ONLINE APPENDIX (NOT FOR PUBLICATION)

A. Summary Statistics and Balance Tests

We present summary statistics of individual characteristics by gender in the chat treatments (Table A1) and the control treatments (Table A2). Men and women do not differ significantly across any of the demographic characteristics in the Control treatments. Table A3 shows that there is balance on demographics across the chat and the control treatments.

Table A1: Comparison of Demographic Characteristics and Experimental Variables by Gender in the Chat Treatment

	Male	Female	Absolute Difference	t-test	Mann-Whitney U test
Demographics					
Age (Years)	24.58	24.42	0.16	0.859	0.968
Never Married (%)	85.0%	89.7%	0.05	0.308	0.307
White (%)	36.0%	36.4%	0.00	0.947	0.947
Black (%)	14.0%	18.7%	0.05	0.365	0.364
Asian (%)	39.0%	40.2%	0.01	0.862	0.862
Hispanic (%)	12.0%	4.7%	0.073*	0.055	0.056
Native English (%)	79.0%	77.6%	0.01	0.804	0.804
Born in US (%)	57.0%	57.9%	0.01	0.891	0.891
US Citizen (%)	64.0%	65.4%	0.01	0.832	0.831
Income > $65,000 (%)	52.0%	46.7%	0.05	0.451	0.450
Currently a Student (%)	99.0%	97.2%	0.02	0.349	0.347
Undergraduate (%)	52.0%	45.8%	0.06	0.375	0.373
Primary Field of Study:					
Arts & Humanities	12.0%	12.1%	0.00	0.974	0.974
Social Sciences & Business	37.0%	29.0%	0.08	0.221	0.220
Natural Sciences & Engineering	33.0%	41.1%	0.08	0.229	0.228
Other Major	18.0%	17.8%	0.00	0.964	0.964
Used Real Name (%)	93.0%	88.8%	0.04	0.296	0.295
GPA (Points)	3.51	3.47	0.04	0.440	0.736
Final Payment ($)	27.81	25.60	2.212*	0.087	0.092
Number of obs.	100	107			
Experimental Variables					
Points in Pre-Group Stage	14.06	13.14	0.92	0.368	0.486
Confidence in Pre-Group Stage	7.73	7.59	0.14	0.351	0.080
Total Probability Chosen	32.98	33.66	0.68	0.425	0.359
Points in Post-Group Stage	18.04	17.99	0.05	0.960	0.913
Number of obs.	400	428			

Notes: GPA averages are based on 95 responding males and 103 responding females. Columns 4 and 5 report p-values. Two-sample tests of proportions for dummy variables produce similar results. Significance levels: *10 percent, **5 percent, ***1 percent.

Table A2: Comparison of Demographic Characteristics by Gender in the Control Treatments

	Male	Female	Absolute Difference	t-test	Mann-Whitney U test
Age (Years)	24.10	23.74	0.37	0.530	0.874
Never Married (%)	87.0%	89.2%	0.02	0.495	0.494
White (%)	35.4%	42.6%	0.07	0.150	0.150
Black (%)	11.5%	18.5%	0.07	0.054	0.054
Asian (%)	43.2%	36.4%	0.07	0.171	0.171
Hispanic (%)	11.5%	8.7%	0.03	0.372	0.371
Native English (%)	78.6%	76.4%	0.02	0.599	0.599
Born in US (%)	55.7%	61.5%	0.06	0.247	0.247
US Citizen (%)	63.5%	67.7%	0.04	0.391	0.391
Income > $65,000 (%)	45.8%	46.2%	0.00	0.950	0.950
Currently a Student (%)	97.4%	97.4%	0.00	0.980	0.980
Undergraduate (%)	51.0%	48.7%	0.02	0.649	0.648
Primary Field of Study:					
Arts & Humanities	9.4%	12.8%	0.03	0.282	0.281
Social Sciences & Business	39.1%	33.3%	0.06	0.242	0.242
Natural Sciences & Engineering	34.9%	40.0%	0.05	0.301	0.300
Other Major	16.7%	13.8%	0.03	0.442	0.441
Used Real Name (%)	93.2%	89.7%	0.03	0.221	0.220
GPA (Points)	3.46	3.44	0.02	0.698	0.794
Final Payment ($)	26.59	26.17	0.42	0.646	0.769
Number of obs.	192	195			

Notes: GPA averages are based on 179 responding males and 185 responding females. Columns 4 and 5 report p-values. Two-sample tests of proportions for dummy variables produce similar results.

Table A3: Balance on Demographics Comparing Chat and the Control Treatments
Chat vs No Chat

	Treatment		Absolute		Mann-Whitney
	Chat	Control	Difference	t-test	U test
Age (Years)	24.50	23.92	0.58	0.262	0.259
Female (%)	51.7%	50.4%	0.01	0.763	0.762
Never Married (%)	87.4%	88.1%	0.01	0.811	0.811
White (%)	36.2%	39.0%	0.03	0.506	0.505
Black (%)	16.4%	15.0%	0.01	0.645	0.645
Asian (%)	39.6%	39.8%	0.00	0.966	0.966
Hispanic (%)	8.2%	10.1%	0.02	0.459	0.459
Native English (%)	78.3%	77.5%	0.01	0.836	0.836
Born in US (%)	57.5%	58.7%	0.01	0.784	0.783
US Citizen (%)	64.7%	65.6%	0.01	0.827	0.827
Income > $65,000 (%)	49.3%	46.0%	0.03	0.446	0.446
Current a Student (%)	98.1%	97.4%	0.01	0.619	0.618
Undergraduate (%)	48.8%	49.9%	0.01	0.803	0.802
Primary Field of Study:					
Humanities	12.1%	11.1%	0.01	0.725	0.725
Social Sciences	32.9%	36.2%	0.03	0.419	0.419
Natural Sciences & Engineering	37.2%	37.5%	0.00	0.949	0.948
Other Major	17.9%	15.2%	0.03	0.408	0.407
Used Real Name (%)	90.8%	91.5%	0.01	0.789	0.789
GPA (Points)	3.49	3.45	0.04	0.240	0.126
Final Payment ($)	26.67	26.38	0.29	0.711	0.989
Number of obs.	207	387			

Notes: Control treatments include Answer and Answer + Confidence treatments. GPA averages are based on 198 and 364 responses for the Chat and Control treatments, respectively. Columns 4 and 5 report p-values. Two-sample tests of proportions for dummy variables produce similar results.

B. Robustness Checks

B1: Nonlinear Specifications

Tables B1-B3, B4, B5, and B6 present parallel analysis to Tables 2-4, 6, 7, and 9 in the paper, respectively, using the ordered probit specifications. These alternative specifications produce qualitatively similar results to those from the linear probability specifications in the paper.

