BELIEFS ABOUT GENDER

Online Appendix

Pedro Bordalo

Katherine Coffman

Nicola Gennaioli

Andrei Shleifer

Appendix A: Experimental Instructions and Materials (available in separate file)

Appendix B: Online Experiment on Social Desirability Bias

The beliefs reported in the experiment may be partially shaped by social norms, which may discourage a participant from truthfully reporting believed gender differences in performance. While we use incentives and anonymity to reduce such concerns, we cannot rule them out. To examine this issue, we ran an experiment online. We had two main goals. First, we were interested in understanding whether the patterns of beliefs that we observed in our samples of college students resembled beliefs patterns from a broader population. Second, we wanted to collect data on the role that social desirability bias might play in determining stated beliefs.

The experiment is a simplified version of Part 1 of the laboratory experiments we ran. It was conducted on Amazon Mechanical Turk. We use the same questions from the six categories in the OSU and Harvard experiment: Art, Verbal Skills, Emotion Recognition, Mathematics, Business, and Sports. To reduce the length of the study, each participant answers a subset of five of the ten questions in each of the six categories. They are paid $0.25 for every correct answer they submit.

After, they are asked about their own and others' performance. Specifically, they are asked to guess their own score (out of 5) in each category. They are then asked to guess the score in each category for a randomly-drawn female MTurk worker and a randomly-drawn male MTurk worker. The order of these two beliefs questions about others is randomized at the individual level. These beliefs questions are unincentivized.

Finally, we attempt to understand whether there may be social desirability bias associated with stating beliefs about gender differences in ability. We adapt the measure proposed by Krupka and Weber (2013) to elicit norms. Participants are asked: "Suppose someone thought that [insert gender] knew more about [insert category] than [insert opposite gender]. How reluctant do you think they would be to announce this to others?". Participants use a sliding scale with 7 places, with 1 labeled "Not at all Reluctant" and 7 labeled "Extremely Reluctant" to indicate their answer. Each participant sees six of these questions, one for each category. We randomize at the participant level whether they see versions of each question that ask about female advantages (women knew more) or male advantages (men knew more). The key is that we care about how participants perceive the social acceptability of reporting beliefs of gender differences. We are not interested in whether participants believe these statements are likely to be true, or whether they themselves would be reluctant to report such a difference. For those reasons, we phrase the question as "suppose someone believed X". And, like Krupka and Weber (2013), we incentivize participants to provide what they believe the modal answer

among other participants will be. They receive $0.05 for each of the sliding scale questions for which they provide an answer that matches the modal answer among the other workers that completed the HIT.

We ran the experiment in February 2016 in two batches. The first batch of 1,000 posted HITs only collected performance and beliefs data. The second batch, of 800 posted HITS, collected the same information on performance and beliefs but also asked about reluctance to report gender differences. Average participation time was approximately 30 minutes and average earnings were approximately $5.50. We present summary statistics in Table A1.

Table A1: Summary Statistics for Mturk Experiment			
	Men	Women	p-value
Mean Age	38.0	36.7	0.66
Proportion Finished High School	0.997	0.994	0.18
Proportion Finished College	0.577	0.591	0.52
Proportion White	0.802	0.808	0.76
Proportion East Asian	0.081	0.043	0.001
Proportion Black or African-American	0.043	0.081	0.001
Proportion Hispanic	0.057	0.043	0.17
N	987	844	

Performance (out of 5 questions)				
	Men	Women	Gap (M-W)	p value
Emotion Score	3.79	3.92	-0.13	0.02
Art Score	3.18	3.18	-0.001	0.99
Verbal Score	3.31	3.32	-0.01	0.88
Math Score	2.30	1.81	0.49	0
Business Score	3.14	2.69	0.45	0
Sports Score	3.37	2.90	0.46	0

Notes: We include data from all participants who finished the Qualtrics link, independent of whether they submitted their performance for payment on Amazon Mechanical Turk. We posted 1,800 HITs in two batches (1,000 and 800).

Figure A1 graphs the raw data collected from Amazon Mechanical Turk. We define exaggeration of believed gaps as the difference between the believed gender advantage in the category and the observed gender advantage in the category. Larger exaggeration reflects believed gaps that exceed observed gaps – in the direction of a female advantage in female-typed categories and in the direction of a male advantage in male-typed categories. The figure below plots exaggeration across categories, and overlays them with our measures of reluctance to report a believed male (female) advantage in male (female) typed categories.

