ONLINE SUPPLEMENT FOR “TEAM TALK”

Rows in Table S1 distinguish the 73 LIWC categories and rows distinguish the five intervals of adjacent three trials. Cell entries are the percent of words during each column interval that come from each row category of words (see Srivastava et al., 2018:1361-1362, for example profiles extracted from management email by an earlier LIWC version) A data note here is that subjects typed messages on computers, so their messages contain typos that subjects have come to expect their smart phones to correct (e.g., “sumbit” versus "submit,” “traingle” versus “triangle”). Far and away, the most common typo is a misplaced space within or between words (e.g., “tha tone” versus “that one,” “noone” versus “no one,”sittin gman” versus “sitting man”). LIWC’s use of word stems makes it robust to some of the typos, but there are likely more words in the messages than the raw text implies. In the spirit of data sanctity, we use text here exactly as it was typed by subjects.

However, as a reliability check on the word counts, we went through the 74,861 messages to correct typos. We did not do global replacements, or insert missing apostrophes, many of which are automatically corrected by LIWC (e.g., “oclock” is not changed to “o’clock,” “dont” is not changed to “don’t,” “im” is not changed to “I’m”), or replace internet contractions, some of which are included in the LIWC dictionary (e.g., “plz” was not replaced with “please,” “nvm” not replaced with “never mind,” “omg” not replaced with “oh my God”). We also preserved word combinations that teammates were using as symbol labels (e.g., “sittingman” is sometimes used as a label, and sometimes used as a description “sitting man.” Our rule was to be consistent with the most common use within a trial. We ran the 685 pages of message text through Microsoft Word’s spell checker to identify suspect words, then read context around each ostensibly misspelled word to make sure a suspect word was in fact a typo before correcting it. After doing this once, we repeated the process a second time to check for now more visible typos.

The LIWC results are surprisingly robust. As expected, the corrected text contains more words (445,743 words in the raw text, 449,149 in the corrected. But the percentage of words captured by the LIWC dictionary is little changed (79.69% of words in raw text versus 80.80% of words in the corrected text), and the graph in Figure S1 shows negligible change in the individual LIWC category percentages. The 1.00 correlation in the graph registers no change in LIWC percentages. There are changes, but they are small: a linear regression equation predicting the vertical from the horizontal axis in the graph has a zero intercept and a slope slightly greater than one (1.014), so LIWC percentages for more prominent categories are slightly higher in the corrected text.

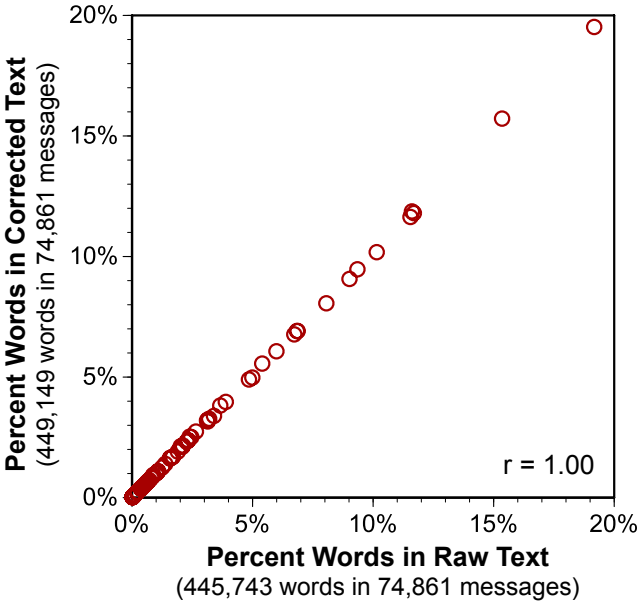


Figure S1. Percent Words in Raw and Corrected Text

We went a level deeper into the data to determine LIWC stability at the level of individual subjects within trials. The result is two bags of words for each subject in each trial (a total of 4,778 person-trials): One bag contains words of raw text the subject sent during the trial. The second bag contains words of corrected text for the same messages. The average bag of raw text contains about one hundred words of which about three quarters are in the LIWC dictionary (92.88 mean words, 73.55%). The average bag of corrected text is about the same, but one word larger with slightly more words from the LIWC dictionary (93.59 words, 75.43% in dictionary). Differences between the averages are small, but vary considerably between subjects and trials. The one word difference varies from zero for many individuals up to a maximum of seven. The two percent difference in LIWC dictionary words varies from zero for many individuals up to a 66.67 maximum. Therefore it is useful to see the reliability results in the right-most column of Table S1. The column contains for each LIWC category, the correlation across subjects and trials between the percent words in the category based on raw versus corrected text. Correlations are high within the 73 categories (.97 mean), especially for function words (.99 correlation), the key category in our analysis.

There might appear to be contradictory reports on percent words captured by the LIWC dictionary. Average percentages in the preceding paragraph state that 73.55% of words in the raw text are captured, and 75.43% in the corrected text. Earlier in this Appendix, and in the text of the paper, we said that 80% of words are captured by the LIWC dictionary. Both statements are correct. The graph in Figure S2

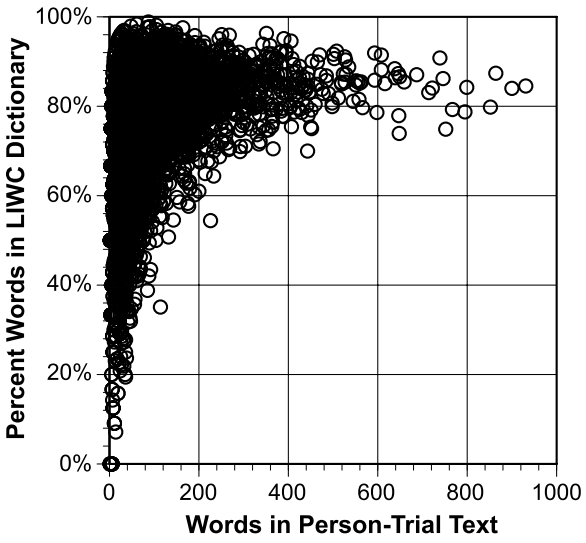


Figure S2. Longer Messages Contain a Higher Percent LIWC Words

shows how percent words captured by the LIWC dictionary varies with the volume of words in a subject's messages during a trial: The LIWC dictionary captures higher proportions of words in the text of subjects sending more words during a trial. In this experiment, subjects in later trials send fewer, shorter messages that contain fewer words in the LIWC dictionary (Figure 3, and bottom left in Figure S2). The percent words captured by LIWC when all messages are treated as one bag of words gives more weight to longer messages since they contain more words (upper right in Figure S2), so the percent words captured by LIWC in the text of all messages is the higher 80%. Average percent across person-trial observations gives equal weight to later and earlier trials, which means later trials are given more weight than if the messages were treated as one text, so the percent captured by LIWC in the messages sent by an average person during an average trial is the lower 75%. Percentages in Table S1 are LIWC results combining all messages during the column trials as a single bag of words.

