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

Contents

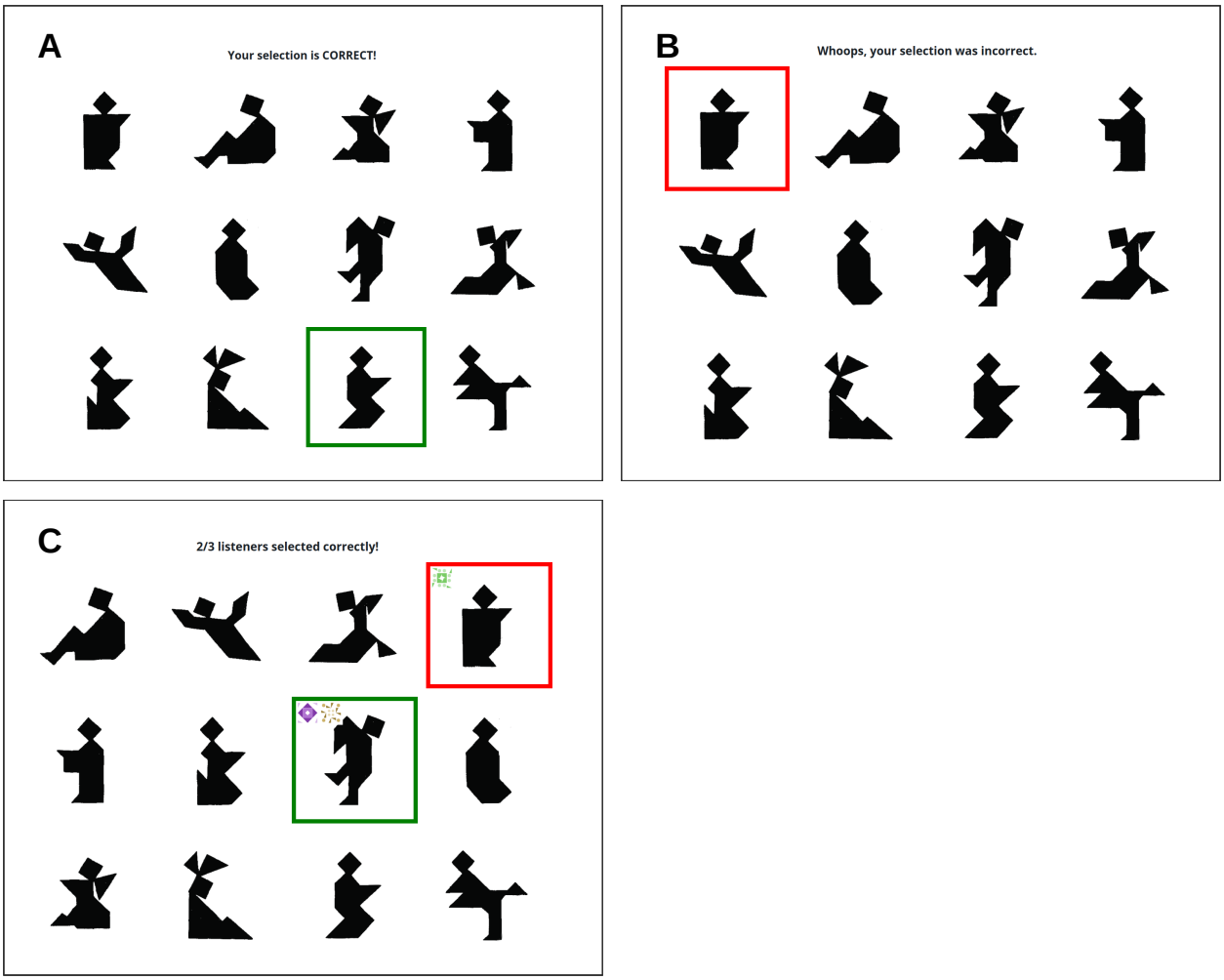

Figure 1: Feedback shown to participants. In low-coherence games, matchers saw feedback as shown in A and B. In all games, describers saw feedback like in C; icons were associated with players and indicated who selected which. In high feedback games, matchers saw grids like C, but with text as in A and B.

1 Feedback

Figure 1 shows the type of feedback shown to participants in different conditions.

2 Semantic content of descriptions

As an exploratory, qualitative analysis of the describer's language, we performed dictionary searches to classify whether the describers utterances on each trial contained the following types of words.

Geometric/literal words: squar*, triangle, triangular, diamond, shape, trapez*, angle, degree, parallel*, rhomb*, box, cube, line, white, black

Words for body parts: face, head, heads, back, shoulder, shoulders, arm, arms, leg, legs, foot, feet, body, knee, knees, toe, toes, hand, hands, body, butt, heel, heels, ear, ears, nose, neck, chest, hair

Positional words: right, left, above, below, under, over, top, bottom, behind, side, beneath

Posture words: kick*, crouch*, squat*, kneel*, knelt, stood, stand*, sit*, sat, lying, walk*, facing, fall*, looking, lean*, seat*, laying

Qualitatively, descriptions that contain none of these words are usually shorter, more abstract, more conventionalized descriptions. Figure 2 shows the fraction of trials where the speaker used these words. Generally, the more concrete word categories become less frequent over time, although this effect seems weaker in larger, thinner games.

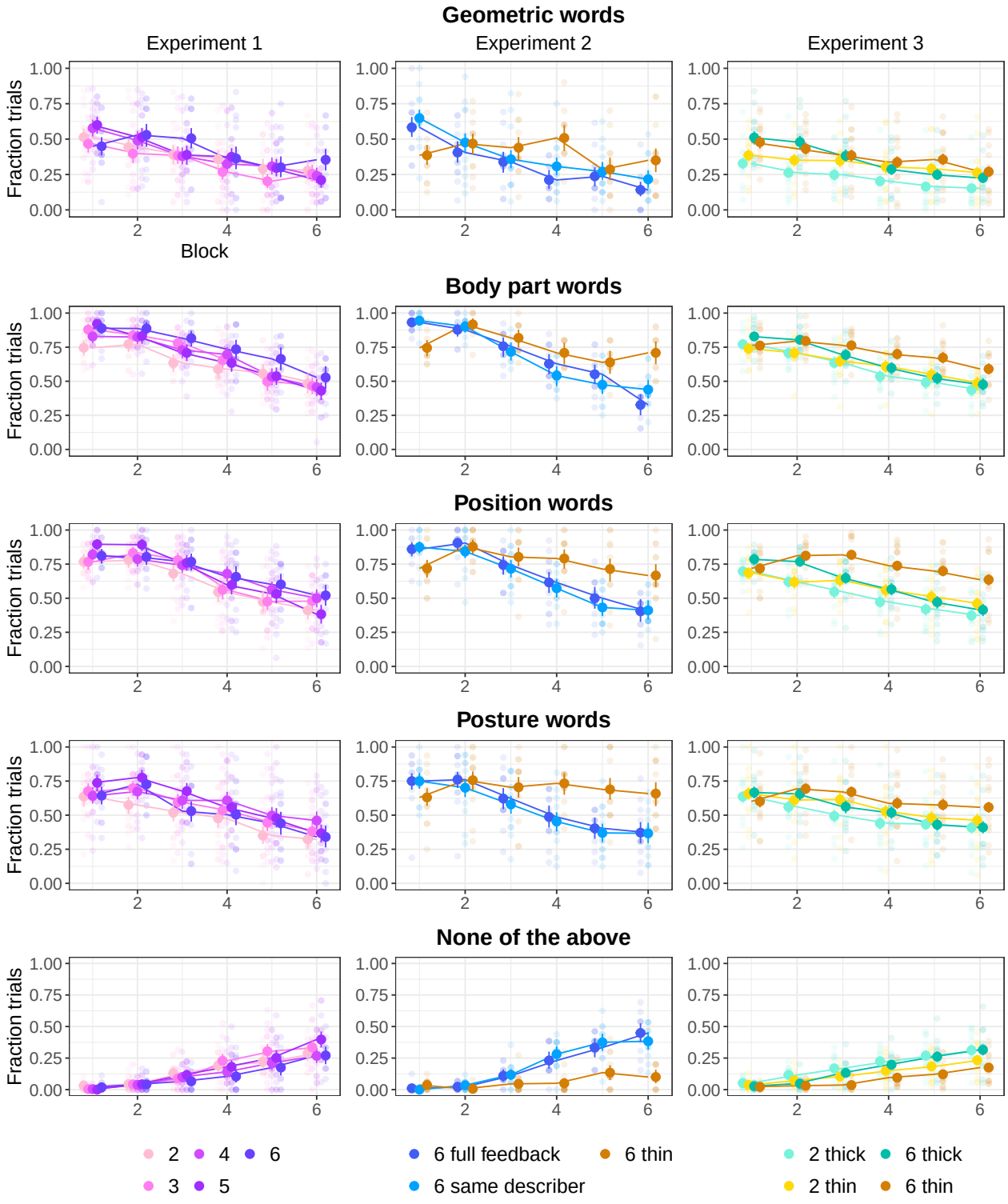

Figure 2: Fraction of trials where the describer used certain types of words. See text for word lists. Faint dots are per tangram, larger dots and error bars are means and boostrapped 95% CIs.

