NBER WORKING PAPER SERIES

BELIEFS ABOUT GENDER

Pedro Bordalo
Katherine B. Coffman
Nicola Gennaioli
Andrei Shleifer

Working Paper 22972
http://www.nber.org/papers/w22972

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2016

We are grateful to James Pappas, Annie Kayser, and Paulo Costa for significant help with experiments, to Benjamin Enke, Josh Schwartzstein, and Neil Thakral for comments and to the Pershing Square Venture Fund for Research on the Foundations of Human Behavior for financial support of this research. Gennaioli thanks the European Research Council for financial support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research. Research funding and support was provided by the Economics department at Ohio State University.

Beliefs about Gender
Pedro Bordalo, Katherine B. Coffman, Nicola Gennaioli, and Andrei Shleifer
NBER Working Paper No. 22972
December 2016
JEL No. C91,D01,J16

ABSTRACT

We conduct a laboratory experiment on the determinants of beliefs about own and others' ability across different domains. A preliminary look at the data points to two distinct forces: miscalibration in estimating performance depending on the difficulty of tasks and gender stereotypes. We develop a theoretical model that separates these forces and apply it to analyze a large laboratory dataset in which participants estimate their own and a partner's performance on questions across six subjects: arts and literature, emotion recognition, business, verbal reasoning, mathematics, and sports. We find that participants greatly overestimate not only their own ability but also that of others, suggesting that miscalibration is a substantial, first order factor in stated beliefs. Women are better calibrated than men, providing more accurate estimates of ability both for themselves and for others. Gender stereotypes also have strong predictive power for beliefs, particularly for men's beliefs about themselves and others' beliefs about the ability of men. Our findings help interpret evidence on gender gaps in self-confidence.

Pedro Bordalo
Saïd Business School
University of Oxford
Park End Street
Oxford, OX1 1HP
United Kingdom
pedro.bordalo@sbs.ox.ac.uk

Katherine B. Coffman
445 Baker Library
Harvard Business School
Boston, MA 02163
kcoffman@hbs.edu

Nicola Gennaioli
Department of Finance
Università Bocconi
Via Roentgen 1
20136 Milan, Italy
nicola.gennaioli@unibocconi.it

Andrei Shleifer
Department of Economics
Harvard University
Littauer Center M-9
Cambridge, MA 02138
and NBER
ashleifer@harvard.edu

1. Introduction

Beliefs about ourselves and others are at the heart of many economic and social decisions. One critical area where such beliefs are often found to be biased is abilities of men and women. Controlling for performance, women are less confident about their own ability in math and science then men, contributing to differences in financial decision-making, academic performance, and career choices (Barber and Odean 2001, Buser, Niederle, and Oosterbeek 2014). But what explains these gender gaps in self-confidence? Do they vary by field? How are beliefs about oneself related to beliefs about others? The existing research typically focuses on beliefs in limited domains, and on beliefs either only about oneself or only about others, making it hard to isolate the sources of gaps in self-confidence. In this paper, we try to uncover the sources of differences in beliefs that women and men hold about their own ability and that of others in multiple domains of knowledge.

We do so using new experimental evidence on beliefs that extends earlier findings of Coffman (2014). In our experiments, participants answer multiple-choice trivia questions in a variety of categories: art and literature, emotion recognition, business, verbal skills, mathematics, and sports and games. They are then asked to estimate either their total score on each of the 10-question category quizzes, or the probability of answering a given question correctly. They also provide beliefs about the performance of a randomly-selected partner. For some participants, the gender of this partner is revealed, although we take some pains not to focus attention on gender. For every participant, we thus have direct measures of their own performance in multiple categories of knowledge, but also their estimates of both their own performance and that of a partner. We can also estimate the difficulty of the questions from how many people answer them correctly, and relate beliefs about performance to question difficulty.