FUNCTION WORDS

In complement to Figure 3 in the text, Figure S3 is a graphic display of change between the initial and last trials in the use of LIWC word categories. Categories are distinguished on the horizontal axis by the frequency with which they are used during the first three trials (first LIWC profile in Table S1). The vertical axis orders categories by their use during the last three trials (last LIWC profile in Table S1).

Word categories close to the diagonal dashed line in Figure S3 are used about as often in early messages as in later messages. For example, the LIWC category "verb" is high and close to the diagonal dashed line, indicating that verbs are used often in messages during the initial trials as well as during the final trials (Table S1 shows 20% verbs in trial 1-3 messages, 18% verbs in trial 13-15 messages).

Word categories above the diagonal are used more often in later messages than in initial messages. The LIWC categories that stand out for increased use by more experienced teams are "relativity," "body" and "biology" (which includes "body" as a subcategory), and "motion." These

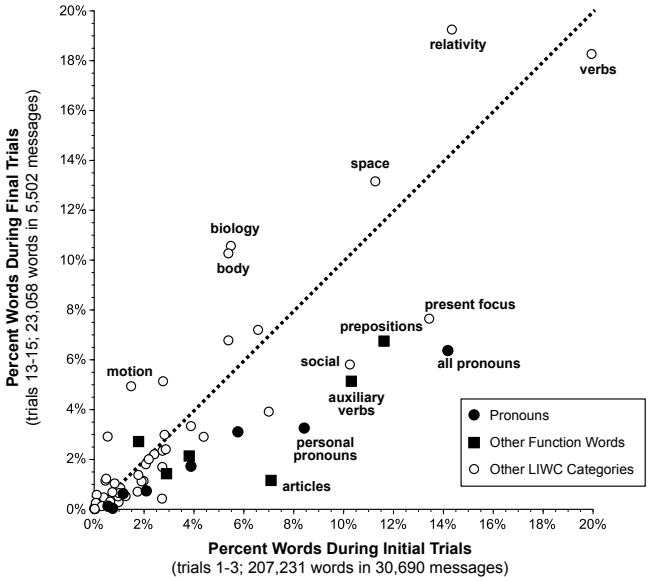


Figure S3.
What Kinds of Words Change Most?

categories do not indicate coordination so much as they indicate subject preferences for anthropomorphic labels.

Word categories below the diagonal in Figure S3 are used less often in later messages than in initial messages. We identify in the graph symbols for LIWC categories of function words. Most subcategories of function words in Figure S3 lie below the dashed-line diagonal, especially pronouns, auxiliary verbs, and articles. In other words, teams make less use of function words as they gain experience.

Further complementing Figure 3 in the text, Figure S4 shows how the aggregate drift away from function words in Figure 3 involves drift away from most subcategories of function words. Prepositions decline from 12 to seven percent of message words. Pronouns of all kinds decline. Verb modifiers decline. Articles and conjunctions decline. The one category of function words that enjoys continued, slightly increasing, use is negations. Teammates succinctly indicate with negations symbols that are not the correct answer.

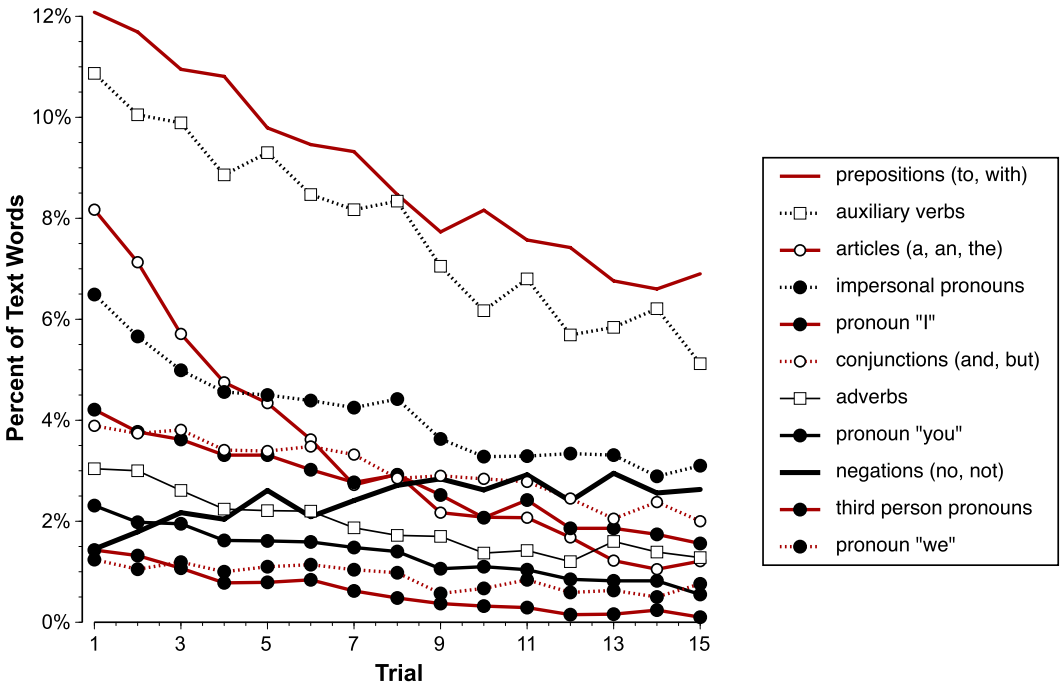


Figure S4.
Most Subcategories of Function Words Are Used Less in Later Trials.

With respect to coordination on certain function words, Figure S5A shows that no one stands out for leading or lagging in the drift away from function words. The line through solid dots shows for each trial the average function word use by the person in each team who uses the lowest percent function words during the trial. Call that subject the minimum user. The line through hollow dots shows the same thing for the teammate who uses the highest percent function words during the trial. Call that subject the maximum user. The dashed line with no dots is the average percent function word use by the other three teammates during a trial. The three lines have similar slopes, and the dashed line is about equidistant from the two solid lines. The dashed line would be closer to the upper line if the person using the least function words was well ahead of the team, leading the drift away from function words. The dashed line would be closer to the lower line if the person using the most function words was a laggard in the team average drift away from function words. The fact that the dashed line goes down the middle of the space between the solid lines shows that the person using the most function words, and the person using the least, are both outliers to the average use of function words in their team. More specifically, 10.5 percent of teammates using function words the least in this trial are maximum user in the next trial. Of teammates using function words the most in this trial, 9.1 percent are minimum user in the next trial. These cross-overs are less than the 20 percent expected if subjects move at random between maximum and minimum use, but higher than expected if subjects behave consistently across trials. Finally, the number of trials during which a subject is minimum user is independent of the extent to which teammates view the subject as team leader.