3 Additional game transcripts

Table 1: Additional examples from 6-player thick games for the same image across repetitions. Describers are indicated with an asterisk.

6-person thick game	
Rep 1: 5/5 correct	
Q*	oh no. okay. this one is similar to the one i called minnesota, so like pretty simple, with just a distinct "head" diamond and then a body beneath
Q*	the body has sort of sloping shoulders
P	BEAN SHAPED BODY?
O	with a small triangle on the bottom right?
Q*	yeah i think so! bean is a good description lol
Q*	bean body
Rep 2: 5/5 correct	
Q*	bean body!!
Rep 3: 5/5 correct	
Q*	bean body!
Rep 4: 5/5 correct	
Q*	bean body!
Rep 5: 5/5 correct	
Q*	you guys are doin' great. :) bean body for this one
Rep 6: 5/5 correct	
Q*	bean body for this one

6-person thick game	
Rep 1: 2/5 correct	
V*	This shape hasn't got much sticking out. Small triangle pointing right, tilted square on top
Y	Is the triangle on the right at the bottom?
V*	Long tall shape, not far off symmetrical.
V*	Yes triangle bottom right
W	Are their two triangles on it?
X	does it look like a rabbit?
V*	Not the rabbit
X	does it look like someone on their back kicking a square soccer ball
V*	It's like one block like a gemstone
V*	but square on top
V*	Not on back
V*	It's the one with the least sharp corners I think
V*	One long gem shape with small triangle bottom right, tilted square on top
Rep 2: 4/5 correct	
V*	Long tall shape. Little triangle sticking out bottom right. Diamond on top
Rep 3: 3/5 correct	
V*	Long tall hexagon, triangle bottom right, diamond on top
Rep 4: 4/5 correct	
V*	Long hexagon
V*	little triangle bottom right
V*	Diamond top
Rep 5: 4/5 correct	
V*	Long hexagon.
V*	Diamond on top
V*	little triangle bottom right
Rep 6: 4/5 correct	
V*	Long hexagon
V*	Triangle bottom right

Table 2: Additional examples from 6-player thin games for the same image across repetitions. Describers are indicated with an asterisk.

6-person thin game

Rep 1: 2/5 correct

J* Looks like a skinnier candle but with a shadow at the bottom
K 🤔
H ✅
J* The top square looks like the flame
J* It has a triangle leading to that flame
I ❌

Rep 2: 3/5 correct

K 🤔
I* the simplest image in the whole, no much issues
J 🤔
L 🤔
I* 1 square as the head
L 🤔
J 🤔
H 🤔
K 🤔
I* candle with more missing chunks

Rep 3: 1/5 correct

L* looks like a candle with more chunks missing
J 🤔
I 🤔
L* The bottom part faces the right and forms a triangle
L* The top is a square
I ✅
J ❌

Rep 4: 3/5 correct

H* Candle one with shadow
H* The straight one not the one with bits missing
L 🤔
I 😂
L 🤔
H* Lower case j but the bottom swings to the right
I ✅ ✅ ✅ ✅

Rep 5: 0/5 correct

K* Candle with chunks missing on the right
I ✅ ✅ ✅

Rep 6: 4/5 correct

G* candle shape with a chunk of the bottom left missing and added to the bottom right instead
K 🤔 🤔
H ✅
I ✅

6-person thin game

Rep 1: 2/5 correct

A* Square on top
A* At the bottom of the base part of it sticks out to the right
A* The base has 8 faces (8 sides of the shape)
C ✅
A* The square balances on a triangular point of the base
A* It is a thick base shape, the bottom curves right
D ✅

Rep 2: 5/5 correct

B* tombstone and its shadow with a rotated square on top
D ✅
F ✅
C 🤔

Rep 3: 5/5 correct

D* tombstone casting a shadow (sorry i've pinched that from before)
C 🤔
B 😂
E 🤔
C 🤔
E 🤔 🤔
D* square on top sitting on top of a triangle which is on top of a square, with 5 side shape at the bottom

Rep 4: 4/5 correct

C* looks mummified
C* person with no limbs
D ✅
B 🤔
C* square atop, chunky with triangle bottom right

Rep 5: 5/5 correct

F* Gravestone casting shadow. Square on top
D ✅

Rep 6: 5/5 correct

E* tombstone
D ✅

4 Recruitment and number of games

The number of games in each condition varied due to the realities of online recruitment for real-time games. We had control over how many participants we recruited on Prolific. Because of the real-time nature of the games, we relied on the fact that most Prolific workers who accepted the task would rapidly open it and go to our experiment site, where they would read the instructions and be assigned into games. Due to technical details about how games filled, we only ran one condition of games at any one time.

In experiments 1 and 2, games would only start if the game was full. Thus, if for example, we recruited 28 people, and 27 people went to the experiment when we were running a 4-person game, three people would be unable to make a complete game, and we also wouldn't have groupmates available for a 28th person who opened the experiment late.

We ran games in large batches to try to minimize the fraction of people who wouldn't be matched into a game. If we got many fewer (complete) games than anticipated, we could recruit another batch of participants, taking the expected attrition rate into account. However, the logistics of needing participants to all open the study at roughly the same time to match into games meant that it was infeasible to recruit fewer than 20 or so participants at a time, so we could not precisely achieve the targeted number of games. Thus, the number of games varied somewhat across conditions.

In addition to not having control over exactly how many games started, some games also ended prematurely when participants disconnected. These partial games were more common in larger groups. We speculate it could be due either to just having more people who might have connection issues or need to leave the game, or more frustration causing people to quit the game (sometimes very early in the game).

Both of these factors led to a low yield rate in terms of completed games per participants recruited, especially in experiment 2 where all the games were 6-player. To try to maximize the data collected per money spent on recruitment, we made adjustments for experiment 3. In experiment 3, we allowed games to them to start and continue with fewer than 6 people. Thus, while many of the "6-player" games did not have 6 players the entire game, we at least have data from the entire course of the game.

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.

In Figure 3, the number of games goes up in some cases because only complete blocks (where the describer said something every trial) are analysed. If there was initial confusion and a describer missed a trial, that

Table 3: 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: same describer	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

block was excluded.

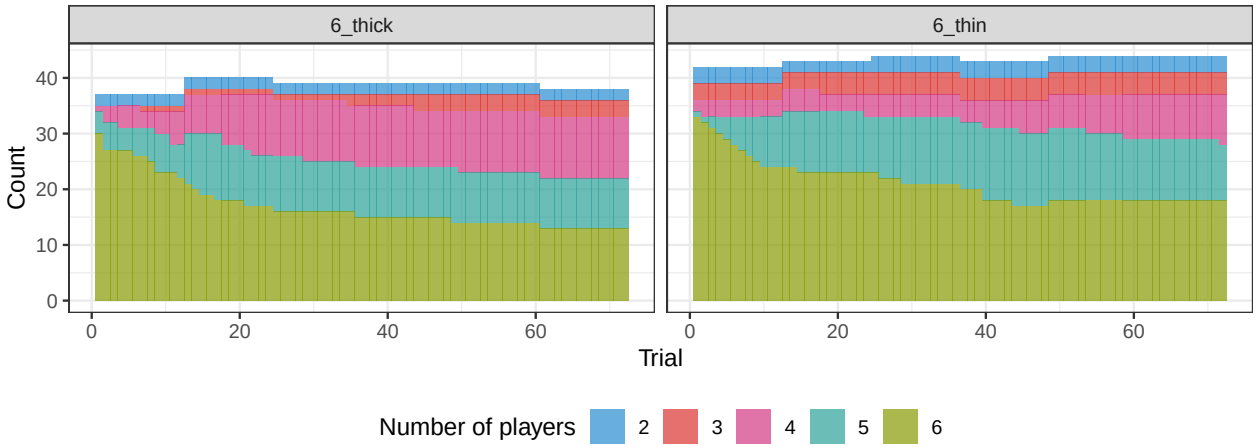

Figure 3: Number of players during 6 thin and 6 thick games in experiment 3. Blocks that were incomplete were excluded, so if a describer said nothing during a trial, that block was excluded.

5 Sensitivity analyses

As (post-hoc) sensitivity analyses, we re-ran our main analyses on the subset of games that completed all 6 blocks with a full set of players. This is a check on whether any of the results are driven by a) partial games that were included in early blocks but not later blocks or b) experiment 3 games that started without a full set of players or had players drop out.