A preliminary look at the data reveals two striking findings that shape our more detailed analysis. First, both men and women overestimate performance in all domains, particularly for

difficult questions (for which the share of correct answers is low). Such overestimation occurs both when evaluating oneself and evaluating others. This suggests that misperception of ability of both self and others plays a role in belief formation. Second, there are differences in overestimation across domains. Conditional on both question difficulty and own performance, men overestimate own performance more in male-typed domains than in female-typed domains. When evaluating others, participants also overestimate the performance of men more in male-typed domains than in female-typed domains. This suggests that stereotypes – exaggerations of true differences in ability between genders -- also play a role in shaping beliefs.

To understand this evidence, we construct a model that takes into account both misperceptions of ability related to question difficulty and stereotype distortions and estimate it with our experimental data. To incorporate misperceptions – which we call Difficulty Influenced Miscalibration or DIM – we follow Moore and Healy (2008), who analyzed this distortion in a context unrelated to gender. To incorporate gender stereotypes, we follow Bordalo et al. (2016) and model belief formation based on Kahneman and Tversky's (1973) representativeness heuristic. The model treats beliefs about self and others in a unified framework.

We then bring the model to the data. For identification, we adopt the perspective that the effects of DIM on beliefs about performance are orthogonal to those of stereotypes. The former depends only on task difficulty, whereas the latter depends only on task domain. Thus, comparing easy and difficult questions in mathematics should reveal the role of misperceptions. In contrast, comparing easy questions in mathematics with those in verbal tasks should reveal the role of stereotypes. Because respondents answer questions of varying difficulty within categories but stereotypes differ across categories, we can identify both DIM and stereotypes under this orthogonality assumption. While an approximation, this approach captures the essential difference between the two hypotheses.

Consistent with Moore and Healy (2008), we find a consistent and strong role for DIM in shaping beliefs. In evaluating both themselves and others, participants overestimate ability, particularly in more difficult questions and categories. At the same time, women are better calibrated than men, providing more accurate estimates of their own ability and that of others. We also find that gender stereotypes are an important source of belief distortions, in many cases significantly improving the fit of the model beyond DIM alone. Stereotypes are especially important in categories perceived as male-typed, such as mathematics and sports. Stereotyping has more predictive power for men, both in terms of beliefs about their own ability and others' beliefs about men, than it does for women. Our results thus show that two important but distinct forces contribute to the often-discussed gender gap in self-confidence.

Our data and our model speak directly to beliefs about absolute ability: estimated own performance and estimated performance of others. In many contexts of interest, however, beliefs about relative ability may best predict decisions – an individual's estimate of her ranking in a pool of potential job candidates, or her assessment of her ability relative to peers or competitors. We consider the implications of our findings for beliefs of relative ability and decision-making using our data on beliefs about both oneself and others. In our data, men are much more likely than women to believe their own ability exceeds that of a partner. For both men and women, beliefs about relative ability and decisions are a function of both the category and the gender of one's partner, over and above true gender gaps in performance, suggesting a prominent role for stereotypes. Moreover, these exaggerated beliefs translate into decisions: in our group decision-making environment, women contribute answers less often when paired with male than with female partners, particularly in male-typed categories.

Our paper follows an enormous literature on beliefs about gender. Coffman (2014) shows that decisions about willingness to contribute ideas to a group are predicted by gender stereotypes. Conditional on performance, participants' beliefs about own relative ability and

decisions to contribute ideas to a group are predicted by the gender-type of the category, revealing a major role for self-stereotyping in self-assessments of ability. We expand on this work by formally modeling stereotypes and miscalibration, collecting more data on beliefs about self and others, including in situations where the gender of one's partner is known, and estimating and evaluating a formal model of belief formation.

Other past work suggests a role for both stereotypes and DIM in shaping beliefs about both one's own and others' ability. Many studies find that gender stereotypes in math and science influence academic performance (see Kiefer and Sekaquaptewa 2007 and Nosek et al 2009 on implicit bias and test performance and Spencer, Steele and Quinn 1999 on stereotype threat). Both experimental and field evidence document a widespread belief that women have lower ability than men in math (Eccles, Jacobs, and Harold 1990, Guiso, Monte, Sapienza and Zingales 2008, Carrell, Page and West 2010, Reuben, Sapienza and Zingales 2014), although the differences have been shrinking and now only exist at the upper tail (Goldin, Katz and Kuziemko 2006). Guiso et al. (2008) find that actual male advantage in math disappears in cultures where gender stereotypes are weaker.