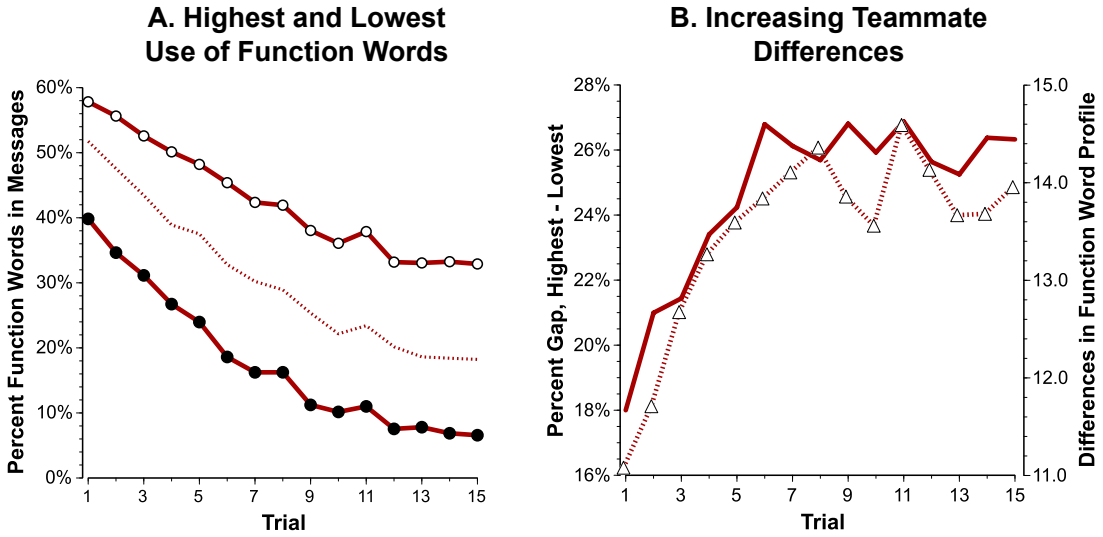


Figure A5.
Details on the Shift Away from Function Words.

No subject was told they were team leader, but teammates voted for who they thought operated most like a leader during the trials. The proportion of votes a subject received is used in Burt et al. (2021) to measure the extent to which a subject is perceived as team leader. The measure is independent of the number of trials in which a subject was the minimum user of function words (-.77 Poisson regression test statistic for leadership predicting frequency minimum user, .58 for frequency maximum, both supporting the null hypothesis, P > .4). Looking ahead to the analysis of network structure, we also found no prediction from the network structure assigned to the team (1.94 chi-square, 3 d.f., P ~ .59).

Figure S5B shows that teammates do not become more similar in their use of the function words they continue to use. Rather, they become more different. The solid line in Figure S5B describes the gap between the two solid lines in Figure S5A. The gap between teammates most and least using function words is smallest in the first trial, when language most resembles typical conversational language (Figure 2 in text). The gap expands from 18 percent on average in the first trial to 26 percent during the last trial. That is not a large difference in magnitude, but it is a large marginal difference given the lower use of function words in the last trial. And the trend is statistically significant in the wrong direction if teammates are aligning on function words: Across 945 team-trials, the gap described for trial averages by the solid line in Figure S5B increases systematically with the log of the trial (4.41 regression coefficient, 9.58 t-test, P < .001; .92 correlation in Figure S5B).

We reach the same conclusion if we compare teammates for their relative use of kinds of function words. Table S1 shows 12 LIWC categories of function words. Prepositions are distinguished from articles, are distinguished from kinds of pronouns, and so on (rows 4 through 15 in Table S1). For each subject in each trial, there are 12 p_{jk} variables measuring the percent of subject j's words during the trial that come from category k function words. The more similar the profiles for two teammates, the more similarly they use kinds of function words. The dashed line through hollow triangles in Figure S5B shows the average Euclidean distance between teammate profiles during each trial. Distance is lowest during the first trial, then increases during the next five or so trials. Distance continues at about that level for the remainder of the experiment. In other words, differences between teammates in how often they use kinds of function words increase systematically with the log of the trial: 1.04 regression coefficient, 5.71 t-test, P < .001; .88 correlation in Figure S5B).

We note a measurement detail. Euclidean distance between teammates i and j during a trial is the square root of the summed squared differences in their use of function words during the trial: $d_{ij} = [\ \sum_k (p_{ik} - p_{jk})^2\ ]^{.5}$, where p_{jk} is the percent of teammate j's words during the trial that come from category k, and summation is across the 12 LIWC categories of function words. The dashed line in Figure S5B shows the average value of d_{ij} during each trial. Profile differences

can be measured to control for weighting LIWC categories and population distributions (e.g., Goldberg et al., 2016: 1200-1201; Srivastava et al., 2018: 1354; Kovacs and Kleinbaum, 2020: 204; see Muller-Frommeyer et al., 2019, for review). We use an unweighted Euclidean distance measure because our point in Figure S5B is not subtle, and Euclidean distance is widely familiar, easy to interpret, and implicitly weights each category of function words by the frequency with which a category is used. The weighted measure originally proposed as a "Language Style Match" index, LSM, normalizes differences to give equal weight to each LIWC category of function words (Gonzales et al., 2010:9-10). LSM similarity between teammates i and j on category k is: $1 - |p_{ik}–p_{jk}|/(p_{ik}+p_{jk})$, which is averaged across K categories to get a fractional measure of similarity between the teammates. To stay close to the team discussions as experienced, we weight by the relative frequency with which categories of words are used (e.g., 11.62% of words during the first three trials are prepositions versus 1.16% that are first person plural pronouns). To be sure we did not miss something, however, we computed LSM scores as Gonzales et al. propose. LSM for function words as a whole decreases as distance in Figure S5B increases, decreasing from an average of .90 during the first trial to .55 during the final trial (the .88 correlation for log trial with distance in Figure S5B is -.94 for log trial with LSM on function words as a whole). However, certain kinds of function words show increasing LSM. Third-person pronouns stand out in particular. LSM on third-person plural pronouns such as "they" and "them," increases from an average of .54 during the first trial to .96 during the final. LSM on third-person singular pronouns such as "he" and "she" increases from an average of .52 during the first trial to a perfect 1.00 during the final trial — which is based on the near absence of words in the coordinated category. Figure 3 in the text shows that the two categories of third-person pronouns combined are rarely used, decreasing from 1.43% of words in the first trial to 0.1% of words in the final. In short, the closer to zero p_{ik} and p_{jk}, the closer to 1.00 the above LSM computation for category k.