Versions of main paper Figures 2 and 4 with just these complete games are shown here in SI Figures 4 and 5. The full model output for the sensitivity analyses is in SI Tables 54-73.

The only qualitative differences in the results between these models and those on the full dataset are in experiment 3. Note that large experiment 3 games were the ones most likely to be excluded, both for incompleteness and for not having the full 6 players throughout.

In the accuracy model, the block:gameSize interaction is not robust, and there is a marginal effect of block:gameSize:thin instead. The main effect of gameSize is also weaker. In the reduction model, the block:gameSize interaction is not robust. In the convergence model, the earlier:gameSize interaction is not robust.

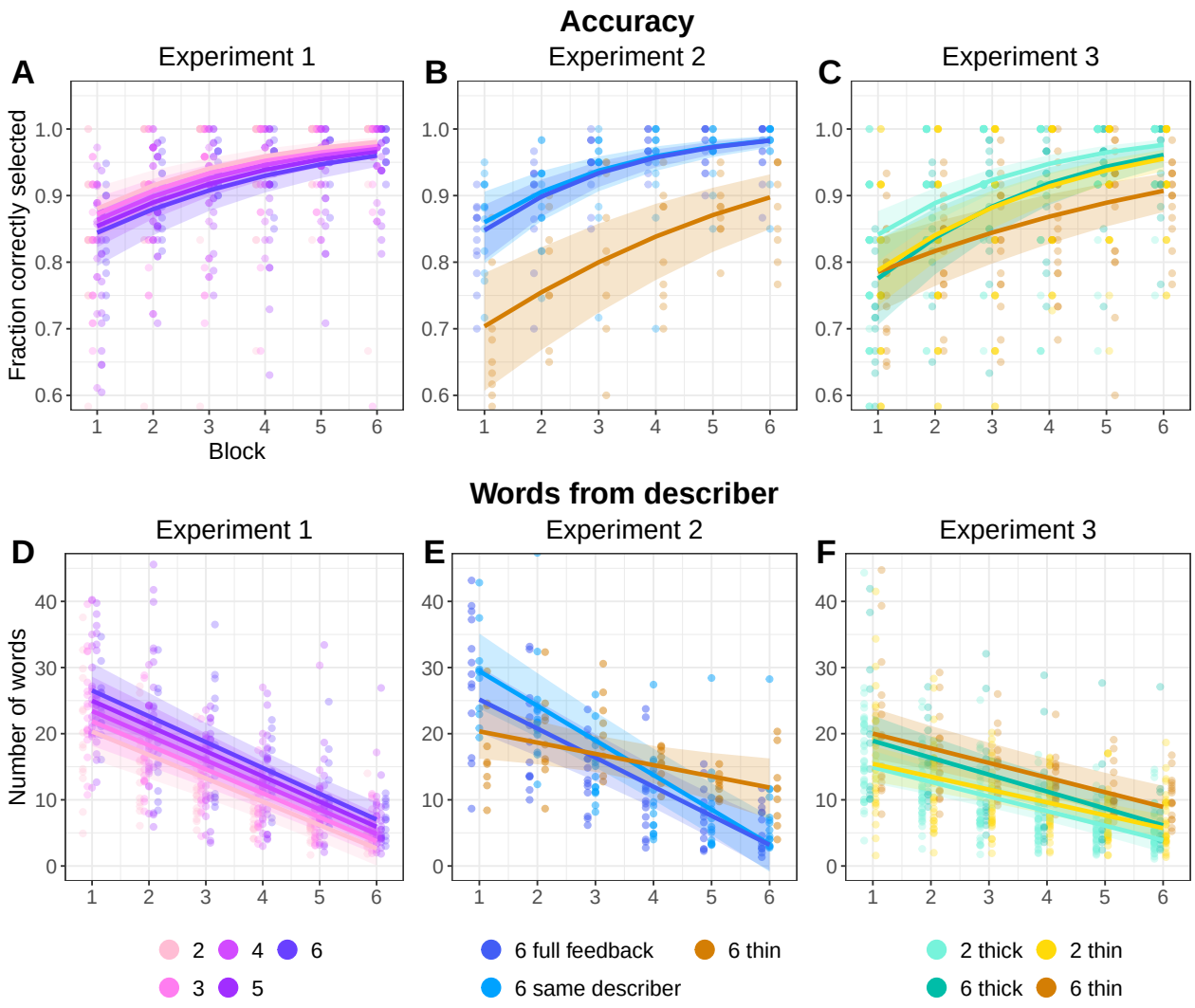

Figure 4: Equivalent of Main text Figure 2, with only full complete games. Behavioral results across all three experiments. (A-C). Matcher accuracy at selecting the target image. (D-F). Number of words produced by the describer each trial. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a few outliers points are not visible.

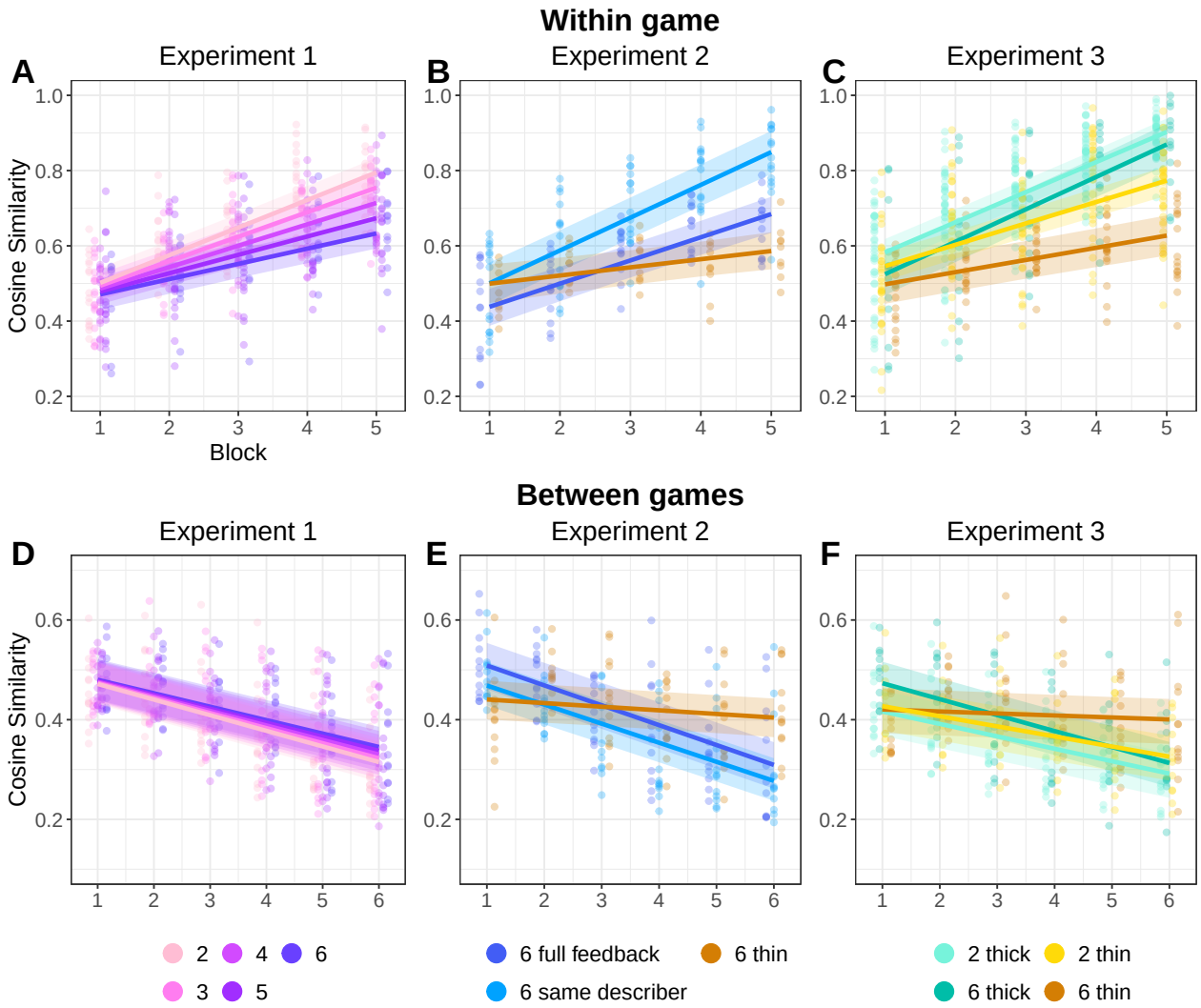

Figure 5: Equivalent of Main text Figure 4 with only full complete games. Language similarity results measured with pairwise cosine similarity between embeddings of two utterances. (A-C). Convergence of descriptions within games as measured by similarity between an utterance from block 1-5 to the block 6 utterance in the same game for the same image. (D-F). Divergence of descriptions across games as measured by the similarity between two utterances produced for the same image by different groups in the same block. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a few outliers points are not visible.