Table B1: The Determinants of the Probability of Being Chosen as the Group Representative in the Post-Group Stage, Ordered Probit Estimates

	All Groups			Mixed-Gender Groups		
Sample	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.00542	0.145	0.134	-0.000796	0.220*	0.198*
	(0.104)	(0.110)	(0.103)	(0.119)	(0.132)	(0.120)
Gender Congruence of Question	0.324**	-0.0903	-0.0829	0.395**	-0.0820	-0.0748
	(0.151)	(0.166)	(0.156)	(0.176)	(0.199)	(0.186)
Own Gender Share in Group	0.245*	0.0843	0.0995	0.491**	0.341	0.338
	(0.139)	(0.144)	(0.137)	(0.248)	(0.255)	(0.241)
KG Treatment			0.0317			0.0569
			(0.154)			(0.181)
Female x KG			-0.125			-0.172
			(0.146)			(0.171)
Gender Congruence x KG			0.394*			0.466*
			(0.219)			(0.260)
Own Gender Share x KG			0.102			0.128
			(0.199)			(0.352)
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using an ordered probit model (marginal effects). Sample is restricted to chat treatment data only. Dependent variable mean is 33.33. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, dummy for whether the US is the country of citizenship and birth; and controls for pre- and post-chat individual performance, as well as performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B2: The Determinants of Probability of Favorable Self-Ranking in the Post-Group Stage, Ordered Probit Estimates

Sample	All Groups			Mixed-Gender Groups		
	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.0928	0.0580	0.0722	-0.0326	0.233	0.196
	(0.127)	(0.136)	(0.132)	(0.143)	(0.164)	(0.158)
Gender Congruence of Question	0.140	-0.180	-0.112	0.209	-0.0600	-0.0326
	(0.177)	(0.205)	(0.196)	(0.203)	(0.248)	(0.238)
Own Gender Share in Group	0.440**	0.143	0.139	0.597**	0.389	0.413
	(0.181)	(0.179)	(0.175)	(0.296)	(0.331)	(0.321)
KG Treatment			-0.151			-0.0861
			(0.183)			(0.205)
Female x KG			-0.167			-0.210
			(0.181)			(0.212)
Gender Congruence x KG			0.239			0.242
			(0.264)			(0.313)
Own Gender Share x KG			0.236			0.132
			(0.249)			(0.432)
Dep. Var. Mean	46.67	49.26	47.95	46.04	49.12	47.50
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using an ordered probit model (marginal effects). Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, dummy for whether the US is the country of citizenship and birth; and controls for pre- and post-chat individual performance, as well as performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B3: The Determinants of Probability of Favorable Ranking of Group Member j Relative to Group Member k in the Post-Group Stage, Ordered Probit Estimates

	All Groups			Mixed-Gender Groups		
Sample	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Gender j v k	0.133	0.144	0.132	0.130	0.184*	0.156*
	(0.0921)	(0.0975)	(0.0942)	(0.0936)	(0.0989)	(0.0948)
Gender Congruence of Question j v k	0.331**	0.0198	0.0209	0.333**	0.00290	0.0125
	(0.137)	(0.156)	(0.148)	(0.139)	(0.155)	(0.148)
Own Gender Share in Group j v k	0.194	0.0535	0.0540	0.152	-0.0149	0.0220
	(0.183)	(0.194)	(0.186)	(0.185)	(0.198)	(0.189)
KG Treatment			-0.00831			-0.0686
			(0.0960)			(0.113)
Gender j v k x KG			0.00619			-0.0181
			(0.131)			(0.132)
Gender Congruence j v k x KG			0.323			0.352*
			(0.200)			(0.201)
Own Gender Share j v k x KG			0.119			0.126
			(0.259)			(0.262)
Dep. Var. Mean	53.57	55.39	54.47	52.20	55.09	53.57
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using an ordered probit model (marginal effects). Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Performance controls the differences in pre- and post-group individual scores of j relative to kinclude pre- and post-group ability of j v. k. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B4: The Effect of Conversation Variables, Ordered Probit Estimates

	Probability of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)				Probability of Ranking j higher than k (Explanatory variables represent the difference between j and k)			
	KG	UG	KG	UG	KG	UG	KG	UG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Assertive	-0.0469	0.282**	-0.0298	0.299**	0.181*	0.229**	0.171	0.234**
	(0.142)	(0.137)	(0.144)	(0.139)	(0.106)	(0.110)	(0.107)	(0.111)
Competent	0.0260	0.0260	0.0337	0.0270	0.263**	0.188	0.249**	0.185
	(0.144)	(0.143)	(0.148)	(0.143)	(0.124)	(0.115)	(0.124)	(0.116)
Warm	-0.241	0.264*	-0.251	0.255	0.204	-0.153	0.226*	-0.150
	(0.155)	(0.155)	(0.157)	(0.160)	(0.127)	(0.121)	(0.127)	(0.123)
Open	0.125	-0.259*	0.104	-0.253*	-0.0902	0.00684	-0.0656	0.0175
	(0.159)	(0.142)	(0.162)	(0.144)	(0.129)	(0.113)	(0.132)	(0.113)
Difficult	0.0566	-0.0622	0.0597	-0.0465	0.0945	-0.0137	0.140	0.0111
	(0.137)	(0.135)	(0.142)	(0.133)	(0.163)	(0.197)	(0.167)	(0.198)
Female			-0.0987	0.0608			0.101	0.152
			(0.130)	(0.141)			(0.0972)	(0.101)
Gender Congruence of Question			0.190	-0.142			0.344**	-0.00424
			(0.182)	(0.219)			(0.141)	(0.162)
Own Gender Share in Group			0.535***	0.160			0.116	0.0648
			(0.185)	(0.183)			(0.189)	(0.203)
Dep. Var. Mean	46.67	49.30	46.67	49.30	53.68	55.47	53.68	55.47
Observations	408	402	408	402	408	402	408	402

Notes: Coefficients obtained using an ordered probit model (marginal effects). Explanatory variables represent the difference between j and k. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for gender, age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individivual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B5: The Effect of Beliefs, Ordered Probit Estimates

	Probability of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)		Probability of Ranking j higher than k (Explanatory variables represent the difference between j and k)	
	KG	UG	KG	UG
	(1)	(2)	(3)	(4)
Female	-0.112	0.141	0.0287	0.200*
	(0.131)	(0.159)	(0.0982)	(0.111)
Gender Congruence of Question	0.243	-0.303	0.318**	0.121
	(0.193)	(0.228)	(0.146)	(0.178)
Own Gender Share in Group	0.605***	-0.0360	0.231	-0.135
	(0.196)	(0.183)	(0.194)	(0.231)
Predicted Points of Member 1	0.410***	0.550***		
	(0.0518)	(0.0687)		
Predicted Points of Member 2	-0.178***	-0.232***		
	(0.0417)	(0.0442)		
Predicted Points of Member 3	-0.146***	-0.187***		
	(0.0374)	(0.0485)		
Predicted Points of Member j v. k			0.522***	0.440***
			(0.0806)	(0.0759)
Dep. Var. Mean	46.67	49.26	53.57	55.39
Observations	408	402	408	402

Notes: Coefficients obtained using an ordered probit model (marginal effects). Explanatory variables represent the difference between j and k. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for gender, age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individivual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B6: The Effect of Third-Party Beliefs, Ordered Probit Estimates

	Probability of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)		Probability of Ranking j higher than k (Explanatory variables represent the difference between j and k)	
	KG	UG	KG	UG
	(1)	(2)	(3)	(4)
Made Valuable Contributions	0.309	0.494*	0.179	0.226
	(0.342)	(0.296)	(0.184)	(0.176)
Deserves Extra Compensation	0.654**	-0.0146	-0.00253	-0.0946
	(0.331)	(0.302)	(0.196)	(0.188)
Bet on Being Chosen	0.252	0.528*	0.500***	0.252
	(0.281)	(0.281)	(0.177)	(0.164)
Female	-0.111	0.0866	0.134	0.141
	(0.132)	(0.140)	(0.0960)	(0.0990)
Gender Congruence of Question	0.153	-0.0789	0.288**	0.0517
	(0.181)	(0.212)	(0.140)	(0.160)
Own Gender Share in Group	0.498***	0.144	0.113	0.0670
	(0.186)	(0.187)	(0.189)	(0.199)
Dep. Var. Mean	46.67	49.30	53.68	55.47
Observations	408	402	408	402

Notes: Coefficients obtained using an ordered probit model (marginal effects). Explanatory variables represent the difference between j and k. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part; demographic controls for gender, age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individivual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

B2: Specifications without Demographic Controls

Tables B7-B9, B10, B11, and B12 present parallel analysis to Tables 2-4, 6, 7, and 9 in the paper, respectively, omitting demographic controls. These alternative specifications produce qualitatively similar results to those controlling for demographics represented in the paper.