Table S1. LIWC Data Profiles During Experiment.

	Category	Examples (from LIWC manual)	Trials 1-3	Trials 3-6	Trials 7-9	Trials 10-12	Trials 13-15	Reliability
	LINGUISTIC DIMENSIONS							
1	Total function words	it, to, no, very	48.72	39.74	33.86	28.18	24.51	0.99
2	Total pronouns	I, them, itself	14.18	11.2	9.67	7.42	6.37	0.99
3	Personal pronouns	I, them, her	8.42	6.71	5.52	4.12	3.26	0.98
4	1st pers	I, me, mine	3.88	3.23	2.76	2.13	1.73	0.96
5	1st pers plural	we, us, our	1.16	1.07	0.9	0.71	0.63	0.99
6	2nd person	you, your, thou	2.09	1.61	1.35	1.02	0.74	0.99
7	3rd pers	she, her, him	0.74	0.53	0.26	0.13	0.04	0.99
8	3rd pers plural	they, their, they'd	0.55	0.27	0.25	0.14	0.13	1.00
9	Impersonal pronouns	it, it's, those	5.76	4.49	4.15	3.3	3.11	0.98
10	Articles	an, a, the	7.09	4.31	2.65	1.97	1.16	1.00
11	Prepositions	to, with, above	11.62	10.11	8.62	7.77	6.75	0.99
12	Auxiliary verbs	am, will, have	10.31	8.9	7.95	6.26	5.14	0.99
13	Common Adverbs	very, really	2.9	2.22	1.77	1.34	1.43	0.98
14	Conjunctions	but, and, whereas	3.81	3.42	3.05	2.72	2.14	0.99
15	Negations	not, no, never	1.78	2.24	2.62	2.67	2.72	0.96
	OTHER GRAMMAR							
16	Common verbs	eat, come, carry	19.94	19.29	19.63	18.62	18.27	0.97
17	Common adjectives	free, happy, long	2.91	2.52	1.95	1.82	1.46	0.97
18	Comparisons	best, greater, after	2.72	1.59	1.07	0.75	0.43	0.98
19	Interrogatives	when, how, what	0.98	0.79	0.73	0.4	0.29	0.96
20	Numbers	thousand, second	6.57	7.7	7.8	7.45	7.2	0.99
21	Quantifiers	many, few, much	1.26	1.04	0.88	0.74	0.53	0.99
	AFFECTIVE PROCESSES							
22	Affective processes	happy, cried	2.41	2.65	2.46	2.31	2.21	0.96
23	Positive emotion	love, nice, sweet	2.07	2.21	2	1.92	1.81	0.95
24	Negative emotion	hurt, ugly, nasty	0.34	0.44	0.46	0.39	0.4	0.97
25	Anxiety	fearful, worried	0.03	0.03	0.04	0.02	0.01	1.00
26	Anger	kill, hate, annoyed	0.05	0.08	0.1	0.09	0.07	0.95
27	Sadness	crying, grief, sad	0.1	0.12	0.14	0.14	0.14	0.94
28	**SOCIAL PROCESSES**	mate, talk, they	10.25	10.01	8.73	7.38	5.81	0.98
29	Family	dad, daughter, aunt	0.01	0.02	0.02	0.04	0.02	1.00
30	Friends	neighbor, buddy	1.05	1.62	1.44	1.21	0.85	0.99
31	Female references	girl, her, mom	0.07	0.11	0.15	0.26	0.21	1.00
32	Male references	boy, his, dad	2.74	3.98	3.49	3.23	2.35	0.99
33	**COGNITIVE PROCESSES**	cause, know, ought	7	5.69	5.09	4.33	3.92	0.99
34	Insight	know, think	1.74	1.28	1.12	0.69	0.71	0.98
35	Causation	effect, because	0.42	0.44	0.31	0.28	0.22	0.99
36	Discrepancy	would, should	0.48	0.37	0.29	0.27	0.21	0.91
37	Tentative	perhaps, maybe	1.98	1.44	1.33	1.19	1.14	0.99

Table S1. LIWC Data Profiles During Experiment (continued).

	Category	Examples (from LIWC manual)	Trials 1-3	Trials 3-6	Trials 7-9	Trials 10-12	Trials 13-15	Reliability
38	Certainty	never, always	0.96	0.88	0.8	0.64	0.52	0.97
39	Differentiation	but, else, hasn't	2.19	2	2.05	2.04	2.01	0.99
40	**PERCEPTUAL PROCESSES**	look, heard, feeling	4.39	3.83	3.42	3.46	2.91	0.98
41	See	saw, seen, view	2.73	1.99	1.82	1.92	1.7	0.98
42	Hear	hearing, listen	0.99	1.14	1.04	0.97	0.65	0.98
43	Feel	touch, feels	0.73	0.81	0.6	0.62	0.69	0.99
44	**BIOLOGICAL PROCESSES**	eat, blood, pain	5.48	7.56	8.15	9.37	10.57	0.99
45	Body	hands, spit, cheek	5.38	7.4	7.96	9.15	10.27	0.99
46	Health	flu, pill, clinic	0.04	0.04	0.03	0.03	0.06	0.99
47	Sexual	love, incest, horny	0.01	0.04	0.04	0.06	0.06	0.99
48	Ingestion	eat, pizza, dish	0.07	0.1	0.13	0.16	0.24	0.99
49	**DRIVES**		5.38	5.49	5.47	6.37	6.78	0.97
50	Affiliation	friend, social, ally	1.9	1.58	1.37	1.27	1.12	0.95
51	Achievement	success, better, win	0.56	0.48	0.4	0.39	0.37	0.97
52	Power	bully, superior	2.76	3.15	3.52	4.63	5.14	0.97
53	Reward	prize, benefit, take	0.38	0.48	0.47	0.41	0.48	0.99
54	Risk	doubt, danger	0.07	0.08	0.06	0.05	0.06	0.99
	TIME ORIENTATION							
55	Past focus	ago, did, talked	2.82	3.7	3.79	3.47	2.99	0.97
56	Present focus	today, is, now	13.43	11.14	10.03	8.05	7.65	0.99
57	Future focus	may, will, soon	0.64	0.59	0.47	0.43	0.31	0.97
	RELATIVITY							0.98
58	Relativity	bend, exit, area	14.34	16	17.22	18.82	19.25	0.98
59	Motion	car, go, arrive	1.48	2.3	3.31	4.58	4.94	0.98
60	Space	in, thin, down	11.27	12.24	12.71	13.11	13.16	0.98
61	Time	until, season, end	1.77	1.61	1.31	1.25	1.38	0.97
	PERSONAL CONCERNS							
62	Work	majors, xerox, job	0.45	0.7	0.67	1.16	1.15	0.99
63	Leisure	chat, movie, cook	0.48	0.65	0.86	1.08	1.23	0.96
64	Home	landlord, kitchen	0.02	0.02	0.04	0	0.01	0.94
65	Money	cash, owe, audit	0.02	0.04	0.03	0.03	0.02	0.99
66	Religion	church, alter	0.54	1.13	1.66	2.4	2.92	0.99
67	Death	coffin, kill, bury	0.1	0.17	0.36	0.42	0.58	0.97
68	**INFORMAL LANGUAGE**		3.88	3.87	3.69	3.26	3.34	0.94
69	Swear words	fuck, damn, shit	0.07	0.1	0.22	0.2	0.26	0.98
70	Netspeak	lol, thx, btw	0.82	0.95	0.91	0.98	1.04	0.95
71	Assent	OK, yes, agree	2.87	2.75	2.63	2.4	2.41	0.94
72	Nonfluencies	hm, umm, er	0.31	0.22	0.18	0.12	0.13	0.95
73	Fillers	youknow, imean	0.02	0.04	0.02	0.01	0.02	0.96