6 Matcher contributions

We can think about matcher utterances either at the level of one matcher or at the level of a group. In larger groups, the total number of matchers who could contribute is higher, so matcher utterances might be more common. On a per-matcher basis, having more group-mates could potentially increase the amount a given matcher contributes (more people to potentially have lateral conversations with) or it could decrease it (others are already asking the questions you were going to).

We analyzed matcher utterances both ways – on a game level and on an individual level. We broke it down into two steps: 1) are there matcher contributions on a given trial and 2), for trials with matcher contributions, how long are the contributions.

Matchers' contributions declined over the course of the game. The use of emoji in the thin games is not directly comparable to matcher language use in thick games, since some emoji usage (such as the green checkmark) is most likely equivalent to non-referential matcher 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 matchers having a lower threshold for sending emojis compared to writing out clarifications.

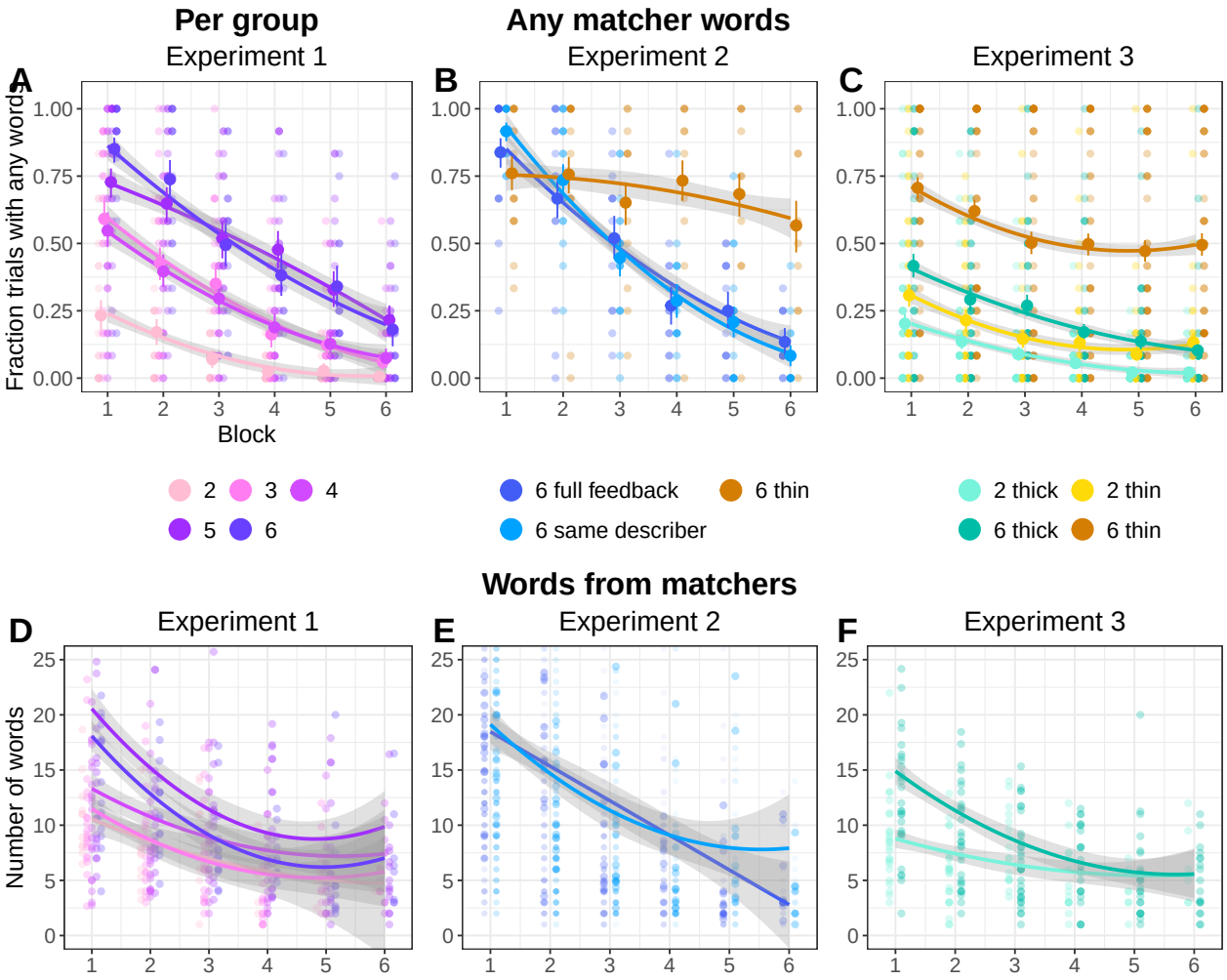

Figure 6: Matcher contributions at the group level. A-C: Fraction of trials where any matcher said anything that was referential. Dots are per game averages. Smooths are binomial fit lines. D-F: On trials where at least one matcher contributed, the number of words of referential language produced by matchers. Dots are per game averages. Smooths are quadratic fit lines. Y-axis is truncated, and a few outliers points are not visible.

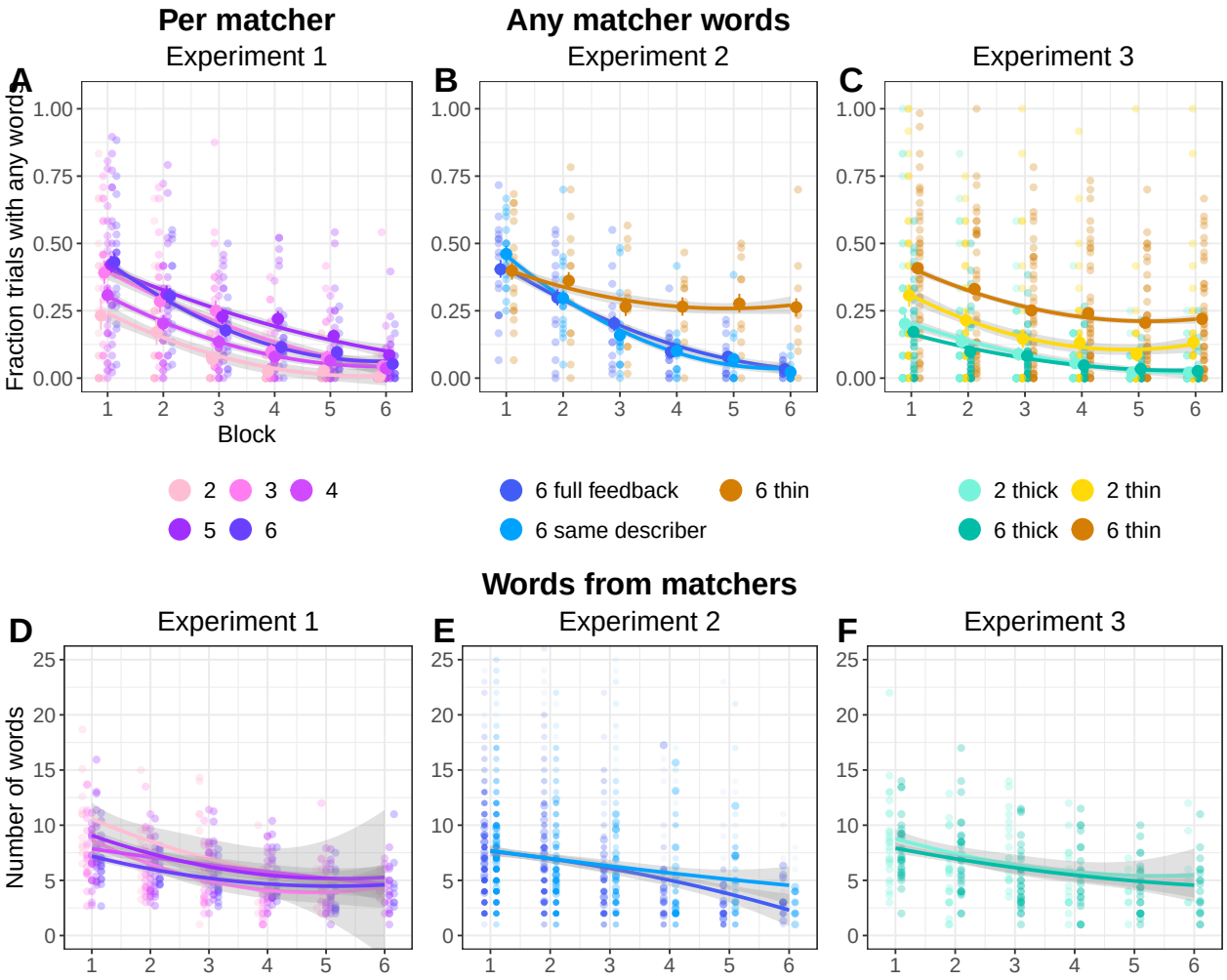

Figure 7: Matcher contributions at the individual level. A-C: Fraction of trials where an individual matcher said anything that was referential. Dots are per game averages. Smooths are binomial fit lines. D-F: On trials where a matcher contributed, the number of words of referential language that matcher produced. Dots are per game averages. Smooths are quadratic fit lines. Y-axis is truncated, and a few outliers points are not visible.