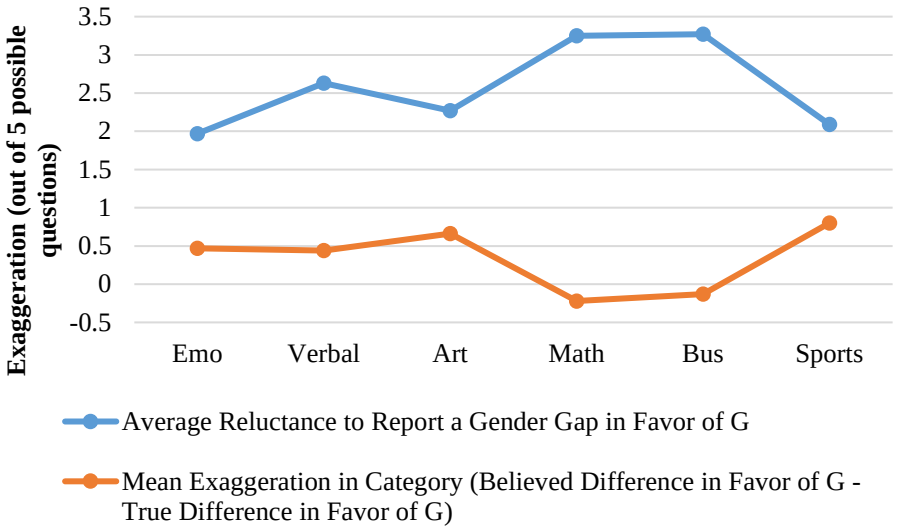

Figure A1. Exaggeration versus Reluctance to Report a Gender Difference

Notes: Emotion, Verbal, and Art have true gaps in favor of women and we report average reluctance to report female advantages in these categories; Math, Business, and Sports have true gaps in favor of men and we report average reluctance to report male advantages in these categories.

Figure A1 shows that: i) believed gaps exaggerate true gaps except in math and business, ii) reluctance to report a gender's true advantage (men in this case) is large in precisely these two categories. While hardly definitive, this evidence suggests that social norms may be an important factor driving stated beliefs.

Appendix C: Additional Tables and Empirical Analysis

C1. First Stage of Two-Stage Analysis

Below are the results for the first stage of the two-stage analysis presented in Table III, specifications I and II.

Table A2: OLS Predicting $I_{i,j}$ Dummy for whether Individual Answered Question Correctly

	I (Men)	II (Women)
Share of Correct Answers to Question Overall (Excluding individual *i*)	1.012**** (0.014)	0.954**** (0.015)
Share of Correct Answers in Category *J* by Individual *i* (Excluding question *j*)	0.400**** (0.017)	0.421**** (0.017)
Own Gender Advantage in Category	0.481**** (0.027)	0.117**** (0.035)
Constant	-0.215**** (0.009)	-0.195**** (0.009)
R-squared	0.23	0.24
Clusters	548	504
N	23,438	21,840

Notes: Pools OSU, Harvard, and UCSB data across all treatments. Standard errors are clustered at the individual level.

C2. Kitchen Sink Regressions for Self-beliefs in Part 3

Table A3 presents the "kitchen sink" specifications for predicting self-beliefs in question-level data. We predict own believed probability of answering correctly from our measures of ability: a dummy for whether the individual answered the specific question correctly, share of correct answers in category provided by individual in the bank of questions other than *j*, and the share of correct answers on question *j* by all individuals other than individual *i*. While we cannot recover our parameter estimates for DIM from this specification, the estimates for the effect of stereotypes are similar to the main specifications presented in Table III, repeated here as specifications I and II.

Table A3: Predicting Own Believed Probability of Answering Correctly

	I Two-Stage Least Squares (Men)	II OLS (Men)	III Two-Stage Least Squares (Women)	IV OLS (Women)
Own Gender Adv.	-0.039 (0.026)	0.093**** (0.025)	0.49**** (0.028)	0.42**** (0.030)
Fitted Value of $\hat{I}_{i,j}$	0.60**** (0.011)		0.61**** (0.011)	
Dummy for Individual Answered Qn. Correctly, $I_{i,j}$		0.21**** (0.005)		0.18**** (0.005)
Individual's Share of Correct Answers in Category excluding question *j*		0.36**** (0.017)		0.35**** (0.015)
Overall Share of Correct Answers to question *j*		0.31**** (0.010)		0.33**** (0.011)
Constant	0.33**** (0.009)	0.19**** (0.012)	0.30**** (0.009)	0.17**** (0.011)
Clusters	548	548	504	504
N	23,438	23,438	21,840	21,840

Notes: Pools OSU, UCSB, and Harvard data across all treatments. Standard errors are clustered at the individual level.

C3. Gender of Evaluator

Appendix Tables A4 presents the results on beliefs about others separated by the gender of the evaluator. In both sets of data, female evaluators seem to rely on stereotypes more than male evaluators, particularly when evaluating women. In both the question-level and bank-level data, we estimate that women stereotype female partners significantly more than men do. We see no consistent differences in DIM parameters for male and female evaluators.

Table A.4: Beliefs about Others by Gender of Evaluator					
Question-Level Beliefs OLS Predicting Belief of Partner's Probability of Answering a Question Correctly			**Bank-Level Beliefs** OLS Predicting Belief of Partner's Probability of Answering a Question Correctly		
	I (Beliefs About Men)	II (Beliefs About Women)		III (Beliefs About Men)	IV (Beliefs About Women)
Partner's Gender Adv.	0.045 (0.029)	0.35**** (0.048)	Partner's Gender Adv.	0.35**** (0.058)	0.052 (0.074)
Share of Partner's Gender Answering Qn. Correctly	0.36**** (0.017)	0.36**** (0.047)	Partner's Gender Avg. Score in Category (0 to 1 scale)	0.63**** (0.061)	0.62**** (0.051)
Female Evaluator	-0.011 (0.020)	0.034 (0.023)	Female Evaluator	-0.043 (0.047)	-0.011 (0.042)
Female Evaluator x Share of Answering Qn. Correctly	-0.034 (0.025)	-0.065** (0.031)	Female Evaluator x Partner's Gender Avg. Score in Category	0.026 (0.086)	-0.014 (0.074)
Female Evaluator x Partner's Gender Adv.	-0.046 (0.053)	0.27**** (0.072)	Female Evaluator x Partner's Gender Adv.	0.19* (0.100)	0.18* (0.110)
Constant	0.40**** (0.014)	0.42**** (0.015)	Constant	0.18**** (0.033)	0.22**** (0.030)
Clusters	395	398	Clusters	395	398
N	18,020	18,179	N	2,590	2,630