JARGON DATA

We each coded the message texts for jargon independently so we could assess reliability. We found slightly different strategies appropriate for the task, which warrants attention in future research. We will refer to the two strategies as “literal” versus “figurative.” The words are more extreme than the actual differences between the strategies used, but the words accurately capture the nature of the difference between the strategies. The literal strategy was to code exact phrases subjects used to reference symbols in the final trials. The figurative strategy was to code key words subjects used to identify and distinguish symbols in the final trials.

To illustrate the difference between the strategies, here are the results of the literal coding strategy for a team referencing the “sitting” symbol in Figure 4:

“knee” (used in one message),

“knee up” (used in three messages),

“knees up” (used in one message),

“man with knee” (used in six messages), and

“guy with knee” (used in two messages).

The coding is a roster of specific terms used in messages during the final trials. The term used most often was coded the team jargon for the symbol: “man with knee.” The figurative coding for the same team referencing the same symbol is “knee.” Knee as a jargon term is short, is used often by the teammates, and is not used by the team to reference any of the other five symbols. When a teammate saw the word “knee,” he or she knew immediately it was a reference to the “sitting” symbol — regardless of adjacent modifiers such as “up”, or “man with,” or “guy with.”

We decided to use the results of the figurative strategy for its closer analogy to jargon as the term is used in the paper, but extensive overlap between the results of the two strategies is evidence of reliability reassuring for future research, so we offer a few details here.

Table S2, on the next page, is a tabulation of all 288 symbols (48 teams referencing six symbols). Rows distinguish number of possible jargon terms generated by the literal coding strategy. For example, the above “man with knee” example shows five possible jargon terms. The first content column in Table S2 shows that there are 43 instances in which the literal strategy generated five possible jargon terms. There are 72 instances in which the literal strategy generated a single possible jargon term, 68 instances in which the strategy generated two possible jargon terms, and so on.

The next column, “Mean Max Percent of Messages,” shows the extent to which messages are concentrated in one of the possible jargon terms. In the above example, “man with knee” is used in six messages, which is more frequent than any of the other four literal jargon terms.

There are 13 messages in which literal jargon terms occurred, so the maximum percent of messages containing the jargon term most often used is six divided by 13, or 46%. The "man with knee" example contains five possible jargon terms. If each possible jargon term occurred in the same number of messages, the maximum portion would be 20%. The fact that 46% is considerably larger than 20% shows that the most frequently used jargon term occurs in a disproportionate number of messages. The second column in Table S2 shows that, on average, subjects focused on one literal jargon term. When there is only one possible term, that one occurs in 100% of messages containing jargon. When there are two possible terms, the more frequent one occurs in 78% of messages on average, which makes the alternative term clearly secondary. When there are three possible terms, the most frequent occurs in 66% of messages on average, leaving a third of messages for the other two alternatives.

Table S2. Unreliability Is Concentrated in Extreme Cases.

Number of Literal Jargon Terms	N Symbols	Mean Max Percent of Messages	Reliability	
			Codings Do Not Match	Codings Match (%)
One	72	100%	0	72 (100%)
Two	68	78%	1	67 (99%)
Three	58	66%	1	57 (98%)
Four	27	54%	1	26 (96%)
Five	43	38%	16	27 (63%)
Six or more	20	20%	12	8 (40%)
Total	288	69%	31	257 (89%)

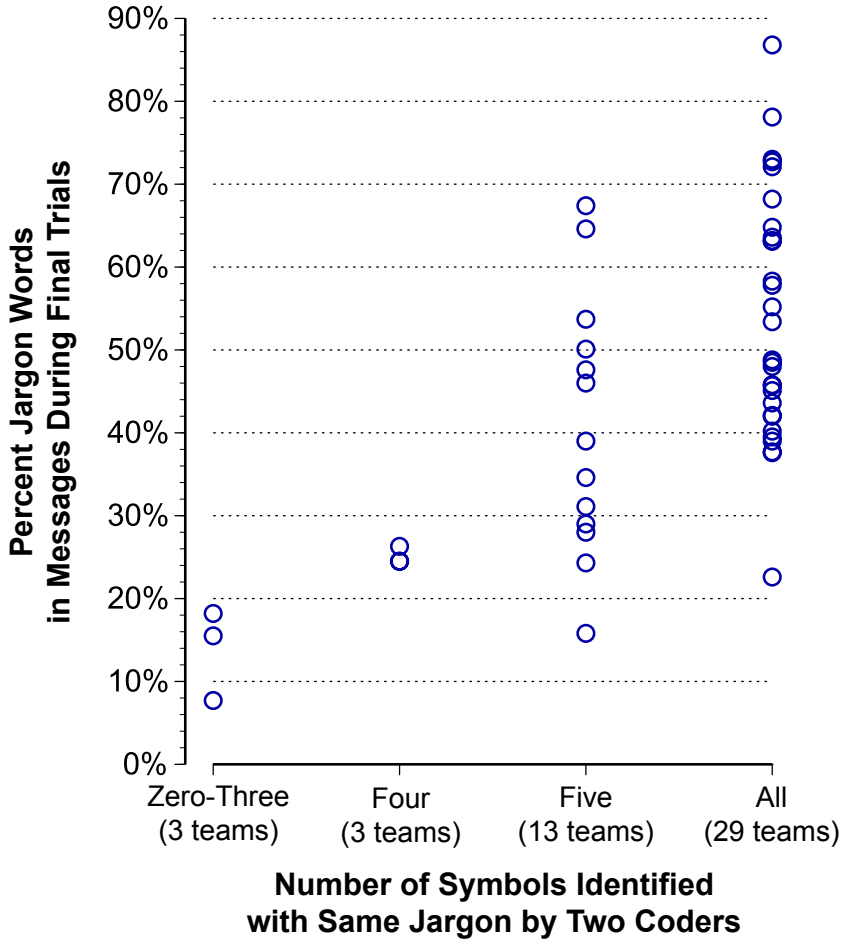
Given teammates focused on one literal jargon terms, it is not surprising that the literal and figurative coding strategies often generate the same jargon terms. The final column in Table S2 shows the frequency with which the jargon term generated by the figurative strategy is the same as the most frequent jargon term generated by the literal strategy. In the above example, "man with knee" occurs six times so it is the jargon term generated by the literal strategy. The figurative strategy generated "knee." The two strategies do not match, which adds one observation to the "Codings Do Not Match" column in Table S2.