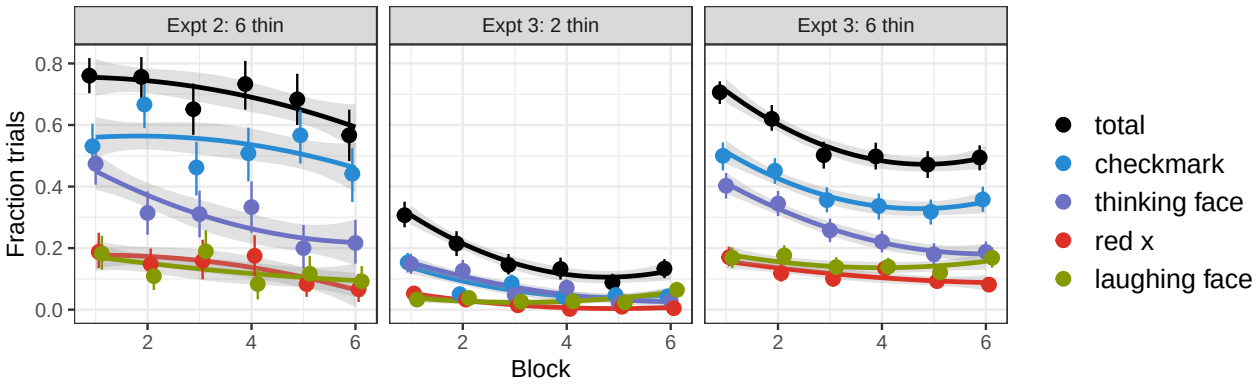

Figure 8: Fraction of trials on which at least one matcher produced the labelled emoji. Fraction of trials when any emoji was produced are shown in black. Dots are per condition, per block estimates with 95% bootstrapped CIs. Smooths are binomial fit lines.

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.

7 Reaction times

As the number of words produced by the speaker each trial decreases over repetition, so does the time each trial takes. In SI Figure 9, the time to selection (per-matcher) is shown in panels A-C, and the time per trial (equivalent to how long the slowest matcher took) is shown in panels B-D.

Matchers sped up over time across conditions, although thin large groups may not have speed up as much. Larger groups appear to be a bit slower at the level of individual matchers, which compounds to overall slower games as there are more matchers to wait for.

We also confirm in SI Figure 10 that there is a strong relationship between the number of words from the describer and the length of a trial.

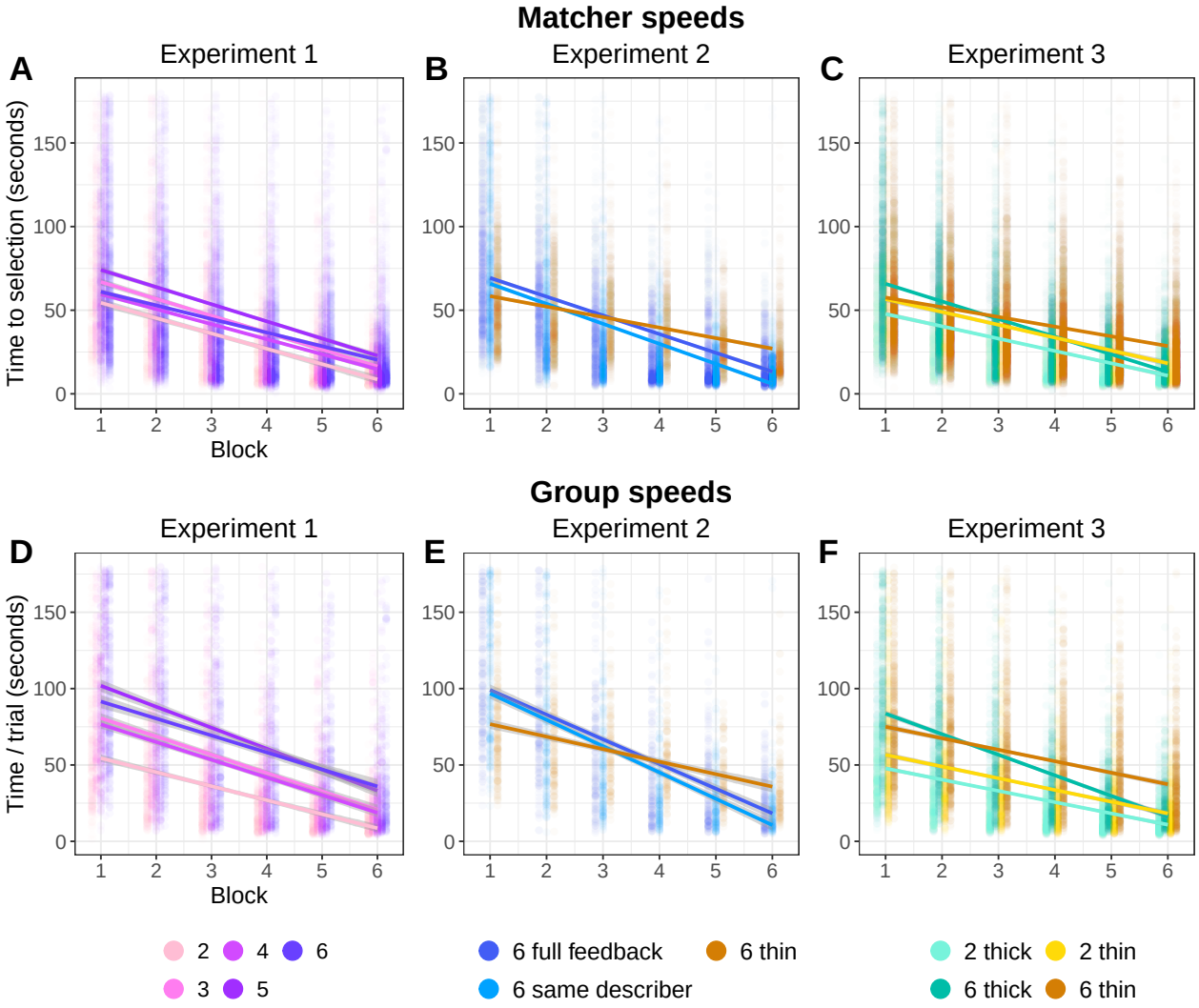

Figure 9: TODO Behavioral results across all three experiments. (A-C). Matcher accuracy at selecting the target image. (D-F). Number of words produced by the describer each trial. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a few outliers points are not visible.

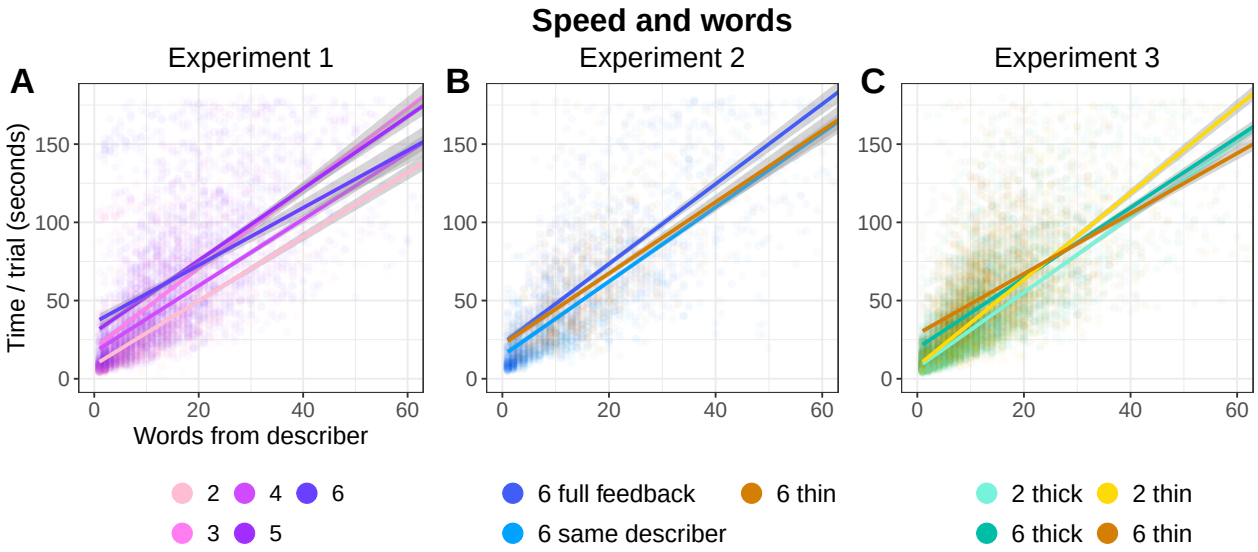

Figure 10: Relationship between number of words produced by describer and trial length. Across conditions, a similar pattern holds. X-axes are truncated, and a few outliers points are not visible.