Table B7: The Determinants of the Probability of Being Chosen as the Group Representative in the Post-Group Stage, No Demographic Controls

	All Groups			Mixed-Gender Groups		
Sample	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.0491	1.589	1.515	0.180	2.049	2.061
	(1.223)	(1.150)	(1.130)	(1.420)	(1.357)	(1.340)
Gender Congruence of Question	3.501*	-0.985	-1.002	4.506**	-1.255	-1.357
	(1.807)	(1.889)	(1.836)	(2.150)	(2.204)	(2.186)
Own Gender Share in Group	2.206	0.608	0.849	5.582*	3.152	3.256
	(1.674)	(1.595)	(1.573)	(2.975)	(2.824)	(2.807)
KG Treatment			0.240			0.176
			(1.729)			(2.033)
Female x KG			-1.496			-1.894
			(1.635)			(1.939)
Gender Congruence x KG			4.503*			5.899*
			(2.559)			(3.042)
Own Gender Share x KG			1.232			2.280
			(2.286)			(4.075)
R-squared	0.0636	0.108	0.0798	0.0956	0.0965	0.0950
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. Dependent variable mean is 33.33. All specifications include fixed effects for round and part and controls for pre- and post-chat individual performance, as well as performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B8: The Determinants of Probability of Favorable Self-Ranking in the Post-Group Stage, No Demographic Controls

	All Groups			Mixed-Gender Groups		
Sample	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-1.871	0.279	0.369	-0.846	1.808	1.636
	(1.871)	(1.721)	(1.713)	(2.154)	(2.140)	(2.114)
Gender Congruence of Question	1.971	-2.212	-1.389	3.650	-1.294	-0.989
	(2.713)	(2.825)	(2.781)	(3.188)	(3.474)	(3.448)
Own Gender Share in Group	5.350*	1.916	2.078	8.358*	4.837	4.919
	(2.723)	(2.393)	(2.385)	(4.698)	(4.623)	(4.577)
KG Treatment			-2.683			-2.359
			(2.696)			(3.153)
Female x KG			-2.113			-2.530
			(2.504)			(3.003)
Gender Congruence x KG			3.204			4.467
			(3.885)			(4.687)
Own Gender Share x KG			3.108			3.351
			(3.582)			(6.529)
R-squared	0.0732	0.0649	0.0641	0.0777	0.0638	0.0707
Dep. Var. Mean	46.67	49.26	47.95	46.04	49.12	47.50
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part and controls for pre- and post-chat individual performance, as well as performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B9: The Determinants of Probability of Favorable Ranking of Group Member j Relative to Group Member k in the Post-Group Stage, No Demographic Controls

Sample	All Groups			Mixed-Gender Groups		
	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Gender j v k	5.333	5.068	4.781	5.188	5.567	5.230
	(3.369)	(3.633)	(3.634)	(3.383)	(3.635)	(3.619)
Gender Congruence of Question j v k	13.33***	2.209	1.172	12.70**	2.219	0.966
	(5.085)	(5.647)	(5.649)	(5.060)	(5.657)	(5.658)
Own Gender Share in Group j v k	6.307	0.0241	0.878	5.208	-0.439	0.446
	(6.712)	(7.091)	(7.118)	(6.709)	(7.130)	(7.127)
KG Treatment			-0.436			-2.802
			(3.535)			(4.165)
Gender j v k x KG			0.751			-0.232
			(4.926)			(4.934)
Gender Congruence j v k x KG			12.48*			12.45*
			(7.545)			(7.532)
Own Gender Share j v k x KG			4.501			3.955
			(9.667)			(9.652)
R-squared	0.0578	0.0736	0.0504	0.0848	0.0642	0.0640
Dep. Var. Mean	53.57	55.39	54.47	52.20	55.09	53.57
Observations (groups)	420 (140)	408 (136)	828 (276)	318 (106)	285 (95)	603 (201)

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part. Performance controls the differences in pre- and post-group individual scores of j relative to kinclude pre- and post-group ability of j v. k. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B10: The Effect of Conversation Variables, No Demographic Controls

	Probability of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)				Probability of Ranking j higher than k (Explanatory variables represent the difference between j and k)			
	KG	UG	KG	UG	KG	UG	KG	UG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Assertive	0.318	2.951*	0.391	3.118*	6.005	9.077**	5.624	8.112**
	(2.174)	(1.754)	(2.185)	(1.785)	(4.011)	(4.102)	(3.988)	(4.055)
Competent	1.017	0.406	1.092	0.386	8.680*	9.071**	8.191*	6.983*
	(2.168)	(1.828)	(2.164)	(1.833)	(4.660)	(4.031)	(4.639)	(3.994)
Warm	-2.340	3.604*	-2.289	3.628*	5.117	-4.813	5.863	-4.739
	(2.331)	(1.927)	(2.298)	(1.954)	(4.835)	(4.713)	(4.795)	(4.548)
Open	1.741	-3.011*	1.284	-2.893	-1.962	2.286	-1.028	1.648
	(2.394)	(1.807)	(2.415)	(1.822)	(4.792)	(4.075)	(4.873)	(3.991)
Difficult	0.0706	-0.472	0.0492	-0.274	2.663	5.224	4.482	2.358
	(2.016)	(1.863)	(2.047)	(1.857)	(6.047)	(6.859)	(6.106)	(6.838)
Female			-1.969	-0.0749			4.412	5.349
			(1.911)	(1.764)			(3.515)	(3.691)
Gender Congruence of Question			2.611	-1.491			13.47***	1.611
			(2.750)	(2.909)			(5.165)	(5.838)
Own Gender Share in Group			6.604**	2.209			4.357	0.905
			(2.762)	(2.440)			(6.958)	(7.236)
R-squared	0.0679	0.0843	0.0856	0.0870	0.0524	0.0667	0.0748	0.106
Dep. Var. Mean	46.67	49.30	46.67	49.30	53.68	55.47	53.68	55.47
Observations	408	402	408	402	408	402	408	402

Notes: Coefficients obtained using a linear probability model. Explanatory variables represent the difference between j and k. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individivual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B11: The Effect of Beliefs, No Demographic Controls

	Probability of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)		Probability of Ranking j higher than k (Explanatory variables represent the difference between j and k)	
	KG	UG	KG	UG
	(1)	(2)	(3)	(4)
Female	-1.631	0.665	2.557	5.244*
	(1.591)	(1.446)	(2.838)	(3.100)
Gender Congruence of Question	2.322	-1.784	10.10**	3.770
	(2.313)	(2.368)	(4.261)	(4.807)
Own Gender Share in Group	6.344***	-0.842	7.492	-5.032
	(2.392)	(1.958)	(5.687)	(6.072)
Predicted Points of Member 1	4.708***	4.780***		
	(0.383)	(0.410)		
Predicted Points of Member 2	-1.719***	-1.319***		
	(0.313)	(0.298)		
Predicted Points of Member 3	-1.430***	-1.187***		
	(0.378)	(0.328)		
Predicted Points of Member j v. k			9.565***	9.047***
			(0.577)	(0.591)
R-squared	0.324	0.335	0.312	0.321
Dep. Var. Mean	46.67	49.26	53.57	55.39
Observations	408	402	408	402

