic model of listener accuracy:
correct.num ∼ block + (1|gameId)

Term	Est.	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.	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.	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.	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

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.	CrI
block	-3.37	[-4.54, -2.24]
block:numPlayers	-0.10	[-0.36, 0.17]
Intercept	16.79	[11.96, 21.93]
numPlayers	1.66	[0.66, 2.61]

Extra reduction model

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

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

Term	Est.	CrI
block	-5.39	[-6.46, -4.31]
Intercept	29.93	[24.92, 34.84]

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

Term	Est.	CrI
block	-4.68	[-5.88, -3.52]
Intercept	26.03	[21.12, 30.58]

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

Term	Est.	CrI
block	-2.15	[-3.44, -1.12]
Intercept	20.50	[17.26, 23.76]

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

Term	Est.	CrI
block	-2.29	[-2.95, -1.6]
block:channelthin	0.32	[-0.65, 1.24]
block:channelthin:gameSize6	0.64	[-0.61, 1.89]
block:gameSize6	-1.21	[-2.06, -0.3]
channelthin	0.63	[-3.18, 4.73]
channelthin:gameSize6	-2.11	[-7.41, 2.98]
gameSize6	7.41	[3.57, 11.18]
Intercept	14.99	[11.86, 17.89]

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.

SBERT models

For all of the models using cosine similarity, we used linear models with the priors normal(.5,.2) for intercept, normal(0,.1) for beta, and normal(0,.05) for sd.

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

Term	Est.	CrI
block	-2.15	[-3.35, -0.98]
block:numPlayers	-0.23	[-0.51, 0.06]
block:wasINcorrect	0.25	[-0.24, 0.74]
Intercept	11.98	[6.31, 17.7]
numPlayers	2.15	[0.93, 3.36]
wasINcorrect	3.08	[1.69, 4.42]

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

Term	Est.	CrI
block	-0.17	[-1.63, 1.24]
block:numPlayers	-0.41	[-0.72, -0.09]
Intercept	4.67	[0.09, 9.32]
numPlayers	2.12	[1.03, 3.12]

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

Term	Est.	CrI
block	-0.80	[-0.97, -0.63]
block:numPlayers	0.03	[0, 0.07]
Intercept	-2.65	[-3.5, -1.83]
numPlayers	0.78	[0.58, 0.98]

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 (compare to first, next, and last), and divergence between tangrams within games.

Convergence within games: comparison to last round

This is the convergence metric presented in the paper.

Table 15: Experiment 1:sim ∼ earlier × condition + (1|tangram) + (1|gameId)

Term	Est.	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 16: Experiment 2: 6 single speaker:sim ∼ earlier + (1|tangram) + (1|gameId)

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

Table 17: Experiment 2: 6 full feedback:sim ∼ earlier + (1|tangram) + (1|gameId)

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

Table 18: Experiment 2: 6 thin:sim ∼ earlier + (1|tangram) + (1|gameId)

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

Table 19: Experiment 3:sim ∼ earlier × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	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 20: Experiment 1:sim ∼ block × condition + (1|tangram)

Term	Est.	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]

Divergence across tangrams

Table 21: Experiment 2: 6 single speaker:sim ∼ block + (1|tangram)

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

Table 22: Experiment 2: 6 full feedback:sim ∼ block + (1|tangram)

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

Table 23: Experiment 2: 6 thin:sim ∼ block + (1|tangram)

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

Table 24: Experiment 3:sim ∼ block × channel × gameSize + (1|tangram)

Term	Est.	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]

Table 25: Experiment 1:sim ∼ block × condition + (1|gameId)

Term	Est.	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 26: Experiment 2: 6 single speaker:sim ∼ block + (1|gameId)

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

Table 27: Experiment 2: 6 full feedback:sim ∼ block + (1|gameId)

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

Table 28: Experiment 2: 6 thin:sim ∼ block + (1|gameId)

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

Table 29: Experiment 3:sim ∼ block × channel × gameSize + (1|gameId)

Term	Est.	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 round utterance: 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 30: Experiment 1:sim ∼ earlier × condition + (1|tangram) + (1|gameId)

Term	Est.	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 31: Experiment 2: 6 single speaker:sim ∼ earlier + (1|tangram) + (1|gameId)

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

Table 32: Experiment 2: 6 full feedback:sim ∼ earlier + (1|tangram) + (1|gameId)

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

Table 33: Experiment 2: 6 thin:sim ∼ earlier + (1|tangram) + (1|gameId)

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

Table 34: Experiment 3:sim ∼ earlier × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	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 round utterance. This is not very informative because first round utterances tend to be pretty noisy with lots of hedges and filler words.

Table 35: Experiment 1:sim ∼ later × condition + (1|tangram) + (1|gameId)

Term	Est.	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 36: Experiment 2: 6 single speaker:sim ∼ later + (1|tangram) + (1|gameId)

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

Table 37: Experiment 2: 6 full feedback:sim ∼ later + (1|tangram) + (1|gameId)

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

Table 38: Experiment 2: 6 thin:sim ∼ later + (1|tangram) + (1|gameId)

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

Table 39: Experiment 3:sim ∼ later × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	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]