Notes: Includes data only from participants who knew the gender of their partner. We pool observations from OSU, Harvard, and UCSB experiments. Standard errors are clustered at the individual level.

C4. More on Context Dependence

In Section 5.5, we presented results on context dependence in beliefs of own ability. In Appendix Table A5, we extend this analysis by presenting pooled specifications that increase statistical power by examining men and women jointly. Context dependence predicts that both men and women should react more to the male advantage in a category, increasing beliefs of own ability, when paired with a female partner than when paired with a male partner. This is indeed what we find in the question-level data, demonstrated by the significant interaction of partner female and male advantage in specification I. We find a directionally similar result in the bank-level data, though it is only marginally significant (p=0.10).

Table A5. Self-beliefs with Context-Dependence, Pooled			
Question-Level Beliefs OLS Predicting Believed Probability of Answering Correctly		**Bank-Level Beliefs** OLS Predicting Believed Score	
	I (Pooled)		II (Pooled)
Male Adv.	-0.13**** (0.032)	Male Adv.	0.12** (0.048)
Fitted Value of $\hat{I}_{i,j}$	0.59**** (0.009)	Score in Bank	0.70**** (0.015)
Partner Female	-0.001 (0.008)	Partner Female	-0.007 (0.010)
Partner Female x Male Adv.	0.096** (0.039)	Partner Female x Male Adv.	0.095* (0.056)
Female	-0.032**** (0.008)	Female	-0.018* (0.010)
Female x Male Adv.	-0.45**** (0.040)	Female x Male Adv.	-0.64**** (0.060)
Constant	0.33**** (0.010)	Constant	0.12**** (0.012)
Clusters	793	Clusters	793
N	36,199	N	5,220

Notes: Includes laboratory data from OSU, Harvard, and UCSB samples, using only observations for individuals who knew partner's gender. Standard errors are clustered at the individual level.

We can also consider other evidence of context dependence in our data by considering reactions to partner ethnicity in the Ohio State sample, where participants received photographs of their partners. While the experiment was not designed to consider ethnic stereotypes, the fact that a substantial fraction of the Ohio State sample is composed of Asian and Asian American students may have activated ethnic as well as gender stereotypes within the experiment. To explore this, we follow our approach to studying gender. We construct the average Asian advantage within each category for both banks of questions for each category (average Asian performance – average performance of all non-Asians in sample). We proxy for ability as we did for gender: in bank-level analysis, we simply use Part 1 score in category and in the question-level analysis, we follow our two-stage approach, creating fitted values, $\hat{I}_{i,j}$ in a first stage that is performed separately on the Asian and non-Asian samples.

Recall that we have four categories in the Ohio State data: art, verbal skills, math, and sports. Asians have an advantage on average in math but are at a disadvantage on average in the other three categories. Compared to the gender gaps, the ethnicity gaps are quite large: among the 10 questions in Part 1, the gaps are -1.10 in art, -1.55 in verbal, 1.62 in math, and -1.23 in sports. Our test of context dependence asks whether participants report less optimistic self-beliefs as the Asian advantage increases when paired with an Asian partner than when paired with a non-Asian partner. Appendix Table A6 demonstrates that is indeed what we find

for non-Asian participants, both in question-level and bank-level data. Asian participants react to partner ethnicity as expected in question-level data, but not bank-level data.

Table A.6 Self-beliefs with Context Dependence with Ethnic Stereotypes

Question-Level Beliefs OLS Predicting Own Believed Probability of Answering Correctly in Part 3			**Bank-Level Beliefs** OLS Predicting Own Believed Part 1 Score		
	I (Non-Asians)	II (Asians)		III (Non-Asians)	IV (Asians)
Asian Adv. in Pt. 3	0.025	0.43****	Asian Adv. in Pt. 1	0.33****	0.63****
	(0.047)	(0.087)		(0.071)	(0.113)
Fitted Value of $\hat{I}_{i,j}$	0.61****	0.71****	Part 1 Score	0.68****	0.70****
	(0.020)	(0.038)		(0.041)	(0.068)
Partner Asian	-0.027	0.032	Partner Asian	-0.00	0.047
	(0.021)	(0.026)		(0.215)	(0.030)
Partner Asian x Asian Adv. in Part 3	-0.20*	-0.21*	Partner Asian x Asian Adv. in Part 1	-0.27**	0.22
	(0.106)	(0.121)		(0.125)	(0.135)
Constant	0.35****	0.27****	Constant	0.22****	0.18****
	(0.016)	(0.034)		(0.023)	(0.044)
Clusters	131	62	Clusters	131	62
N	5,240	2,480	N	524	248

Notes: Includes laboratory data from OSU sample, using only observations for individuals who received photograph of partner. Standard errors are clustered at the individual level. In the question-level specification, we instrument for own ability using a two-stage approach, instrumenting for whether or not an individual answered correctly with her own share of correct answers in other questions in that bank excluding question *j* and the share of correct answers to that particular question by other non-Asian participants or Asian participants, excluding individual *i.*

C5. Willingness to Contribute Analysis

In Section 6, we explored the differences in willingness to contribute by gender. Here, we further explore this data using regression analysis and provide robustness checks on the results we presented.