Notes: Coefficients obtained using a linear probability model. Explanatory variables represent the difference between j and k. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individivual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table B12: The Effect of Third-Party Beliefs, No Demographic Controls

	Probability of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)		Probability of Ranking j higher than k (Explanatory variables represent the difference between j and k)	
	KG	UG	KG	UG
	(1)	(2)	(3)	(4)
Made Valuable Contributions	3.946	6.185	6.317	9.147
	(4.847)	(3.801)	(6.990)	(6.270)
Deserves Extra Compensation	9.365**	-0.808	0.0854	-2.808
	(4.563)	(4.178)	(7.045)	(6.958)
Bet on Being Chosen	4.351	6.776**	16.70***	9.406
	(4.257)	(3.310)	(6.395)	(5.867)
Female	-2.021	0.287	5.107	4.697
	(1.864)	(1.723)	(3.455)	(3.620)
Gender Congruence of Question	2.247	-1.128	11.87**	3.575
	(2.670)	(2.881)	(5.149)	(5.740)
Own Gender Share in Group	6.082**	1.789	4.359	1.017
	(2.708)	(2.391)	(6.843)	(7.102)
R-squared	0.121	0.0953	0.0896	0.0945
Dep. Var. Mean	46.67	49.30	53.68	55.47
Observations	408	402	408	402

Notes: Coefficients obtained using a linear probability model. Explanatory variables represent the difference between j and k. Sample is restricted to chat treatment data only. All specifications include fixed effects for round and part. Columns 1-4 control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individivual scores are also included. Columns 5-8 control for the differences in pre- and post-group individual scores of j relative to k. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

C. Results for the Control Treatments

In the chat treatments, we find that individuals are significantly more likely to be chosen to represent the group for more gender congruent questions. In the section, we investigate whether this result holds when the level of interaction is greatly controlled and diminished.

In particular, group members in the Answer Only treatment only have the information about the participants' answer to the pre-group question. That is, they do not chat or interact. Their pre-group answers are simply shared in a table visible on the screen. In the Answer + Confidence treatment, the subjects can see the pre-group answer of each member, as well as the associated confidence in that answer. These control treatments shut down any potential mechanisms behind the gender gap that have to do with gender stereotypes about how men and women express ideas, leaving only the effect of objective information about one's ability to answer the question (Answer Only) and subjective beliefs of ability (Answer + Confidence).

Table C1 shows the predicts the total likelihood of being chosen as the group representative from gender, gender congruence, and share of same gender group members, mirroring Table 2 in the main text. We do this separately for each control treatment. We find no significant gender-related biases in the Answer Only or the Answer + Confidence treatment. We conclude that the way in which men and women interact with one another produces significant gender-related biases that are not present in the absence of chat.

Table C1: The Determinants of the Probability of Being Chosen as the Group Representative in the Post-Group Stage of the Control Treatments

	Answer Only Treatment						Answer + Confidence Treatment					
	All Groups			Mixed-Gender Groups			All Groups			Mixed-Gender Groups		
Sample	KG	UG	Pooled	KG	UG	Pooled	KG	UG	Pooled	KG	UG	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	0.283	1.109	0.888	-0.0885	1.027	0.796	-0.678	1.284	1.232	-0.754	1.117	1.711
	(1.376)	(1.353)	(1.298)	(1.732)	(1.564)	(1.485)	(1.341)	(1.610)	(1.504)	(1.567)	(1.819)	(1.705)
Gender Congruence of Question	-0.666	-0.298	-0.301	-0.378	0.0710	0.0230	1.910	0.0129	-0.0649	2.908	0.258	0.414
	(1.756)	(1.740)	(1.715)	(2.201)	(1.987)	(1.962)	(1.642)	(2.021)	(1.991)	(1.903)	(2.341)	(2.265)
Own Gender Share in Group	1.231	0.416	0.581	3.685	0.401	1.139	1.225	-0.565	-0.741	1.459	-3.370	-3.117
	(1.794)	(1.892)	(1.857)	(3.310)	(3.189)	(3.139)	(1.816)	(2.216)	(2.165)	(3.231)	(3.870)	(3.710)
KG Treatment			0.0721			0.0697			0.755			0.348
			(1.836)			(2.188)			(2.016)			(2.449)
Female x KG			-0.529			-0.410			-2.379			-2.816
			(1.836)			(2.183)			(2.016)			(2.322)
Gender Congruence x KG			-0.335			-0.554			1.880			2.100
			(2.435)			(2.922)			(2.574)			(2.960)
Own Gender Share x KG			0.362			1.281			1.410			3.934
			(2.532)			(4.469)			(2.816)			(4.987)
R-squared	0.101	0.125	0.101	0.140	0.144	0.114	0.156	0.108	0.118	0.167	0.131	0.127
Observations (groups)	420 (140)	432 (144)	852 (284)	306 (102)	333 (111)	639 (213)	348 (116)	348 (116)	696 (232)	261 (87)	267 (89)	528 (176)

Notes: Coefficients obtained using a linear probability model. Dependent variable mean is 33.33. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, dummy for whether the US is the country of citizenship and birth; and controls for pre- and post-chat individual performance, as well as performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

D. Conversation Data

D.1. Exploring the Objective Conversation Measures and Their Effect on Group Behavior

Table D1: Comparison of Conversation Variables by Gender and Treatment

	Male	Female	Absolute Difference	t-test p-value	Mann-Whitney U test p-value
All Chat Treatments					
Number of Total Words	11.13	11.50	0.37	0.482	0.580
Number of Engagements	3.72	3.59	0.13	0.275	0.244
Intensity of Engagement	19.67	20.79	1.13	0.148	0.174
Suggested Group Answer	0.46	0.40	0.07*	0.058	0.058
Share of Others Convinced	0.40	0.36	0.04	0.216	0.197
Switched Answer	0.45	0.47	0.02	0.575	0.575
New Ideas per Conversation	1.08	1.05	0.04	0.441	0.665
Weak Words	0.69	0.80	0.11*	0.067	0.104
Confident Words	1.10	1.11	0.01	0.825	0.921
KG Treatment					
Number of Total Words	11.43	11.95	0.52	0.471	0.628
Number of Engagements	3.69	3.40	0.29*	0.076	0.076
Intensity of Engagement	20.19	22.28	2.09**	0.047	0.059
Suggested Group Answer	0.49	0.39	0.09*	0.059	0.059
Share of Others Convinced	0.41	0.34	0.06	0.145	0.120
Switched Answer	0.43	0.50	0.07	0.176	0.176
New Ideas per Conversation	1.06	1.00	0.06	0.353	0.431
Weak Words	0.78	0.84	0.06	0.496	0.541
Confident Words	1.19	1.18	0.01	0.941	0.740
UG Treatment					
Number of Total Words	10.79	11.08	0.29	0.713	0.698
Number of Engagements	3.76	3.77	0.01	0.942	0.971
Intensity of Engagement	19.09	19.39	0.30	0.792	0.766
Suggested Group Answer	0.44	0.40	0.04	0.461	0.461
Share of Others Convinced	0.39	0.38	0.01	0.757	0.775
Switched Answer	0.47	0.45	0.03	0.573	0.573
New Ideas per Conversation	1.11	1.09	0.02	0.769	0.994
Weak Words	0.59	0.77	0.18**	0.038	0.069
Confident Words	1.00	1.05	0.05	0.577	0.736

Significance levels: *10 percent, **5 percent, ***1 percent.