Notice how often the alternative codings do match. When there is one literal jargon term, the two strategies match every time (100%). When there are two possible literal jargon terms, there is only one instance of mismatch (99% match). When there are three possible, there is again only one instance of mismatch (98% match). The match between literal and figurative coding does not deteriorate until there are five or more possible literal jargon terms. Even then, the figurative coding can be secure. Of the 43 instances in which literal coding generated five possible jargon terms, the two coding strategies match on 27 team symbols (63%) and do not

match on 16, one of which is the “man with knee” example, which is well represented by the “knee” figurative coding.

The concentration of unreliability in extreme cases is further apparent when we aggregate to the team level for the analysis in the paper. Four team levels of jargon reliability are distinguished on the horizontal axis of Figure S6: 29 teams in which the two coding strategies identified the same jargon terms for all six symbols, 13 teams in which the strategies agreed on jargon for five of the six symbols, 3 teams in which the strategies agreed on jargon for four symbols, and a set of 3 teams in which reliability is low (two teams in which the strategies generate inconsistent jargon for three of the six symbols, and one team in which the literal strategy generated no jargon terms, so all six symbols are coded as not matching — that one team is the one whose messages are summarized in Figure 7C).

Figure S6. Jargon Reliability Is Higher for Teams that Reach Higher Convergence on Jargon



Two points are illustrated in Figure S6: First, reliability increases with the extent to which a team converged on jargon terms. Each dot is a team, located on the vertical axis by the percent jargon in team messages during the final trials. The graph shows a positive association between the number of symbols assigned similar jargon terms by the two coding strategies (horizontal) and the percent jargon in team messages (vertical).

Second, the association is asymmetric. When reliability is low within a team (two or more mis-matches between the two coding strategies), it is always in a team that did not converge on

jargon (all dots are under 30% in first two columns in Figure S6). When reliability is high within a team (one or no mis-matches between the two coding strategies), average percent jargon is high, but there are also teams that did not converge on jargon terms (wide distribution of dots in last two columns in Figure S6). In sum, unreliability is concentrated in a few teams that converged least on jargon terms (messages for one of which is in Figure 7C).

Tying the above points together, Table S3 shows the results of a logit model predicting when the two coding strategies generate the same jargon term. The strongest predictor is the number of alternatives generated by literal coding, which is the rows in Table S2. Match between coding strategies is less likely in messages that contain more alternatives (-4.53 test statistic). The other strong predictor is team convergence on jargon (vertical axis in Figure S6). The more prevalent the jargon in a team's messages during the final trials, the more likely the coding strategies identify the same jargon (2.71 test statistic). Reliability does not differ between symbols (3.71 summary chi-square, 5 d.f., P ~ .59), nor between the four network structures imposed on teams (2.27 summary chi-square, 3 d.f., P ~ .52).

Table S3. Predicting Reliability.

Predictors	Logit Coefficient	Z-Score Test Statistic
Literal Options	-1.44	-4.53 ***
Percent Jargon	.05	2.71 **
Symbol		
Kicker (reference)	—	
Priest	.26	.25
Falling	1.63	1.49
Kneeling	.52	.49
Bunny	.39	.46
Sitting	-.17	-.20
Network Structure		
Clique (reference)	—	
Wheel	-.49	-.54
DB Network	.44	.51
CB Network	.14	.16
Intercept	5.79	

NOTE — Logit regression predicting symbols (six symbols in each of 48 teams, N = 288) for which literal and figurative coding strategies identify the same jargon term. Test statistics are adjusted down for autocorrelation between symbols in same team (Stata "cluster" option). "Literal Options" is the number of possible jargon terms generated by the literal strategy (rows in Table S2). "Percent Jargon" is the percent of message words in final trials that are jargon (vertical axis in Figure S6). The six symbols are given in Figure 4 and the four team network structures are given in Figure 6). Pseudo R^2 is .51. * P < .05 ** P < .01 *** P < .001

Table S4. Selected Messages Leading to "Angelmouse."

- a sitting person with a bow in their hair
- sitting down with two wings
- angel with wings
- mountain with a bow on the top
- someone crouching
- a person who looks like kneeling in a kimono
- theres a mouse shaped one
- mouse with bow
- sitting girl with a bow in her hair
- sitting girl with bow
- mouse with bow
- mouse with ears on right
- bow on right
- girl sitting on the ground with bow on her head
- sitting girl with bow
- pointy nose looking left
- 5 sided body and a little triangle tail to left
- angel with triangle wings
- girl with bow has triangle body, square head and ears
- mouse
- bow on right
- triangles
- mouse
- kimono person/dragon
- angel with triangle wings
- triangle/angel
- we need consistent names (in trial 8)
- angel = triangle wings
- oh wait, maybe the mouse I see is an angel
- it has a triangle for a nose and a square and triangle on the top (looks like a bow to the right)
- thats angel = mouse
- ohh
- everyone angel = mouse
- triangle = angel = mouse
- it doesn't even look like a mouse
- pick that one!
- angel/mouse (in trial 9)
- mouse/angel (in trial 10)
- angel/mouse (in trial 11)
- mouse
- mouse (in trial 12)
- angel mouse
- angel mouse (in trials 13-14)
- angel mouse (in trial 15)
- angelmouse
- angelmouse
- angelmouse
- angelmouse

NOTE — These are selected messages within the team that settled on "angelmouse" as the jargon label for the "kneeling" symbol in Figure 5. Messages to the left are from trials 1-8. Messages to the right are from trials 8-15, as indicated in parentheses.