8 Group level variation in models

One problem with precisely estimating the fixed effects of interest is that there is substantial variation at the game and tangram levels leading to wide confidence intervals on some of the fixed effects.

Models were fit with the maximum mixed effect structure for which the model would run in a reasonable amount of time.

As shown in SI Figure 11, the standard deviations for the grouping levels are substantial compared to the non-Intercept coefficient estimates. This means that the between-game or between-tangram variation is large compared to the effects of interest. This variation causes wide estimates as there is model uncertainty in how to apportion observed effects between group-level idiosyncrasies and population-level fixed effects.

This variability is diminished in the mega-analytic models which were fit with less extensive mixed effects (to aid in model convergence) and with more data.

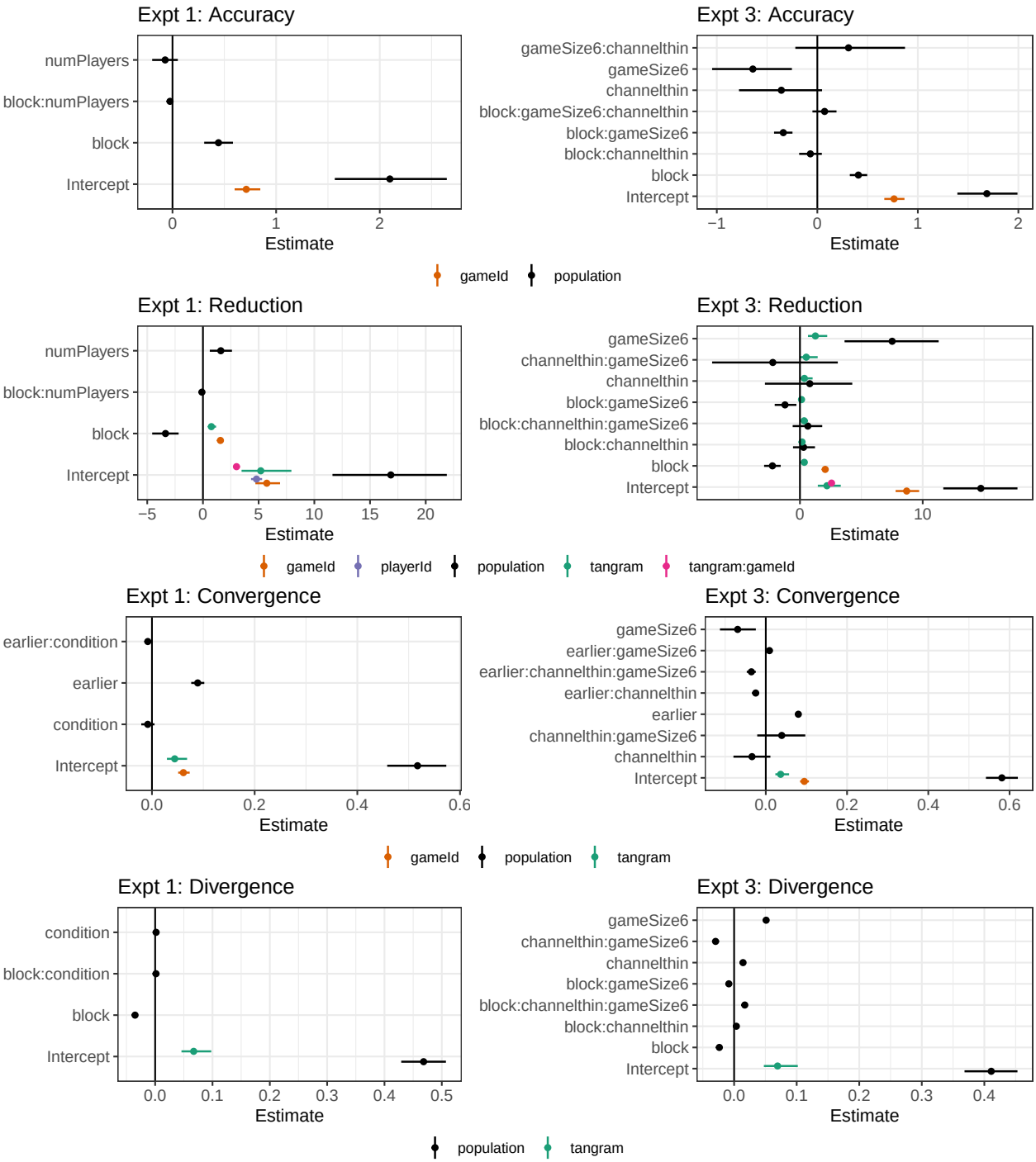

Figure 11: Comparison between the population-level coefficients (point estimate and 95% CI) and the group-level standard deviations (point estimates and 95% CIs) for the heirarchical effects.

9 Additional measures 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.

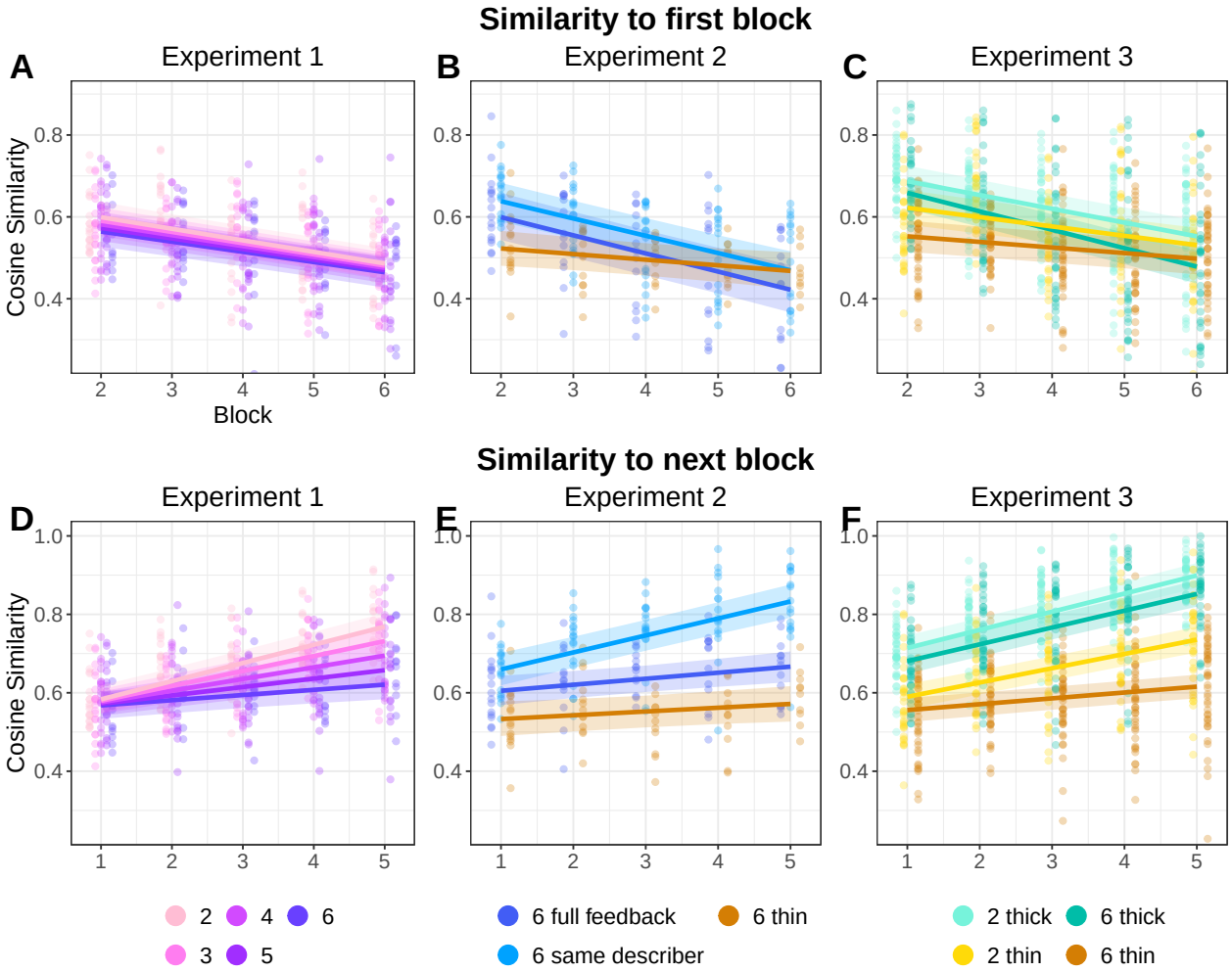

Figure 12: Additional measures of convergence and divergence. A-C is the similarity between utterances on a given block to the first block utterance for the same image, in the same game. Dots are per-game averages, smooths are quadratic. D-F is the similarity between utterances on a given block to the corresponding utterances in the next block. Dots are per-game averages, smooths are quadratic.