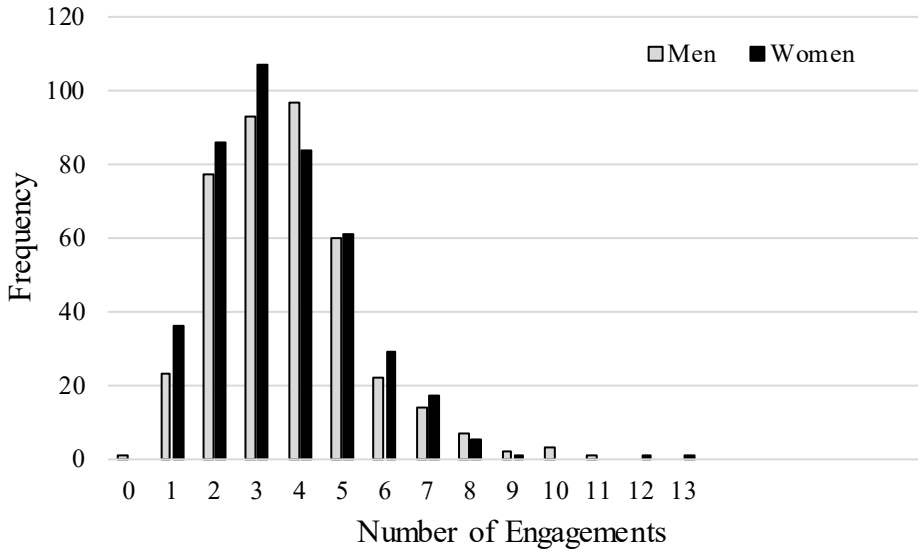

Figure D1. Distribution of Engagements by Gender in Chat

Note that the following words, characters, and expressions appearing in the transcripts of our conversations were classified as "weak" (which can be interpreted as under-confident):

maybe, just, unsure, probably, sort of, ?, sorry, not sure, agree, possible, no idea, or something, idk, guess

On the other hand, the following words, characters, and expressions appearing in the transcripts of our conversations were classified as "confident":

definitely, sure, certain, should, opinion, believe, !, let's

Table D2 asks how these objective chat characteristics impact the probability of being chosen as the group representative. We find that there are three important factors in predicting a favorable ranking in both treatments. First is the number of times a participant enters a statement into the chat interface (the level of engagement). More engaged participants are chosen more often as group representatives. Second is the ability to convince others to adopt one's initial answer. Note that these regressions condition on pre-group answer quality, so the fact that convincing others positively predicts being chosen is not simply picking up on some individuals having better pre-group answers. Third, individuals who switch from their initial answer to a new answer in the post-group stage are significantly less likely to be chosen in both treatments.

Table D3 explores the heterogeneity of these effects reporting two regressions: one for KG and one for UG. We interact each of these factors of interest with gender to ask whether they are similarly predictive for men and women (coefficients on these interactions recorded in the second set of estimates). We also interact each of these objective factors with the gender congruence of the question (third set of estimates) and the same gender share for the member (fourth set of estimates). Overall, we see no large differences. It is clear that engagements, the share of others convinced, and not switching from your original pre-group answer positively predict being chosen

as the group representative. The extent to which these matter does not seem to vary much by gender, gender stereotype, or same gender share.

Table D2: The Effect of Objective Conversation Measures on the Probability of Being Chosen as the Group Representative in the Post-Group Stage

	All Groups		Mixed-Gender Groups	
Sample	KG	UG	KG	UG
	(1)	(2)	(3)	(4)
Engagements	0.684*	0.637**	0.319	0.888**
	(0.411)	(0.310)	(0.469)	(0.389)
Intensity of Engagement	0.0291	0.0968*	0.00919	0.107*
	(0.0546)	(0.0550)	(0.0665)	(0.0626)
Suggested Group Answer	4.535**	1.154	3.590	3.204
	(2.187)	(1.975)	(2.543)	(2.442)
Share of Others Convinced	5.967***	7.602***	8.214***	6.808**
	(2.162)	(2.120)	(2.520)	(2.629)
Switched Answer	-5.733***	-4.724***	-6.626***	-3.672*
	(1.667)	(1.480)	(1.868)	(1.965)
New Ideas per Conversation	1.537	0.677	1.908	1.221
	(0.969)	(0.860)	(1.185)	(1.070)
Weak Words	-0.290	0.920	-0.0876	0.420
	(0.677)	(0.621)	(0.761)	(0.729)
Confident Words	-0.506	-0.238	-0.418	-0.564
	(0.682)	(0.599)	(0.869)	(0.709)
Female	0.981	0.737	0.748	1.585
	(1.089)	(1.091)	(1.266)	(1.271)
Gender Congruence of Question	2.971*	-0.384	3.698**	-0.421
	(1.545)	(1.634)	(1.834)	(1.936)
Own Gender Share in Group	1.749	1.743	4.441*	3.584
	(1.422)	(1.415)	(2.482)	(2.490)
R-squared	0.355	0.312	0.401	0.334
Observations (groups)	419 (140)	408 (136)	317 (106)	285 (95)

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only. Dependent variable mean is 33.33. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, dummy for whether the US is the country of citizenship and birth; and controls for pre- and post-chat individual performance, as well as performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table D3: The Effect of Objective Conversation Measures on the Probability of Being Chosen as the Group Representative in the Post-Group Stage, Interacted Model

	Level Effects		Interactions with Female		Interactions with Gender Congruence of Question		Interactions with Own Gender Share	
Sample	KG	UG	KG	UG	KG	UG	KG	UG
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Engagements	0.208	0.717	0.513	1.284	-1.678	-1.404	1.300	-1.394
	(0.916)	(0.658)	(0.860)	(0.793)	(1.206)	(1.151)	(1.277)	(0.895)
Intensity of Engagement	0.0508	0.174*	-0.0812	0.0867	0.0909	0.0772	0.120	-0.195
	(0.111)	(0.0915)	(0.114)	(0.105)	(0.177)	(0.144)	(0.146)	(0.140)
Suggested Group Answer	7.902*	6.924**	-8.052*	-3.948	5.092	-3.633	-1.458	-5.559
	(4.747)	(3.501)	(4.557)	(4.060)	(6.407)	(6.535)	(7.249)	(4.578)
Share of Others Convinced	8.959**	4.785	5.482	2.723	-3.528	-3.108	-8.407	2.350
	(4.472)	(3.976)	(4.538)	(4.368)	(6.542)	(6.797)	(6.712)	(5.173)
Switched Answer	-5.158	1.701	-3.419	-7.936***	-3.080	-2.766	1.471	-3.511
	(3.633)	(3.145)	(3.333)	(2.966)	(4.812)	(4.516)	(4.747)	(3.809)
New Ideas per Conversation	0.357	1.123	0.978	2.455	3.861	0.627	0.689	-2.778
	(1.942)	(1.369)	(1.804)	(1.764)	(2.634)	(2.814)	(2.685)	(2.136)
Weak Words	0.884	0.995	-2.882**	-0.909	-0.587	1.344	-0.0496	0.0538
	(1.269)	(1.356)	(1.257)	(1.460)	(1.885)	(2.073)	(1.681)	(1.806)
Confident Words	0.0949	-2.000	-0.629	0.580	-0.0374	1.794	-1.293	2.755
	(1.325)	(1.342)	(1.331)	(1.292)	(1.970)	(2.144)	(1.624)	(1.737)
Female	5.663	-3.137						
	(4.651)	(4.112)						
Gender Congruence of Question	4.163	3.203						
	(6.943)	(7.053)						
Own Gender Share in Group	-1.074	13.97***						
	(5.952)	(5.050)						
R-squared	0.408	0.364						
Observations (groups)	419 (140)	408 (136)						

D.2. Additional Analysis Using Subjective Conversation Data

Table 6 in the paper shows that controlling for the third-party coders' assessments of conversations does not explain away the effects of gender stereotypes and group composition. Table 9 in the paper further reveals that gender stereotypes and group composition remain important predictors of self- and other-rankings even after we control for third-party recognition of valuable group members.