First, we ask how reported beliefs map into willingness to contribute ideas to the group. Such analysis provides insights into the consequences of beliefs for group decision-making. Accordingly, we regress a participant's place in line on their beliefs about self and on the observed gender gap.[1] We first regress place in line on a set of ability proxies: own performance – instrumented for as described in Section 5.1 – and ability of the partner, proxied by male advantage in the category, partner female, and a partner-female dummy interacted with the male advantage in the category. This regression is captured by Columns I (men) and III (women) in Table A.7. We then add reported self-beliefs in Columns II and IV.

The first specification (columns I and III) shows that ability proxies are highly predictive of place in line in the expected direction. Both men and women move forward by nearly 2

[1] While it would be interesting to run specifications that include both self-belief and beliefs about partner, recall that at Harvard and UCSB participants provided *either* a self-belief *or* a partner-belief for each question. This prevents any question-level analysis that includes both self and other beliefs for most of our data.

places in line when they answer correctly. When a man is paired with a woman, the man moves forward as male advantage increases; he does not do so when paired with a man. Women move back in line as male advantage increases, but this effect is significantly stronger when paired with a man than when paired with a woman. Adding self-beliefs (Columns II and IV) captures much of the explanatory power of ability. Self-beliefs are highly predictive: a 10 percentage point increase in believed probability of answering correctly moves a participant forward in line by approximately 0.2 positions. Controlling for beliefs of own ability reduces the effect of the gender gap but it remains predictive.

Appendix Table A.7 Place in Line Decisions

Two-Stage Least Squares Predicting Place in Line
Lower Numbers Indicate Greater Willingness to Contribute

	Men		Women	
	I – No Beliefs	II – With Self-beliefs	III – No Beliefs	IV – With Self-beliefs
Male Advantage	-0.063 (0.185)	-0.15 (0.144)	2.13**** (0.175)	0.83**** (0.156)
Partner Female	0.054 (0.055)	0.002 (0.056)	-0.083 (0.053)	-0.045 (0.051)
Partner Female x Male Advantage	-0.92**** (0.267)	-0.79**** (0.201)	-1.31**** (0.248)	-0.84**** (0.205)
Own Ability (Fitted Value of $\hat{I}_{i,j}$)	-1.80**** (0.056)	-0.66**** (0.074)	-1.90**** (0.055)	-0.71**** (0.069)
Believed Probability of Self Answering Correctly		-2.01**** (0.137)		-2.09**** (0.094)
Constant	3.08**** (0.057)	3.82**** (0.094)	3.34**** (0.051)	3.97**** (0.057)
Clusters	297	297	288	288
R-squared	0.03	0.48	0.03	0.53
N	13,877	9,118	13,598	8,479

Notes: Includes laboratory data from OSU, Harvard, and UCSB, including only those individuals who knew the gender of their partner during the place in line game. Standard errors are clustered at the individual level. We instrument own ability using Equation (9), just as we do in Table III on self-beliefs.

Next, in Appendix Table A.8., we consider how these place in line decisions map into contribution outcomes. We will say that a participant "contributed" her answer if she submitted a place in line at least as close to the front of the line as her partner. Our first set of results present linear probability models predicting whether or not a participant contributed, exploring the role of gender of partner and gender stereotype of the category. In all specifications we instrument for individual ability, our fitted $\hat{I}_{i,j}$ term from Equation (9), in order to account for any role own ability plays in driving these effects.

Appendix Table A.8: Two-Stage Least Squares Predicting Participant "Contributed" Answer						
	Men			Women		
	I	II	III	IV	V	VI
Partner Female	0.053** (0.021)	0.023 (0.022)	-0.02 (0.017)	0.084**** (0.023)	0.050** (0.023)	0.036** (0.017)
Own Ability -- Fitted Value of $\hat{I}_{i,j}$	0.16**** (0.021)	0.18**** (0.021)	0.48**** (0.022)	0.25**** (0.025)	0.18**** (0.025)	0.52**** (0.023)
Male Adv.		-0.068 (0.077)	-0.004 (0.065)		-1.15**** (0.097)	-0.62**** (0.077)
Partner Female x Male Adv.		0.96**** (0.124)	0.19* (0.101)		1.17**** (0.125)	0.39**** (0.105)
Partner Place in Line			0.20**** (0.005)			0.23**** (0.005)
Constant	0.60**** (0.021)	0.59**** (0.022)	0.000 (0.029)	0.47**** (0.023)	0.54**** (0.023)	-0.15**** (0.025)
Clusters	297	297	297	288	288	288
N	13,877	13,877	13,862	13,598	13,598	13,574

Notes: Includes laboratory data from OSU, Harvard, and UCSB samples, using only observations for individuals who knew partner's gender. Standard errors are clustered at the individual level. We instrument own ability using Equation (9), just as we do in Table III on self-beliefs.