Research shows that men are generally more overconfident than women, with larger differences in domains perceived to be male-typed (Lundeberg, Fox, and LeCount 1994, Deaux and Farris 1997, Pulford and Colman 1997, Beyer 1990, Beyer and Bowden 1997, Beyer 1998, Coffman 2014). Research on gender differences in competition shows that men are more confident and more willing to compete in male-typed tasks, but that these differences are reduced or reversed in female-typed tasks (see Gneezy, Niederle, and Rustichini 2003 and Niederle and Vesterlund 2007 for the original findings, and Grosse and Reiner 2010, Dreber, Essen, and Ranehill 2011, and Shurchkov 2012 for the exploration of male versus female-typed tasks).

Notwithstanding the tremendous amount of work on gender differences in over-confidence, the sources of such differences remain uncertain, and it is our principal goal is to try to unpack

them. From a methodological standpoint, disentangling the sources of gender differences in over-confidence requires a multi-dimensional dataset. We collect data on beliefs about both own and others' ability, across a range of question difficulties and a variety of domains. We examine these data in light of a theoretical model that allows for identification of distinct contributors to biased beliefs. This leads to three principal findings. First, beliefs about others are extremely similar to those about selves; self-motivated beliefs cannot be the principal source of over-confidence. Second, Difficulty Influenced Miscalibration is extremely important in shaping belief distortions, for women but even more so for men. Third, stereotypes play a significant role in belief distortion and, at least in our sample, stereotype distortions are more pronounced for men than they are for women. It is clear from the evidence that the sources of differences in over-confidence are a matter of at least two-fold, and there may be other factors that have escaped our scrutiny.

In Section 2, we describe our experiment. In Section 3, we show some preliminary evidence. Section 4 then presents a formal model, while Section 5 presents our estimates of the model using the experimental data. Section 6 looks at beliefs about relative performance and strategic decisions. Section 7 concludes.

2. Experiments and Performance

2.1 Experimental Design

We ran two laboratory experiments to gather data on stated beliefs, one conducted at Ohio State University and one conducted at Harvard Business School (but with most subjects being Harvard College undergraduates). Our goal is to collect detailed data on beliefs about both own and others' ability across a variety of domains and to link these beliefs to strategic decisions. The full instructions and materials for each experiment are provided in Appendix A.

Both experiments had the same basic three part structure, closely resembling the experimental design of Coffman (2014). In Part 1, each participant answers questions in each

category, giving us a measure of their ability in each category. Participants are then randomized into groups of two. In Part 2, we use the procedure developed by Coffman (2014) to measure willingness to contribute answers to their group. In Part 3, we collect incentivized data on beliefs about their own and their partners' ability in each category. The categories vary in their associated gender stereotype. At Ohio State, we use Arts and Literature, Verbal Skills, Mathematics, and Sports and Games; at Harvard, we replace Verbal Skills and Mathematics with Emotion Recognition and Business.

Our prior was that Arts and Literature, Emotion Recognition, and Verbal Skills would be categories that participants considered female-typed, believing women to have an advantage on average, while Business, Mathematics, and Sports and Games would be considered male-typed. Coffman (2014) found that participants perceived Arts and Literature as female-typed and Sports and Games as male-typed, in line with observed performance differences in the sample. Our priors for Verbal Skills and Mathematics are guided by both observed gender differences on large-scale standardized tests such as the SAT (see http://media.collegeboard.com/digitalServices/pdf/research/2013/TotalGroup-2013.pdf for data) and by prior experimental studies that have successfully varied the perceived gender-type of the task by using verbal versus math tasks (for instance, Shurchkov 2012). Neuroscientists and psychologists have identified a female advantage in the ability to recognize emotion (Hall and Matsumoto 2004). We confirm these priors by asking participants at the end of the experiment which gender knows more about each category.[2]

Participants complete the experiment using a laboratory computer at an individual station and can work at their own pace. In each part, they can earn points. At the end of the experiment,