Table D4 repeats the analysis in Table 6, while Table D5 repeats the analysis in Table 9, with the outcome variable of the probability of being chosen to represent the group, rather than the two separate rankings. Overall, we see once again that assertive and competent group members are more likely to be selected as group representative. Directionally, warmth and openness are correlated with a lower chance of being rewarded. Importantly, we see that the inclusion of these conversation measures does not change the significance of gender stereotypes (gender congruence of question). Own gender share also continues to be an important factor.

Tables D6 and D7 estimate the model from Table 6 that also includes interactions of the conversation style variables with gender, gender congruence of question, and group composition variables. The analysis is performed on the sample with all groups in the chat treatment, separately for self-ranking (Table D6) and ranking of other members (Table D7). Each table consists of two regression specifications, one for each treatment (KG and UG). We document that the interacted model does not change the main results.

Finally, Table D8 shows the relationship between our subjective and objective chat characteristics. We observe that, unsurprisingly, participants who are deemed to be assertive are significantly more likely to be the ones to have more engagements in the conversation. Open-mindedness seems to be positively correlated with the likelihood of switching one's answer, but the correlation is only significant in the KG treatment. On the other hand, more competent participants are less likely to switch (again, in the KG treatment).

Table D4: The Effect of Subjective Conversation Measures on the Probability of Being Chosen as Group Representative Post Group Interaction in the Chat Treatment

	Probability of Favorable Self-Ranking (Explanatory variables represent subject i's characteristics)			
	KG	UG	KG	UG
	(1)	(2)	(3)	(4)
Assertive	1.649	2.444**	1.730	2.510**
	(1.427)	(1.138)	(1.431)	(1.141)
Competent	0.935	2.705**	0.781	2.668**
	(1.372)	(1.318)	(1.382)	(1.314)
Warm	-1.472	-1.488	-1.494	-1.680
	(1.555)	(1.304)	(1.522)	(1.311)
Open	0.306	-1.361	0.284	-1.328
	(1.569)	(1.227)	(1.588)	(1.223)
Difficult	-0.182	-0.0440	0.00288	-0.00831
	(1.359)	(1.214)	(1.375)	(1.202)
Female			0.0970	1.973
			(1.279)	(1.224)
Gender Congruence of Quest.			3.747**	-1.048
			(1.868)	(1.857)
Own Gender Share in Group			3.069*	0.960
			(1.704)	(1.631)
R-squared	0.0987	0.162	0.115	0.169
Observations	408	402	408	402

Notes: Coefficients obtained using a linear probability model. Dependent var. mean is 33.33. All specifications include fixed effects for round and part; demographic controls for gender, age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. We control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individual scores are also included. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table D5: The Effect of Third-Party Recognition Measures on the Probability of Being Chosen as Group Representative Post Group Interaction in the Chat Treatment

	KG	UG
	(1)	(2)
Made Valuable Contributions	3.963	6.241***
	(3.189)	(2.246)
Deserves Extra Compensation	3.853	-2.740
	(3.145)	(2.555)
Bet on Being Chosen	8.630***	5.844**
	(2.778)	(2.422)
Female	0.0594	1.812
	(1.235)	(1.204)
Gender Congruence of Question	3.282*	-0.136
	(1.774)	(1.879)
Own Gender Share in Group	2.389	0.948
	(1.624)	(1.587)
R-squared	0.183	0.180
Observations	408	402

Notes: Coefficients obtained using a linear probability model. Dependent var. mean is 33.33. All specifications include fixed effects for round and part; demographic controls for gender, age, student status, race, English language proficiency, income, use of real name, and a dummy for whether the US is the country of citizenship and birth. We control for performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Pre- and post-group individual scores are also included. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table D6: The Effect of Subjective Conversation Measures on the Probability of Favorable Self-Ranking, Interacted Model (Explanatory variables represent subject i's characteristics)

	Level Effects		Interactions with Female		Interactions with Gender Stereotype of Question		Interactions with Own Gender Share	
Sample	KG	UG	KG	UG	KG	UG	KG	UG
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Assertive	-1.623	1.923	-2.968	8.106**	-7.522	3.442	5.697	-4.993
	(4.420)	(4.129)	(4.525)	(3.815)	(6.281)	(6.026)	(6.410)	(5.324)
Competent	-2.451	3.470	3.365	-2.270	-7.608	-6.196	2.091	-3.731
	(4.700)	(4.387)	(4.457)	(3.993)	(6.363)	(6.448)	(6.359)	(5.709)
Warm	0.163	2.924	-2.620	-2.385	2.334	0.581	-4.578	2.801
	(4.889)	(4.657)	(4.719)	(4.076)	(7.189)	(6.574)	(6.012)	(6.029)
Open	1.169	-4.797	-0.475	2.649	0.879	1.741	2.086	0.361
	(4.607)	(3.985)	(5.028)	(3.896)	(7.068)	(5.027)	(6.690)	(5.080)
Difficult	1.085	-7.010*	-2.620	1.044	6.045	-6.329	2.534	11.82**
	(4.249)	(3.995)	(4.361)	(4.070)	(6.072)	(6.434)	(5.647)	(5.987)
Female	-1.094	0.314						
	(1.973)	(2.100)						
Gender Congruence of Question	1.712	-1.759						
	(2.844)	(3.204)						
Own Gender Share in Group	8.304***	2.309						
	(2.979)	(2.554)						
R-squared	0.149	0.166						
Observations (groups)	420 (140)	408 (136)						

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only, all groups. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, dummy for whether the US is the country of citizenship and birth; and controls for pre- and post-chat individual performance, as well as performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Note that in the Chat treatment, unlike the other two treatments, the pre-group answers of other group members were not displayed to participants. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table D7: The Effect of Subjective Conversation Measures on the Probability of Ranking j higher than k (Explanatory variables represent the difference between j and k)

	Level Effects		Interactions with Female		Interactions with Gender Stereotype of Question		Interactions with Own Gender Share	
Sample	KG	UG	KG	UG	KG	UG	KG	UG
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Assertive	6.057	9.973**	-4.917	5.016	-19.74**	-12.95	13.25	-24.31**
	(4.075)	(4.101)	(6.392)	(6.105)	(8.887)	(8.932)	(12.66)	(12.04)
Competent	9.585**	6.872	9.338	-3.504	5.906	8.995	15.58	-8.040
	(4.695)	(4.354)	(7.299)	(6.964)	(9.873)	(10.35)	(13.12)	(14.52)
Warm	8.872*	-6.662	15.00**	16.45**	12.73	-7.725	15.82	-14.62
	(5.209)	(4.485)	(6.605)	(7.461)	(10.58)	(11.89)	(12.75)	(14.53)
Open	-4.421	-0.0585	-11.74	5.923	2.536	11.07	-15.59	-0.305
	(4.877)	(4.229)	(7.279)	(7.195)	(11.01)	(10.39)	(13.59)	(14.25)
Difficult	6.681	-1.224	-2.871	1.124	-11.58	0.573	4.497	-5.381
	(6.361)	(7.261)	(8.684)	(9.983)	(12.82)	(18.83)	(17.40)	(22.60)
Female	5.364	4.170						
	(3.724)	(3.677)						
Gender Congruence of Question	11.59**	2.628						
	(5.662)	(6.008)						
Own Gender Share in Group	7.834	4.430						
	(7.300)	(7.400)						
R-squared	0.160	0.178						
Observations (groups)	420 (140)	408 (136)						