In Specifications I and IV, we look at the unconditional effect of partner gender and confirm the results reported in the main text in Section 6: both men and women contribute more answers when paired with female partners than when paired with male partners. In Specifications II and V, we add the male advantage in the category and interact it with partner gender. The results reveal that men contribute significantly more answers as male advantage increases, but only when they are paired with female partners. Women contribute significantly fewer answers as male advantage increases when paired with a male partner, but directionally more answers as male advantage increases when paired with a female partner.

Of course, whether an answer is contributed depends both upon a participant's choose of place in line and her partner's choice of place in line. Thus, the results from these specifications likely reflect both adjustments to own place in line and the fact that partners of different genders choose systematically different places line. For example, when we observe that women contribute fewer answers in sports when they are paired with a man than when they are paired with a woman, it could be because (i) the participant chooses a place farther back in line when paired with a man, and/or (ii) the male partner chooses a place closer to the front of the line than the female partner. The last set of specifications (Specifications III and VI) allow us to isolate the impact of force (i) by including a control for partner's choice of place in line. We see that conditional on partner's choice of place in line, gender of partner has a direct impact on place in line chosen by both men and women. In particular, holding fixed partner behavior, men contribute more as male advantage increases, but only when paired with women. And,

women contribute less as male advantage increases, but significantly more so when paired with men.

Appendix D: Robustness Tests

Results by Sample

First, we show the main results tables (Table III on Self-beliefs and Table IV on Beliefs about Others) separately for each laboratory sample. Standard errors are clustered at the individual level in all specifications. A few things are worth noting. First, the impact of stereotypes varies by sample. This is likely a function of the categories used in each sample, although we cannot rule out population-driven differences. Second, the impact of DIM looks quite similar at OSU and UCSB, but is stronger in self-beliefs at Harvard. Again, it is hard to identify where this is a function of the categories or the population.

Appendix Table A.9: Question-Level Self-beliefs by Sample

Two-Stage Least Squares Predicting Own Believed Probability of Answering Question Correctly

	Para-meter	OSU Men	Harvard Men	UCSB Men	Pooled Men	OSU Women	Harvard Women	UCSB Women	Pooled Women
Own Gender Advantage	$\theta\sigma$	0.38**** (0.055)	0.14* (0.078)	-0.14**** (0.029)	-0.039 (0.026)	0.17** (0.070)	0.24*** (0.084)	0.59**** (0.034)	0.49**** (0.028)
Fitted $\hat{I}_{i,j}$	ω	0.62**** (0.016)	0.37**** (0.018)	0.63**** (0.017)	0.60**** (0.011)	0.72**** (0.020)	0.44**** (0.022)	0.59**** (0.015)	0.61**** (0.011)
Constant	c	0.32**** (0.014)	0.49**** (0.017)	0.30**** (0.015)	0.33**** (0.009)	0.26**** (0.016)	0.42**** (0.018)	0.29**** (0.013)	0.30**** (0.009)
Clusters		216	128	204	548	172	124	208	504
N		8,639	2,559	12,240	23,438	6,880	2,480	12,480	21,840

Notes: Standard errors clustered at the individual level. Own gender advantage in all specifications is measured as the average gender difference in the probability of a correct answer on the bank of questions that the question is drawn from, coded so that a positive sign reflects an own gender advantage. Own ability is the fitted value of $\hat{I}_{i,j}$ from Equation (9).

Appendix Table A.10: Bank-Level Self-beliefs by Sample

OLS Predicting Believed Own Score

	Para-meter	OSU Men	Harvard Men	UCSB Men	Pooled Men	OSU Women	Harvard Women	UCSB Women	Pooled Women
Own Gender Advantage	$\theta\sigma$	1.04**** (0.111)	0.23 (0.162)	0.08** (0.034)	0.21**** (0.033)	-0.12 (0.128)	0.32* (0.184)	0.59**** (0.049)	0.44**** (0.046)
Own Score	ω	0.69**** (0.032)	0.69**** (0.050)	0.72**** (0.024)	0.71**** (0.018)	0.88**** (0.035)	0.71**** (0.049)	0.67**** (0.026)	0.71**** (0.020)
Constant	c	0.13**** (0.021)	0.16**** (0.039)	0.08**** (0.016)	0.12**** (0.012)	0.07*** (0.023)	0.14**** (0.036)	0.09**** (0.016)	0.10**** (0.012)
Clusters		216	128	204	548	172	124	208	504
N		864	512	2,448	3,824	688	496	2,496	3,680

Notes: Own gender advantage in all specifications is measured as the average gender difference in the probability of a correct answer on the bank of questions that the question is drawn from, coded so that a positive sign reflects an own gender advantage. Own ability is an individual's average probability of answering correctly in the bank. Bank-level beliefs are re-scaled to a 0 to 1 scale – that is, while an individual predicts her score on a 0 – 10 scale, we divide that belief by 10 here, so that all coefficients can be interpreted in probability points.