10 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.

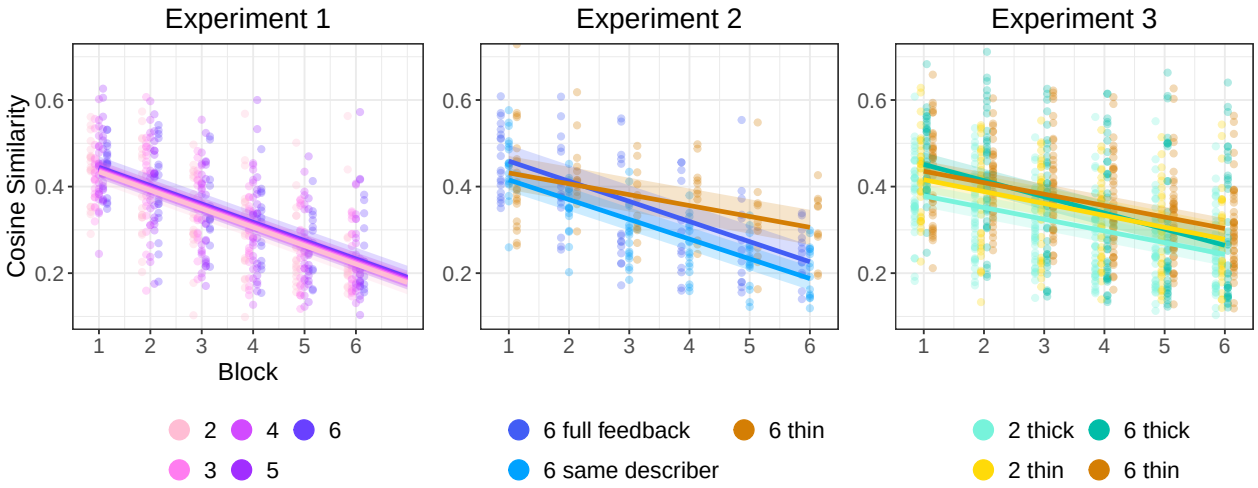

Figure 13: Divergence in descriptions of different tangrams. Cosine similarity between the descriptions of two different tangrams in the same block and group are shown. Dots are per-game averages, smooths are quadratic.

11 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.

12 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 4: Experiment 1 logistic model of matcher accuracy: correct.num ~ block × numPlayers + (1|gameId)

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

Table 5: Experiment 2: 6 same describer logistic model of matcher accuracy: correct.num ~ block + (1|gameId)

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

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

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

Table 7: Experiment 2: 6 thin logistic model of matcher accuracy: correct.num ~ block + (1|gameId)

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

Table 8: Experiment 3 logistic model of matcher accuracy: correct.num ~ block × gameSize × channel + (1|gameId)

Term	Est.	95% CrI
Intercept	1.69	[1.39, 1.99]
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]

13 Reduction models

13.1 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 9: Experiment 1: words ~ block × numPlayers + (block|tangram) + (1|playerId) + (1|tangram:gameId) + (block|gameId)

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

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

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

Table 11: Experiment 2: 6 full feedback: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

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

Table 12: Experiment 2: 6 thin: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

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

Table 13: Experiment 3: words ~ block × channel × gameSize + (block × channel × gameSize|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	14.74	[11.68, 17.72]
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]

13.2 Extra reduction model

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

Table 14: Experiment 1: words ~ block × numPlayers + block × wasINcorrect + (block|tangram) + (1|playerId) + (1|tangram:gameId) + (block|gameId)

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

13.3 Log reduction model

As an exploratory check, we reran some of the primary reduction models using the log number of words as the DV; otherwise model specifications and priors were the same.

Table 15: Experiment 1 log reduction: logwords ~ block × numPlayers + (block|tangram) + (1|playerId) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	2.75	[2.48, 3.01]
block	-0.38	[-0.47, -0.29]
block:numPlayers	0.01	[-0.01, 0.03]
numPlayers	0.07	[0.02, 0.12]

Table 16: Experiment 3 log reduction: logwords ~ block × channel × gameSize + (block × channel × gameSize|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	2.50	[2.32, 2.69]
block	-0.27	[-0.32, -0.22]
block:channelthin	0.06	[0, 0.14]
block:channelthin:gameSize6	0.03	[-0.06, 0.13]
block:gameSize6	-0.02	[-0.09, 0.04]
channelthin	0.06	[-0.17, 0.29]
channelthin:gameSize6	-0.13	[-0.44, 0.18]
gameSize6	0.45	[0.23, 0.67]

13.4 Matcher reduction models

These models were not pre-registered.

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

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

We ran these models both at the group level (did any matcher make a contribution, how many words of matcher contributions were there) and at the individual level (for a given matcher, did they make a contribution, and if so, how many words did they contribute).

Table 17: Experiment 1: group-level: is.words ~ block × numPlayers + (1|gameId)

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

Table 18: Experiment 1: individual-level: is.words ~ block × numPlayers + (1|playerId)

Term	Est.	95% CrI
Intercept	-1.65	[-2.24, -1.05]
block	-0.82	[-0.96, -0.68]
block:numPlayers	0.05	[0.02, 0.08]
numPlayers	0.20	[0.07, 0.33]

Table 19: Experiment 1: group-level: words ~ block × numPlayers + (block|gameId)

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

Table 20: Experiment 1: individual-level: words ~ block × numPlayers + (block|playerId)

Term	Est.	95% CrI
Intercept	9.85	[8.3, 11.4]
block	-1.45	[-2.05, -0.86]
block:numPlayers	0.14	[0.02, 0.27]
numPlayers	-0.52	[-0.84, -0.21]

13.5 Initial utterance reduction model

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

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

Table 21: 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 22: Experiment 2: 6 same describer: 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 23: 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 24: 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]

14 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.

14.1 Convergence within games: comparison to last round

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

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

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

Table 26: Experiment 2: 6 same describer: sim ∼ earlier + (1|tangram) + (1|gameId)

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

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

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

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

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

Table 29: Experiment 3: sim ~ earlier × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.581	[0.542, 0.62]
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]

14.2 Divergence across games

This is the divergence metric presented in the paper.

Table 30: Experiment 1: sim ~ block × condition + (1|tangram)

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

Table 31: Experiment 2: 6 same describer: sim ~ block + (1|tangram)

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

Table 32: Experiment 2: 6 full feedback: sim ~ block + (1|tangram)

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

Table 33: Experiment 2: 6 thin: sim ~ block + (1|tangram)

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

Table 34: Experiment 3: sim ~ block × channel × gameSize + (1|tangram)

Term	Est.	95% CrI
Intercept	0.411	[0.368, 0.453]
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]

14.3 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 35: Experiment 1: sim ~ block × condition + (1|gameId)

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

Table 36: Experiment 2: 6 same describer: sim ~ block + (1|gameId)

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

Table 37: Experiment 2: 6 full feedback: sim ~ block + (1|gameId)

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

Table 38: Experiment 2: 6 thin: sim ~ block + (1|gameId)

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

Table 39: Experiment 3: sim ~ block × channel × gameSize + (1|gameId)

Term	Est.	95% CrI
Intercept	0.378	[0.352, 0.404]
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]

14.4 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 40: Experiment 1: sim ~ earlier × condition + (1|tangram) + (1|gameId)

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

Table 41: Experiment 2: 6 same describer: sim ~ earlier + (1|tangram) + (1|gameId)

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

Table 42: Experiment 2: 6 full feedback: sim ~ earlier + (1|tangram) + (1|gameId)

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

Table 43: Experiment 2: 6 thin: sim ~ earlier + (1|tangram) + (1|gameId)

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

Table 44: Experiment 3: sim ~ earlier × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.714	[0.682, 0.746]
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]

14.5 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 45: Experiment 1: sim ~ later × condition + (1|tangram) + (1|gameId)

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

Table 46: Experiment 2: 6 same describer: 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 47: 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 48: 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 49: Experiment 3: sim ~ later × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.721	[0.681, 0.76]
channelthin	-0.076	[-0.123, -0.026]
channelthin:gameSize6	-0.062	[-0.127, 0.001]
gameSize6	-0.017	[-0.062, 0.03]
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]

15 Exploratory Mega-analytic models

As an exploratory measure, we combined data across all experiments and re-ran the core set of models (same structure and priors as for the individual models). We binarized game thickness: all games that were not thin were counted as thick. The intercept condition is a 2-player game, in the thick condition, for the first block. Thinner means the game was in the thin condition; larger is per additional player, and block is per later block. Group size was coded based on assigned condition, not based on actual number of players at the time.