Notes: Coefficients obtained using a linear probability model. Sample is restricted to chat treatment data only, all groups. All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, dummy for whether the US is the country of citizenship and birth; and pre- and post-group ability of j relative to k. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Table D8: The Relationship between the Objective and Subjective Chat Characteristics

	KG				UG			
Sample	Engagements	Suggest Group	Share Convinced	Switched Answer	Engagements	Suggest Group	Share Convince	Switched Answer
	(1)	(2)	(3)	(4)	(6)	(7)	(8)	(9)
Competent	0.0449	0.0503	0.0570	-0.123**	-0.0861	0.0792	0.0439	0.0204
	(0.175)	(0.0535)	(0.0474)	(0.0546)	(0.191)	(0.0500)	(0.0402)	(0.0489)
Assertive	0.354**	0.0311	0.00173	0.0139	0.829***	0.0465	0.0211	-0.0868*
	(0.178)	(0.0521)	(0.0462)	(0.0519)	(0.197)	(0.0451)	(0.0399)	(0.0443)
Warm	-0.291	0.00744	-0.00852	0.0234	-0.197	0.0250	0.0217	-0.0183
	(0.210)	(0.0557)	(0.0478)	(0.0589)	(0.180)	(0.0541)	(0.0484)	(0.0527)
Open-Minded	0.311	-0.0269	0.0180	0.156***	-0.0320	-0.107**	-0.0535	0.0170
	(0.205)	(0.0545)	(0.0503)	(0.0548)	(0.180)	(0.0501)	(0.0450)	(0.0468)
Difficult	0.157	-0.0383	-0.0554	0.0391	-0.202	-0.0331	-0.0560	0.0433
	(0.196)	(0.0502)	(0.0449)	(0.0504)	(0.165)	(0.0467)	(0.0403)	(0.0457)
Female	-0.189	-0.0799*	-0.0469	0.0492	0.168	0.0244	0.0483	-0.0677
	(0.158)	(0.0462)	(0.0409)	(0.0469)	(0.186)	(0.0466)	(0.0407)	(0.0467)
R-squared	0.206	0.221	0.204	0.212	0.186	0.270	0.323	0.296
Observations	408	408	408	408	402	402	402	402

Notes: All specifications include fixed effects for round and part; demographic controls for age, student status, race, English language proficiency, income, use of real name, dummy for whether the US is the country of citizenship and birth; and controls for pre- and post-chat individual performance, as well as performance distribution pre-chat that includes difference from maximum group score and difference from average group score. Robust standard errors double-clustered at the group and individual level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

E. Original Conversations Coding Study

In the main text, we report the results of a coding experiment, where new participants read the conversations from the original laboratory sessions and provided perceptions. The experiment in the main text was an expanded follow-up of an original coding experiment that we report here. Full instructions for this experiment are available here.

We originally recruited 1000 Amazon Mechanical Turk (AMT) workers to read the conversations and provide impressions of the conversations.[1] We are interested in the AMT raters' perceptions of which of the members made particularly effective contributions to the group, and in how the different group members may have varied in their communication style.

Each AMT participant read three randomly-selected transcripts. Importantly, within each conversation, members were labeled simply as Member 1, 2, or 3. That is, we blind AMT participants to gender. The design was such that each participant saw no more than one conversation for each of the eight *Family Feud* questions used in the study, so as to reduce across-conversation comparisons. Instructions, which can be found in the online appendix, give detailed information about how these conversations were randomized and how the conversations were presented to participants.

For each conversation shown to the participant, she was asked a series of questions about each member of the conversation, both communication-style focused and performance focused. We placed the questions about each member on a separate page, to reduce confusion and to avoid too many questions on a single page. Each page contained the full transcript for the participant's reference. Following the warmth-competence literature (Fiske et al 2007), we asked participants to evaluate members on three dimensions of warmth (warm, tolerant, good-natured) and competence (competent, intelligent, confident). We also asked about how assertive and passive the member was, whether they were supportive or critical of others, and how stubborn they seemed. These 11 personality traits were presented in one block for each member, in an order randomized at the individual level.

We also asked AMT participants performance-oriented questions: to what extent each group member contributed to group success, did a good job voicing their ideas, advocated to be chosen by the group, impeded the group's success, advocated for their preferred answer, and had their ideas listened to by the group. These were again organized into one block and randomized at the individual level.

The 5-point scale ranged from "not at all" to "extremely" for all questions. At the end of each conversation question set, the AMT participants had to choose which of the three members

[1] Workers on AMT have been shown to exhibit similar behavioral patterns and pay attention to the instructions to the same extent as traditional subjects (Paolacci et al. 2010; Germine et al. 2012). Rand (2012) reviews replication studies that indicate that AMT data are reliable. We used randomly placed attention checking questions in order to ensure full attention. The final dataset contains valid responses of 985 AMT raters.

they would vote as the "MVP (most valuable player)" of the group. Finally, after the third (last) conversation question set, the AMT participants were asked to guess the gender of each of the three group members who participated in that chat. Note that we only asked this question once and at the very end of the survey, in order to not give away that our research question concerns gender differences. The main part of the survey was followed by a brief demographic questionnaire.

We provide incentives through matching. Following participation, we matched each participant with another participant who faced one of the same chat transcripts. We then randomly selected one of the questions about that chat and compared the answers. If both participants gave the same answer to that question, the participant received an extra $1.50 in bonus payment, in addition to the $2 participation fee.

With 17 scale questions about each member, we have a wealth of data on perceptions of each individual. In order to categorize questions into broader explanatory factors that are orthogonal to one another, we performed a principal component decomposition analysis, with rotation. The PCA produced three principal components. Table E1 presents the loadings of each factor over the 17 questions asked.

Importantly, while we of course chose the 17 questions that were asked about each participant in each conversation, this principal component analysis aggregates this data in a way that is independent of our judgment. The analysis looks at all of the data collected and organizes it into three orthogonal explanatory factors. Factor 1 loads heavily on competence, confidence, and assertiveness – aligning closely with the competence dimension identified by Fiske et al (2007); Factor 2 on warmth, good-naturedness, being supportive of others, and tolerance – aligning with the warmth dimension identified by Fiske et al (2007); and Factor 3 on the more negative traits of stubbornness, being critical of others, and impeding success. In the analysis below, we use these three independently-identified factors, coded as z-scores.