Appendix Table A.11: Question-Level Beliefs about Others for Different Samples									
OLS Predicting Belief of Partner's Probability of Answering a Question Correctly									
	Para-meter	OSU Beliefs about Men	Harvard Beliefs about Men	UCSB Beliefs about Men	Pooled Beliefs about Men	OSU Beliefs about Women	Harvard Beliefs about Women	UCSB Beliefs about Women	Pooled Beliefs about Women
Partner's Gender Advantage	$\theta\sigma$	0.35**** (0.063)	-0.16** (0.077)	-0.02 (0.029)	0.02 (0.027)	0.04 (0.078)	0.43**** (0.090)	0.55**** (0.042)	0.48**** (0.037)
Share of Partner's Gender Answering Question Correctly	ω	0.40**** (0.022)	0.26**** (0.017)	0.34**** (0.018)	0.34**** (0.013)	0.41**** (0.023)	0.31**** (0.019)	0.31**** (0.022)	0.33**** (0.016)
Constant	c	0.39**** (0.018)	0.53**** (0.017)	0.37**** (0.014)	0.40**** (0.010)	0.43**** (0.019)	0.49**** (0.018)	0.42**** (0.016)	0.43**** (0.012)
Clusters		108	88	199	395	85	100	213	398
N		4,320	1,760	11,940	18,020	3,399	2,000	12,780	18,179

Notes: Includes data only from participants who knew the gender of their partner at the time of providing the belief. Partner gender advantage in all specifications is measured as the average gender difference in the probability of a correct answer on the bank of questions that the question is drawn from, coded so that a positive sign reflects an advantage for the partner's gender. Partner ability is share of individuals of partner's gender that answered that question correctly.

Appendix Table A.12: Score Level Beliefs about Others for Different Samples [Normalizing Everything to 0 to 1 Scale]									
OLS Predicting Belief of Partner's Score									
	Para-meter	OSU Beliefs about Men	Harvard Beliefs about Men	UCSB Beliefs about Men	Pooled Beliefs about Men	OSU Beliefs about Women	Harvard Beliefs about Women	UCSB Beliefs about Women	Pooled Beliefs about Women
Partner's Gender Adv. in Category	$\theta\sigma$	1.69**** (0.154)	0.99**** (0.162)	0.31**** (0.054)	0.45**** (0.052)	-0.33* (0.177)	-0.49*** (0.164)	0.31**** (0.060)	0.14**** (0.055)
Partner's Gender Average Score in Category	ω	0.63**** (0.082)	0.81**** (0.077)	0.66**** (0.066)	0.64**** (0.043)	0.93**** (0.097)	0.69**** (0.064)	0.57**** (0.050)	0.62**** (0.037)
Constant	c	0.14*** (0.049)	0.10** (0.045)	0.13**** (0.036)	0.16**** (0.024)	0.14**** (0.057)	0.22**** (0.042)	0.21**** (0.028)	0.21**** (0.021)
Clusters		108	88	199	395	85	100	213	398
N		432	352	1,806	2,590	340	400	1,890	2,630

Notes: Includes data only from participants who knew the gender of their partner at the time of providing the belief. Standard errors clustered at the individual level. Partner gender advantage in all specifications is measured as the average gender difference in the probability of a correct answer on the bank of questions that the question is drawn from, coded so that a positive sign reflects an advantage for the partner's gender. Partner ability is the average probability of answering correctly in the 10-question bank by members of the partner's gender. Note that bank-level beliefs are re-scaled to a 0 to 1 scale – that is, while an individual predicts her partner's score on a 0 – 10 scale, we divide that belief by 10 here, so that all coefficients can be interpreted in probability points.

Restriction to US High School Sample

Next, we show that results are quite similar when restricted to the sample that attended high school in the United States. Note that we pre-registered this as a restriction at UCSB, so

the exclusion for that sub-sample is already reflected in our main estimates. Appendix Table A.13 shows the results for self-beliefs, which look quite similar to the results for the full sample.

Appendix Table A.13: Replication of Self-beliefs with US HS Participants

Question-Level Beliefs OLS Predicting Own Believed Probability of Answering Correctly US HS ONLY				**Bank-Level Beliefs** OLS Predicting Own Believed Score on 0 to 1 Scale US HS ONLY			
	Para-meter	I (Men)	II (Women)		Para-meter	III (Men)	IV (Women)
Own Gender Adv.	$\theta\sigma$	-0.045* (0.026)	0.53**** (0.029)	Own Gender Adv.	$\theta\sigma$	0.20**** (0.034)	0.48**** (0.047)
Fitted Value of $\hat{I}_{i,j}$	ω	0.59**** (0.011)	0.59**** (0.011)	Individual's Score in Category	ω	0.71**** (0.019)	0.68**** (0.021)
Constant	c	0.34**** (0.010)	0.31**** (0.009)	Constant	c	0.12**** (0.013)	0.11**** (0.013)
Clusters		493	429	Clusters		493	429
N		21,573	19,180	N		3,604	3,380

Notes: Pools observations for OSU, Harvard, and UCSB experiments. Standard errors clustered at the individual level.

In Table A.14, we replicate the results on beliefs about others using only the sub-sample of participants that attended high school in the United States. The results are very similar to the results for the full sample.