SELECTIVE RETENTION

To better understand the negative association between jargon and message concentration, we separated the teams into high versus low concentration to see how jargon develops differently in the two categories of teams. Figure S7 displays word counts over time grouped into five periods of three trials each. Three kinds of words are distinguished: function words, jargon words, and non-jargon content words. Teams are separated at the mean level of message concentration (68.6 points on the horizontal axis in Figure 9).

We take three points from Figure S7. First, the learning curves displayed as message volume in Figure 1, and word proportions in Figure 3, are apparent here in word counts: Teams learn to complete their task using fewer words, and the proportion function words decreases as the proportion content words increases.

To create word counts in Figure S7 we assume that words not in the LIWC dictionary are proportionally function and content words. We do not anticipate a bias from the assumption because we only use the function-content contrast as a frame of reference for the stable counts of jargon words, which we have regardless of the LIWC dictionary. Still, the assumption needs to be explicit. In Figure 3, we use LIWC software to separate three categories of words: percent function words, percent content words, and percent words not in the LIWC dictionary. The percent function and content words are percentages of the words found in the LIWC dictionary. For Figure S7, we aggregate across trials within periods to define percentages on larger word counts, then convert percentages into word counts so we can compare them to the counts of jargon words. Given the total word count, WC, in a team's messages during a period, we define the number of function words to be WCf = WC * (Pf / Pd), where Pd is average trial percent of WC found in the LIWC dictionary, and Pf is the average trial percent of those dictionary words that are function words. The number of content words is then WC – WCf. Thus, we smooth percentages across trials within a period and words not in the LIWC dictionary are allocated to function or content depending on the size of Pf during a period.

Second, concentrating messages in one teammate makes that person a bottleneck for communication, so fewer words get exchanged. The height of the bars in Figure S7B are lower than the corresponding bars in Figure S7A. Teams high in message-concentration exchange an average of 287 words per trial, 854 words per period. Low concentration teams exchange respective averages of 437 and 1300 words (respective t-tests of -4.54 and -4.53 with clustering to adjust down for autocorrelation between a team's activity across trials/periods, $P < .001$).

Third, and most important, the balance of jargon to other content words is less about increasing use of jargon than it is about decreasing use of non-jargon. Jargon words are a relatively constant presence in team discussion, even in the initial trials. This is displayed by the dark areas at the base on each column in Figure S7 being about the same height over time.

There are exceptions in which jargon terms emerge as a synthesis over time, such as “angelmouse” in Table S4, but it is more usual for jargon to enter team discussion early. It is not clear at their entry that the future jargon terms will become jargon because they are buried in numerous function words and other content words competing to become jargon. Teammates initially use a diverse assortment of words to describe the symbols in Figure 4. Jargon emerges as a by-product of teammates neglecting alternatives. Convergence on jargon is more about discontinuing the use of alternatives than it is about creating something new, again acknowledging the occasional “angelmouse.” In short, convergence on jargon is about selective retention.

Low-concentration teams are more thorough in discarding alternatives — as is apparent from the white areas of the bars in Figure S7A shrinking from wide in the initial trials down to a small sliver in trials 10 through 15. White areas in the Figure S7B bars also shrink over time, but not as much, and not as completely. Non-jargon content words are used about as often as jargon words in the Figure S7B final trials. The inset graph in Figure S7 shows percent non-jargon content similarly about 90 percent in initial messages within high- and low-concentration

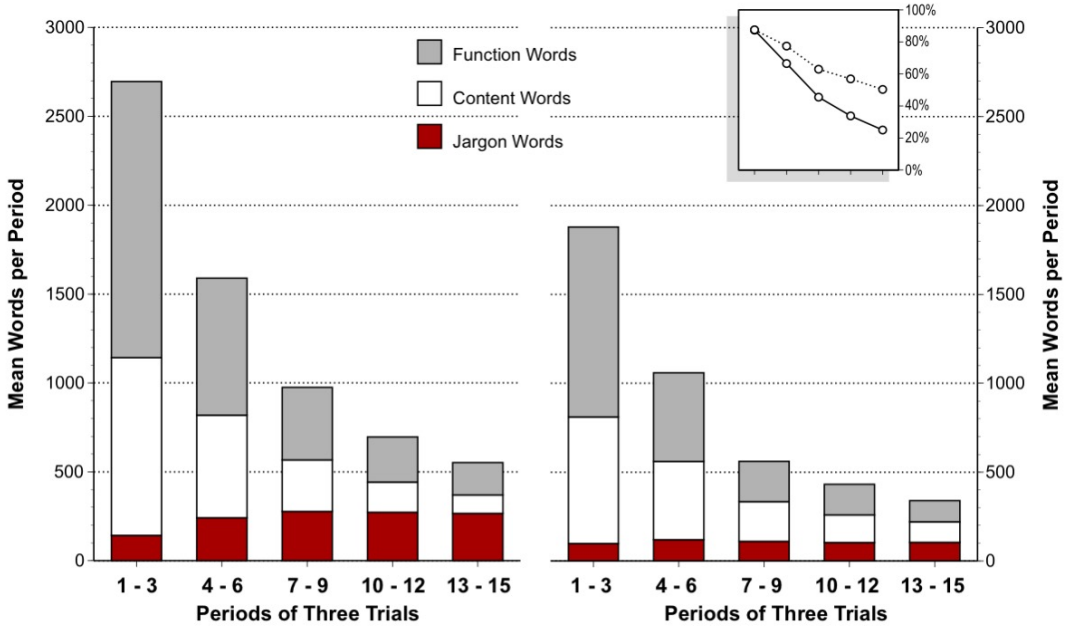


Figure S7. Jargon Use Over Time

Inset graph shows average percent of content words in each period that are not jargon.
Solid line is low-concentration teams.

teams. By the time they reach the final trials, a large gap has opened up between high- versus

low-concentration teams. The low-concentration teams are down to 25 percent of content words versus 50 percent in high-concentration teams (4.82 t-test across the 48 teams, P < .001). Content words that are an alternative to jargon have not been as thoroughly eliminated in teams that concentrate messaging in one teammate.

NETWORK BEHAVIOR, NOT PERCEPTION

Research on the Bavelas-Leavitt-Smith experiment emphasized the clarity of leadership defined by network centrality. We quote Leavitt (1951), and Cohen et al. (1961) in the text. Cohen et al. (1961:428) were particularly emphatic: "“The more a leader is clearly recognized and agreed upon …, the more likely will other members accept influence attempts by him: procedures, answers, etc. Less energy and time will be spent by other members in duplicating the functions of the leader: figuring out answers for themselves, checking on others (once the leader has approved information by passing it on), and trying to set up variations in problem-solving procedures according to their own idiosyncratic evaluations.”