Table E1. Factor Weights

Variable	Factor 1 Weight	Factor 2 Weight	Factor 3 Weight	Uniqueness
competent	0.633	0.492	0.076	0.352
confident	0.779	0.192	0.143	0.335
intelligent	0.590	0.532	0.024	0.368
assertive	0.737	0.030	0.302	0.365
warm	0.215	0.799	0.038	0.314
good-natured	0.233	0.805	0.085	0.291
tolerant	0.078	0.813	0.107	0.321
passive	0.595	0.379	0.348	0.381
supportive of others	0.126	0.802	0.099	0.331
critical of others	0.184	0.144	0.782	0.333
stubborn	0.253	0.256	0.755	0.301
contributed to group's success	0.733	0.392	0.063	0.305
impeded group's success	0.140	0.011	0.803	0.336
voiced ideas	0.786	0.294	0.019	0.295
ideas were heard	0.747	0.257	0.036	0.375
advocated for own answer	0.726	0.050	0.286	0.388
advocated to be chosen as group rep.	0.591	0.092	0.457	0.434

Our main question of interest is whether men and women vary in their communication styles, as rated by our coders. That is, if we consider our three main dimensions of interest – which for simplicity we will call competence, warmth, and negativity – are there gender differences in these dimensions? In Table E2, we show that men and women are rated as identically competent, warm, and negative based upon their conversation contributions. That is, when blind to gender, coders perceive men and women as exactly the same on average on all three dimensions. Note that these results are unchanged if we consider only the known gender treatment (or unknown gender treatment). These results are somewhat in line with the results from our follow-up coding experiment, at least directionally. In our follow-up experiment, we found that when coders were blind to gender, women were perceived as significantly more assertive, more competent, and

warmer than men, and directionally less difficult to work with. We find much smaller estimated differences, but of the same sign, in this experiment.

Table E2: Relationship between Gender and the Level of Personality Trait Factor

Depedent Variable	Factor 1 ("Competence")	Factor 2 ("Warmth")	Factor 3 ("Negativity")
	(1)	(2)	(3)
Female	0.00679	0.000326	-0.0113
	(0.0554)	(0.0417)	(0.0334)
Fixed Effects	YES	YES	YES
Observations (clusters)	1,656 (207)	1,656 (207)	1,656 (207)
R-squared	0.027	0.046	0.069

Notes: Fixed effects include question, round, part, and treatment (gender known or unknown). Robust standard errors clustered at the conversationalist level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Recall that we ask our coders to guess the gender of the members of the conversations. Thus, we can ask what predicts the probability that a coder believes that a member is female. Table E3 reports the estimates from an OLS regression that predicts the likelihood that an AMT rater guessed a given participant was female from their evaluation of that member in terms of the three conversation factors we identified – Competence, Warmth, and Negativity. Importantly, these estimates are not causal: we cannot rule out that an unmeasured factor or conversation feature leads the coder to both evaluate the member in a particular way and guess that he or she is female. These estimates simply tell us which factors are correlated with a coder believing someone is female. Our follow-up experiment improves on this methodology by randomly assigning coders to be blinded to gender or not.

We find members viewed as warm (coded in Factor 2) by the rater are more likely to be believed to be female, while members viewed as negative (coded in Factor 3) are more likely to be believed to be male. Thus, while there are no actual differences in how men and women seem to communicate in these settings (at least as perceived by these coders), the coders hold strong stereotypes about the behavior that is more typical of men or women. Being warm is strongly associated with being perceived as female; being critical is strongly associated with being perceived as male. Given the inaccuracy of these stereotypes, it is perhaps not surprising that the raters are on average quite bad at correctly guessing gender: less than 45% of women are correctly identified as women.

Table E3: The Effect of Personality Trait Factors on the Prediction that a Participant is Female

Factor 1 ("Competence")	-0.016
	(0.010)
Factor 2 ("Warmth")	0.059***
	(0.009)
Factor 3 ("Negativity")	-0.051***
	(0.009)
Rater Was Female	0.084***
	(0.016)
Demographic Controls	YES
Dependent Var. Mean:	0.425
Observations (clusters)	2,961 (984)
R-squared	0.039

Notes: Sample includes all participants. Rater demographics include gender, education, race, and whether the rater attended high school in the US. Robust standard errors clustered at the rater level in parentheses. Significance levels: *10 percent, **5 percent, ***1 percent.

Summing up the evidence on gender stereotypes, we see that raters provide nearly identical ratings of men and women in our data on competence, warmth, and negativity. Yet, when asked to guess gender, the same coders incorrectly believe that those individuals that they rated as warmer or less negative are more likely to be women.

Finally, Table E4 shows the relationship between our three factors and the objective chat characteristics. We observe that participants who are deemed to be competent are significantly more likely to be the ones to have more engagements, suggest the group answer, convince others, and less likely to switch answer. Warmth predicts more engagements. Warmer participants are also more likely to switch answer, but only in the KG treatment.

Table E4: The Relationship between the Objective and Subjective Chat Characteristics

	KG				UG			
Sample	Engagements	Suggest Group	Share Convinced	Switched Answer	Engagements	Suggest Group	Share Convince	Switched Answer
	(1)	(2)	(3)	(4)	(6)	(7)	(8)	(9)
Competent (Factor 1)	0.988***	0.200***	0.209***	-0.224***	1.195***	0.237***	0.177***	-0.184***
	(0.127)	(0.0379)	(0.0340)	(0.0371)	(0.125)	(0.0320)	(0.0314)	(0.0347)
Warm (Factor 2)	0.447**	-0.244***	-0.139***	0.240***	0.620***	-0.100**	-0.0390	0.0756
	(0.190)	(0.0518)	(0.0423)	(0.0542)	(0.170)	(0.0493)	(0.0401)	(0.0494)
Critical (Factor 3)	1.058***	-0.201***	-0.0919*	0.0640	1.045***	-0.0464	-0.0558	-0.0391
	(0.230)	(0.0612)	(0.0523)	(0.0659)	(0.247)	(0.0562)	(0.0504)	(0.0539)
Female	-0.202	-0.0878**	-0.0470	0.0589	0.142	0.0306	0.0506	-0.0690
	(0.138)	(0.0418)	(0.0374)	(0.0433)	(0.159)	(0.0431)	(0.0385)	(0.0438)
R-squared	0.337	0.315	0.286	0.294	0.333	0.349	0.369	0.346
Observations	408	408	408	408	402	402	402	402

F. Instructions for the Coding Experiment Reported in the Main Text

The full instructions for the coding experiment reported in the main text are available under separate cover labeled Appendix F here.

G. AMT Experiment Piloting Family Feud Questions

We piloted 20 Family Feud Questions on Amazon Mechanical Turk in June 2017. We recruited 200 participants to answer 10 Family Feud questions each. For each question, participants were asked to brainstorm high-scoring answers to the question for 30 seconds. Following the question, they were asked to indicate on a sliding scale: "Please use the slider to indicate which gender would be better at answering. The more you drag the slider to one side, the larger you think that gender's advantage is." We code the male end of the scale as 1 and the female end of the scale as -1. They received $2 for participation.

This pilot had two main goals. First, we wanted to select 8 appropriate questions for the full experiment that varied in their gender-type. Second, we wanted to anticipate as many possible answers to each question used, so that we could code correct answers (in many possible variants as empirically collected) in the full experiment program. This allowed us to anticipate common typos, common phrasings, etc., based upon data.

The full instructions for this pilot screening are available under separate cover labeled Appendix G here.

H. Full Instructions for Main Experiment

In Appendix H1, under separate cover, we provide copies of the instructions that participants received. The file shows the introductory language, followed by the language for each of the six possible treatments. Participants would be randomized into receiving one of these six treatments in Part 1. The file then shows the language we used to segue to Part 2, which is treatment specific. Then, we've included the post-experiment questionnaire.

In Appendix H2, under separate cover, we provide screenshots of the full experiment as seen by participants. This includes all of the questions asked.

Both Appendix H1 and H2 are available here.