Appendix Table A.14: Replication of Table IV Beliefs about Others

Question-Level Beliefs OLS Predicting Belief of Partner's Probability of Answering Correctly in Part 3 US HS ONLY				**Bank-Level Beliefs** OLS Predicting Belief of Partner's Part 1 Score US HS ONLY			
	Para-meter	I (Beliefs about Men)	II (Beliefs about Women)		Para-meter	III (Beliefs about Men)	IV (Beliefs about Women)
Partner's Gender Adv.	$\theta\sigma$	0.018 (0.027)	0.49**** (0.038)	Partner's Gender Adv.	$\theta\sigma$	0.41**** (0.052)	0.18**** (0.055)
Share of Partner's Gender Answering Qn. Correctly	ω	0.35**** (0.014)	0.32**** (0.017)	Partner's Gender Avg, Score in Category	ω	0.65**** (0.046)	0.60**** (0.038)
Constant	c	0.39**** (0.011)	0.43**** (0.012)	Constant	c	0.15**** (0.026)	0.22**** (0.022)
Clusters		347	369	Clusters		347	369
N		16,420	17,259	N		2,398	2,514

Notes: Includes data only from participants who knew the gender of their partner. We pool observations from OSU, Harvard, and UCSB. Standard errors are clustered at the individual level.

Robustness to Slider Scale Perceptions and Large Gaps

In Section 5.4, we considered the fact that noisily estimated gender gaps have the potential to complicate our identification of the stereotypes term in our main results tables (Tables III

and IV). In this sub-section, we explore the extent to which are results are robust to (i) replacing observed gaps with slider scale perceptions and (ii) using only using categories with large gaps.

Appendix Table A.15: Self-beliefs using Slider Scale Perceptions

Question-Level Beliefs Two-Stage Least Squares Predicting Own Believed Probability of Answering a Question Correctly				**Bank-Level Beliefs** OLS Predicting Own Believed Score on 0 to 1 Scale			
	Para-meter	I (Men)	II (Women)		Para-meter	III (Men)	IV (Women)
Slider Scale Perception of Own Gender Advantage	$\theta\sigma$	0.01 (0.006)	0.09**** (0.006)	Slider Scale Perception of Own Gender Advantage	$\theta\sigma$	0.04**** (0.007)	0.11**** (0.009)
Own Ability - Fitted Value of $\hat{I}_{i,j}$	ω	0.60**** (0.011)	0.63**** (0.011)	Own Ability –Own Average Probability of Correct Answer in Bank	ω	0.70**** (0.018)	0.69**** (0.019)
Constant	c	0.33**** (0.009)	0.27**** (0.009)	Constant	c	0.13**** (0.012)	0.09**** (0.011)
Clusters		547	504	Clusters		547	504
N		23,398	21,840	N		3,820	3,680

Notes: Pools observations for Ohio State, Harvard, and UCSB. Standard errors clustered at the individual level. We recode the slider scale so that positive numbers indicate a believed advantage for own gender. Own ability for question-level data is the fitted value of $\hat{I}_{i,j}$ from Equation (9) but replacing observed gender gap with the slider scale perception, and, in bank-level data, own ability is an individual's average probability of answering correctly in the bank. Note that bank-level beliefs are re-scaled to a 0 to 1 scale – that is, while an individual predicts her score on a 0 – 10 scale, we divide that belief by 10 here, so that all coefficients can be interpreted in probability points.

Appendix Table A.16: Beliefs about Others using Slider Scale Perceptions							
Question-Level Beliefs OLS Predicting Belief of Partner's Probability of Answering a Question Correctly				**Bank-Level Beliefs** OLS Predicting Belief of Partner's Score			
	Para-meter	I (Beliefs About Men)	II (Beliefs About Women)		Para-meter	III (Beliefs About Men)	IV (Beliefs About Women)
Slider Scale Perception of Partner Gender Advantage	$\theta\sigma$	0.02*** (0.006)	0.08**** (0.009)	Slider Scale Perception of Partner Gender Advantage	$\theta\sigma$	0.09**** (0.011)	0.06**** (0.010)
Partner Ability - Share of Partner's Gender Answering Qn. Correctly	ω	0.34**** (0.013)	0.35**** (0.016)	Partner Ability - Partner's Gender Average Probability of Correct Answer in Bank	ω	0.67**** (0.043)	0.55**** (0.035)
Constant	c	0.40**** (0.010)	0.41**** (0.012)	Constant	c	0.15**** (0.024)	0.24**** (0.018)
Clusters		394	398	Clusters		394	398
N		17,890	18,179	N		2,586	2,630

Notes: Includes data only from participants who knew the gender of their partner at the time of providing the belief. Pools observations for Ohio State, Harvard, and UCSB. Standard errors clustered at the individual level. We recode the slider scale so that positive numbers indicate a believed advantage for partner gender. Partner ability for question-level data is share of individuals of partner's gender that answered that question correctly and, in bank-level data, partner ability is the average probability of answering correctly in the 10-question bank by members of the partner's gender. Note that bank-level beliefs are re-scaled to a 0 to 1 scale – that is, while an individual predicts her partner's score on a 0 – 10 scale, we divide that belief by 10 here, so that all coefficients can be interpreted in probability points.