Teammate agreement on perceived leadership is the dependent variable in Burt et al.'s (2021) analysis of the renovated experiment, so we have it well measured and can test. We conclude that perceived leadership is correlated in the expected way with network structure, but it is uncorrelated with team convergence on jargon so we do not discuss it in the paper.

Here is the measure of perceived leadership: No subject was designated a team leader, but subjects were asked at the end of the experiment: "Did your group have a leader? If so, who?" Some responded that there was no leader. Some cited a teammate. Some cited two. Each subject on a team is perceived to be leader to the extent that he or she received a high percentage of his or her team votes. For the 240 subjects in teams during the final trials, the "perceived leader" percentage varies from a minimum of zero for subjects who received no leader citations, up to a maximum of 100 percent for subjects who received all team leader citations, around a mean of 13.1 percent.

The perception of leadership is associated with the network structure to which a subject was assigned. Subjects assigned to the hub position in a WHEEL network receive 83.5 percent of their team leader cites on average, while no subject assigned to one of the subordinate positions was ever cited.

Behavioral network structure is a stronger predictor. The more concentrated messages are in one teammate, the more likely that teammate is perceived to be leader (Burt et al., 2021). At the team level, message concentration along the horizontal axis in Figure 9 is correlated .50 with perceived leadership. However, perceived leadership has no correlation with percent jargon in the final trials (-.13 correlation, -0.59 t-test, P ~ .56).

In short, the jargon association with network structure is due to network behavior, not to the perception of leadership associated with network behavior. We have no additional evidence to suggest that the lack of jargon in centralized networks is due to teammates consciously reacting to, or resenting, the concentration of messages in one teammate.

ENDOGENOUS VARIATION IN NETWORK PREDICTORS

Table S5 on the last page of this supplement shows how our three network predictors change across trials. The point of the table is to show the extent to which the three network predictors differ in terms of endogenous variation. In the first trial, for example, the 48 teams had an average 62.65% concentration within the networks to which they were randomly assigned (Figure 6), an average 69.86 percent message concentration in terms of messages actually sent during the trial (Figure 9), and teammates sent an average 15.33 messages per minute during the first trial (Figure 11). The bottom row shows the relative extent to which each predictor changes as teammates become more experienced with one another. We obtained the bottom-row results by regressing the column network variable across 47 dummy variables distinguishing the 48 teams for a data deck of 717 team-trial observations (13 x 48 teams + 47 teams in trial 14 + 46 teams in trial 15, see footnote 6 in the paper).

Assigned concentration is exogenous to behavior in the experiment. The network to which a subject is randomly assigned does not change during the experiment (0.0% variation within teams across trials). Behavioral concentration has the potential to change across trials, but once a team has a pattern of messaging, the pattern holds steady across the experiment: 5.2% of variation between teams in behavioral concentration is within teams across trials. Whatever the learned network behavior is that subjects brought into the experiment, that behavior affects network structure right away, then holds steady across the experiment. For example, the bossy teammate to took over the CLIQUE team in Figure 7C and Figure 10C is bossy in other trials too. Behavioral concentration within that team does not increase or decrease significantly across trials (-0.95 t-test across the 15 trials, P ~ .36).

Message rate is where team networks become most endogenous. Almost half of the variance in message rates across trials is within teams. As teams converge on jargon, their messages can be shorter, so they can exchange more of them. It is not surprising to see higher message rates in the later trials. The final column of Table S5 shows the correlation between a team's message rate during a row trial and the team's percent jargon in the final trials. The correlation bounces up and down, but gradually increases toward the final trials (the 15 results in the final column are correlated .44 with trial number).

Therefore, we use message rates in the early trials to predict percent jargon in the final trials. Endogeneity is not removed, but such measurement gives us more confidence in the

Figure 11 association as a fact. We define “early” by a transition in message rates. Panel analysis of percent jargon reveals a distinction between the first two periods versus the later three (third point in the text about the correlation matrix in Figure 5). We did a similar analysis of message rates in which we distinguish the three trials within the second and third periods, looking for a more precise transition, given the more reliable electronic data we have on message rates. The second largest principal component loads on the first seven trials, with subsequent trials negative on the component. Our message-rate predictor in the paper is a team’s average message rate across the initial seven trials.

REFERENCES

Burt, R.S., Reagans, R.E., & Volvovsky, H.C. 2021. Network brokerage and the perception of leadership. ***Social Networks*** 65: 33-50.

Cohen, A. M., Bennis, W. G., & Wolkon, G. H. 1961. The effects of continued practice on the behaviors of problem-solving groups. ***Sociometry*** 24(4): 416-431.

Goldberg, A., Srivastava, S.B., Manian, V.G., Monroe, W., & Potts, C. 2016. Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. ***American Sociological Review*** 81(6): 1190-1222.

Gonzales, A.L., Hancock, J.T., & Pennebaker, J.W. 2010 Language style matching as a predictor of social dynamics in small groups. ***Communications Research*** 31(1) 3–19.

Kovacs, B., & Kleinbaum, A.M. 2020. Language-style similarity and social networks. ***Psychological Science*** 31(2): 202-213.

Leavitt, H.J. 1951. Some effects of certain patterns of communications on group performance. ***Journal of Abnormal and Social Psychology*** 46(1): 38-50.

Muller-Frommeryer, L.C., Frommeyer, N.A.M., & Kauffeld, S. 2019. Introducing rLSM: An integrated metric assessing temporal reciprocity in language style matching. ***Behavior Research Methods*** 51(3): 1343-1359.

Srivastava, S.B., Goldberg, A., Manian, V.G., & Potts, C. 2018. Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. ***Management Science*** 64(3): 1348-1364.

Table S5. Network Structure and Velocity across Trials.

Trial	Assigned Concentration	Behavioral Concentration	Message Rate	Correlation with Percent Jargon
1	62.65	69.86	15.33	.44
2	62.65	70.31	17.10	.31
3	62.65	71.11	17.92	.44
4	62.65	71.32	18.00	.40
5	62.65	69.21	18.87	.57
6	62.65	70.83	19.66	.41
7	62.65	70.14	21.21	.58
8	62.65	69.12	21.72	.43
9	62.65	70.09	22.17	.43
10	62.65	68.79	23.42	.40
11	62.65	69.89	23.50	.47
12	62.65	69.28	25.64	.48
13	62.65	69.36	26.21	.55
14	63.13	69.64	27.51	.51
15	62.33	70.70	28.06	.50
Percent Variation within Teams across Trials	0%	5.2%	42.5%	

NOTE — There are 48 teams in each row, except the 14th trial in which one team dropped out, and the 15th trial in which another team dropped out. Percent variation within teams across trials is $100 - 100*R^2$ from a regression predicting the column variable from 47 dummy variables distinguishing teams in 717 team-trial observations.