Table 50: Mega-analytic on accuracy: correct.num ~ block × thinner × larger + (1|gameId)

Term	Est.	95% CrI
Intercept	1.83	[1.6, 2.07]
block	0.46	[0.4, 0.52]
block:larger	-0.07	[-0.09, -0.05]
block:thinner	-0.12	[-0.21, -0.02]
block:thinner:larger	0.01	[-0.02, 0.04]
larger	-0.07	[-0.15, 0]
thinner	-0.50	[-0.89, -0.08]
thinner:larger	-0.02	[-0.14, 0.11]

Table 51: Mega-analytic of reduction: words ~ block × thinner × larger + (block|gameId)

Term	Est.	95% CrI
Intercept	17.38	[15.59, 19.21]
block	-2.80	[-3.24, -2.36]
block:larger	-0.36	[-0.51, -0.2]
block:thinner	0.79	[0.04, 1.52]
block:thinner:larger	0.26	[0, 0.51]
larger	2.12	[1.5, 2.75]
thinner	-1.59	[-4.61, 1.55]
thinner:larger	-0.90	[-1.92, 0.11]

Table 52: Mega-analytic on divergence between groups: sim ∼ block × thinner × larger

Term	Est.	95% CrI
Intercept	0.428	[0.426, 0.431]
block	-0.026	[-0.026, -0.025]
block:larger	-0.002	[-0.002, -0.002]
block:thinner	0.005	[0.004, 0.007]
block:thinner:larger	0.004	[0.004, 0.005]
larger	0.012	[0.011, 0.013]
thinner	-0.003	[-0.007, 0.001]
thinner:larger	-0.007	[-0.008, -0.006]

Table 53: Mega-analytic on convergence to last: sim ∼ earlier × larger × thinner

Term	Est.	95% CrI
Intercept	0.546	[0.534, 0.558]
earlier	0.072	[0.067, 0.077]
earlier:larger	0.000	[-0.002, 0.002]
earlier:larger:thinner	-0.007	[-0.01, -0.004]
earlier:thinner	-0.016	[-0.025, -0.008]
larger	-0.016	[-0.021, -0.012]
larger:thinner	0.009	[0.002, 0.016]
thinner	0.001	[-0.021, 0.021]

16 Sensitivity analysis models

These are the same specification as the primary models, just on the subset of the data from games that were full and complete.

16.1 Accuracy

Table 54: Experiment 1 logistic model of matcher accuracy: correct.num ∼ block × numPlayers + (1|gameId)

Term	Est.	95% CrI
Intercept	2.15	[1.5, 2.78]
block	0.43	[0.29, 0.57]
block:numPlayers	-0.02	[-0.05, 0.01]
numPlayers	-0.08	[-0.22, 0.07]

Table 55: Experiment 2: 6 same describer logistic model of matcher accuracy: correct.num ∼ block + (1|gameId)

Term	Est.	95% CrI
Intercept	1.81	[1.38, 2.27]
block	0.45	[0.37, 0.52]

Table 56: Experiment 2: 6 full feedback logistic model of matcher accuracy: correct.num ~ block + (1|gameId)

Term	Est.	95% CrI
Intercept	1.72	[1.41, 2.05]
block	0.46	[0.39, 0.54]

Table 57: Experiment 2: 6 thin logistic model of matcher accuracy: correct.num ~ block + (1|gameId)

Term	Est.	95% CrI
Intercept	0.86	[0.43, 1.29]
block	0.26	[0.21, 0.31]

Table 58: Experiment 3 logistic model of matcher accuracy: correct.num ~ block × gameSize × channel + (1|gameId)

Term	Est.	95% CrI
Intercept	1.67	[1.4, 1.96]
block	0.41	[0.32, 0.5]
block:channelthin	-0.06	[-0.18, 0.06]
block:gameSize6	-0.02	[-0.12, 0.09]
block:gameSize6:channelthin	-0.14	[-0.27, 0]
channelthin	-0.36	[-0.77, 0.05]
gameSize6	-0.43	[-0.9, 0.03]
gameSize6:channelthin	0.42	[-0.21, 1.04]

16.2 Reduction

Table 59: Experiment 1: words ~ block × numPlayers + (block|tangram) + (1|playerId) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	17.22	[11.42, 22.74]
block	-3.35	[-4.64, -2.05]
block:numPlayers	-0.09	[-0.39, 0.21]
numPlayers	1.55	[0.43, 2.69]

Table 60: Experiment 2: 6 same describer: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	29.45	[23.78, 35.14]
block	-5.23	[-6.36, -4.05]

Table 61: Experiment 2: 6 full feedback: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	25.17	[20.16, 30.14]
block	-4.40	[-5.5, -3.28]

Table 62: Experiment 2: 6 thin: words ~ block + (block|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	20.35	[16.21, 24.45]
block	-1.70	[-3.03, -0.38]

Table 63: Experiment 3: words ~ block × channel × gameSize + (block × channel × gameSize|tangram) + (1|tangram:gameId) + (block|gameId)

Term	Est.	95% CrI
Intercept	14.93	[12.54, 17.43]
block	-2.21	[-2.77, -1.67]
block:channelthin	0.26	[-0.48, 1.05]
block:channelthin:gameSize6	0.07	[-1.28, 1.45]
block:gameSize6	-0.35	[-1.32, 0.64]
channelthin	0.52	[-2.53, 3.52]
channelthin:gameSize6	0.58	[-4.8, 6.19]
gameSize6	3.99	[0.09, 8.14]

16.3 Convergence within games

Table 64: Experiment 1: sim ~ earlier × condition + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.517	[0.458, 0.575]
condition	-0.008	[-0.02, 0.005]
earlier	0.090	[0.077, 0.103]
earlier:condition	-0.008	[-0.011, -0.005]

Table 65: Experiment 2: 6 same describer: sim ~ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.501	[0.443, 0.556]
earlier	0.087	[0.079, 0.095]

Table 66: Experiment 2: 6 full feedback: sim ~ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.438	[0.388, 0.486]
earlier	0.062	[0.052, 0.072]

Table 67: Experiment 2: 6 thin: sim ~ earlier + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.499	[0.448, 0.55]
earlier	0.022	[0.011, 0.032]

Table 68: Experiment 3: sim ~ earlier × channel × gameSize + (1|tangram) + (1|gameId)

Term	Est.	95% CrI
Intercept	0.583	[0.549, 0.618]
channelthin	-0.037	[-0.084, 0.008]
channelthin:gameSize6	0.010	[-0.069, 0.09]
earlier	0.080	[0.075, 0.086]
earlier:channelthin	-0.023	[-0.032, -0.015]
earlier:channelthin:gameSize6	-0.031	[-0.046, -0.015]
earlier:gameSize6	0.006	[-0.005, 0.017]
gameSize6	-0.058	[-0.116, -0.003]

16.4 Divergence across games

Table 69: Experiment 1: sim ~ block × condition + (1|tangram)

Term	Est.	95% CrI
Intercept	0.468	[0.426, 0.51]
block	-0.034	[-0.037, -0.03]
block:condition	0.001	[0, 0.002]
condition	0.002	[0, 0.005]

Table 70: Experiment 2: 6 same describer: sim ~ block + (1|tangram)

Term	Est.	95% CrI
Intercept	0.469	[0.429, 0.516]
block	-0.038	[-0.041, -0.036]

Table 71: Experiment 2: 6 full feedback: sim ~ block + (1|tangram)

Term	Est.	95% CrI
Intercept	0.509	[0.461, 0.554]
block	-0.040	[-0.043, -0.037]

Table 72: Experiment 2: 6 thin: sim ~ block + (1|tangram)

Term	Est.	95% CrI
Intercept	0.441	[0.401, 0.478]
block	-0.007	[-0.011, -0.004]

Table 73: Experiment 3: sim ~ block × channel × gameSize + (1|tangram)

Term	Est.	95% CrI
Intercept	0.416	[0.368, 0.458]
block	-0.025	[-0.026, -0.024]
block:channelthin	0.005	[0.003, 0.006]
block:channelthin:gameSize6	0.023	[0.019, 0.027]
block:gameSize6	-0.007	[-0.01, -0.004]
channelthin	0.011	[0.006, 0.015]
channelthin:gameSize6	-0.064	[-0.075, -0.052]
gameSize6	0.057	[0.048, 0.066]