Appendix Table A.17 looks at the coefficient on the stereotypes term under different exclusion restrictions: first, restricting to banks of questions that have a gender gap of at least 5 percentage points; second, restricting to banks of questions that have a gender gap of at least 10 percentage points. We do this for both question-level beliefs (Panel a) and bank-level beliefs (Panel b). In general, we estimate a larger effect of stereotypes as we restrict attention to domains with larger gender gaps. However, the estimates are not dramatically changed, with the exception of the bank-level estimates of the extent of stereotyping of women, which are estimated to be much larger when gaps are large. This suggests that noisily estimated gaps are not playing a large role in driving our results.

Appendix Table A.17. Stereotype Coefficient Estimates when Restricted to Large Gender Gaps						
	Question-Level Beliefs			**Bank-Level Beliefs**		
	All data	Gap of at least 5pp	Gap of at least 10pp	All data	Gap of at least 5pp	Gap of at least 10pp
Men						
Self-beliefs	-0.039 (0.026)	-0.020 (0.028)	0.017 (0.029)	0.21**** (0.034)	0.21**** (0.034)	0.21**** (0.035)
Beliefs about Men	0.02 (0.027)	0.04 (0.028)	0.07** (0.028)	0.45**** (0.052)	0.40**** (0.052)	0.41**** (0.051)
Women						
Self-beliefs	0.49**** (0.028)	0.48**** (0.031)	0.47**** (0.032)	0.44**** (0.046)	0.44**** (0.048)	0.48**** (0.051)
Beliefs about Women	0.48**** (0.037)	0.49**** (0.042)	0.45**** (0.041)	0.14**** (0.055)	0.27**** (0.064)	0.55**** (0.078)

Notes: This table reports the estimated coefficient $\theta\sigma$ from a series of regressions that either (i) do not restrict the data, (ii) restrict the data to observations from banks with at least a 5pp gender gap, or (iii) restricts the data to observations from banks with at least a 10pp gender gap. Pools observations for Ohio State, Harvard, and UCSB. Standard errors clustered at the individual level. Own ability for question-level data is the fitted value of $\hat{I}_{i,j}$ from Equation (9), and, in bank-level data, own ability is an individual's average probability of answering correctly in the bank. For beliefs about others, specifications include data only from participants who knew the gender of their partner at the time of providing the belief. Partner ability for question-level data is share of individuals of partner's gender that answered that question correctly and, in bank-level data, partner ability is the average probability of answering correctly in the 10-question bank by members of the partner's gender. Note that bank-level beliefs are re-scaled to a 0 to 1 scale – that is, while an individual predicts her partner's score on a 0 – 10 scale, we divide that belief by 10 here, so that all coefficients can be interpreted in probability points.

MTurk Replication

In Appendix Tables A.18 and A.19, we replicate the bank-level beliefs using the MTurk data. Recall from Appendix B that the MTurk experiment features questions from the six categories from OSU and Harvard: art, emotion recognition, verbal, business, math, and sports. Participants are asked to guess their own score in each 5-question bank, as well as the score of a randomly-chosen man and a randomly-chosen woman. Thus, the paradigm is different than the laboratory paradigm, where participants never assess both a male and female other.

In general, DIM looks much more severe for MTurk participants. This could reflect the increased noise for a 5-question bank, or other features of the population. We estimate that stereotypes shape women's self-beliefs, and beliefs about men, similar to what we find in the laboratory. However, for beliefs about women and men's self-beliefs, we see no evidence of stereotypes.

Table A.18: OLS Predicting Own Believed Score on 0 to 1 scale				
	Laboratory		Mechanical Turk	
	I (Men)	II (Women)	III (Men)	IV (Women)
Own Gender Adv.	0.21**** (0.033)	0.44**** (0.046)	-0.094**** (0.011)	0.29**** (0.0141.58)
Individual's Score in Category on 0 to 1 scale	0.71**** (0.018)	0.71**** (0.020)	0.47**** (0.014)	0.46**** (0.014)
Constant	0.12**** (0.012)	0.10**** (0.012)	0.34**** (0.010)	0.32**** (0.011)
Clusters	548	504	987	843
N	3,824	3,680	5,922	5,064

Notes: Pools observations for OSU, Harvard, and UCSB experiments. Standard errors clustered at the individual level.

Appendix Table A.19: OLS Predicting Belief of Partner's Score on 0 to 1 Scale				
	Beliefs about Men		Beliefs about Women	
	Lab	Mturk	Lab	Mturk
	I	II	III	IV
Partner's Gender Adv.	0.45**** (0.052)	0.21**** (0.010)	0.14**** (0.055)	0.006 (0.004)
Partner's Gender Avg. Score	0.64**** (0.043)	0.65**** (0.020)	0.62**** (0.037)	0.41**** (0.012)
Constant	0.16**** (0.024)	0.12**** (0.014)	0.21**** (0.021)	0.63**** (0.014)
Clusters	395	1,826	398	1,826
N	2,590	10,986	2,630	10,986

Notes: Laboratory specifications include laboratory data from OSU, Harvard, and UCSB samples, using only observations for individuals who knew partner's gender. Standard errors are clustered at the individual level.