[2] We ask participants at the end whether they believe men or women, on average, know more about each category. They use a sliding scale ranging from -1 to 1 to indicate their answer, with -1 labeled as "women know more" and 1 labeled as "men know more". Indeed, participants report Arts and Literature, Verbal Skills, and Emotion Recognition as areas of female advantage (means of -0.30, -0.28, -0.18, respectively) and Business, Mathematics, and Sports and Games as areas of male advantage (means of 0.15, 0.18, 0.50, respectively).

one part is randomly chosen for payment; participants receive a fixed show-up fee and additional pay for every point earned in the selected part.[3]

The key departure from Coffman's (2014) experiment is that when participants are assigned to groups, we randomly vary whether the gender of one's partner is revealed. This allows us to collect direct measures of beliefs about male and female performance. To do this, we must reveal the gender of one's partner. We sought to avoid messages that explicitly referred to gender (for instance, "your partner is a woman"), as we worried about possible experimenter demand effects and making participants aware of the gender focus of the experiment. At Ohio State, we used photos to convey gender. By providing a photo of a partner, we convey gender but may also introduce confounds – for instance, by triggering a reduction in social distance between partners (Bohnet and Frey 1999) or by rendering race or attractiveness top of mind.

For that reason, in the Harvard experiment, we introduced a novel method for revealing gender – and only gender. At the moment of assignment to groups, the experimenter announced each pairing by calling out the two participant numbers. In the treatment where gender was not revealed, the experimenter simply announced the pairings. In the treatment where gender is revealed, participants were asked to call out, "Here", when their participant number was announced. Because of the station partitions within the laboratory, it is highly likely that in this treatment, a participant could hear the voice of his or her assigned partner, but not see them. By restricting to the word, "Here", we limit the amount of conveyed information (through tone of voice, friendliness, etc.). In this way, only gender is likely to be revealed.[4] In analyzing the data,

[3] At Ohio State, participants earned a $5 show-up fee plus an additional dollar for every point earned in the selected part. At Harvard, participants earned a $10 show-up fee, $15 for completing the experiment, and an additional $0.25 for every point earned in the selected part. These differences reflect the requirements related to minimum and average payments at the two laboratories.

[4] We validate this approach by asking a subset of participants at the conclusion of the experiment to guess the gender and ethnicity of their partner. Participants are significantly more likely to identify the gender of their partner in treatments where the voice is heard (correctly identified in 92% of cases where voice is revealed compared to 67% of cases where voice is not revealed, $p<0.0001$); they are not significantly more likely to identify ethnicity (correctly identified in 47% of cases where voice is revealed compared to 39% of cases where voice is not revealed, $p=0.25$).

we group all participants who received a photo or heard a voice as our "knew gender" treatment, performing an intent-to-treat analysis.

We designed the experiment to minimize the extent to which participants were focused on gender. Participants see no questions that refer to gender until the final demographic and debriefing questions at the end of the experiment. Relative to a paradigm where gender is made prominent, our findings may underestimate the importance of stereotypes, but we can be more confident that the effects we observe are not due to experimenter demand.

Part 1 Participants answer 40 multiple-choice questions, 10 in each category. Each question has five possible answers. Participants receive 1 point for a correct answer and lose 1/4 point for an incorrect answer; they must provide an answer to each question. All questions appear on the same page, labeled with their category, in a random order. The goal of Part 1 is simply to collect a measure of individual ability in each category.

Intervention Following completion of Part 1, participants are told that they have been randomly assigned to groups of two. Participants in the control condition receive no further information about their partners. Treated participants at Ohio State are given a photo of the partner, and at Harvard they hear the partner answer a role call with the single word "here". At Ohio State, treatment assignment occurs at the participant level; due to the nature of the intervention, treatment assignment occurs at the session level at Harvard.

Following the intervention in each experiment, participants are asked to estimate their own and their partner's performance in Part 1. For each category, they are asked to guess the number of correct Part 1 answers both they and their partner had. Participants receive an additional point for every correct guess, incentivizing them to give the guess they think is most likely to be correct.

Part 2 Participants make decisions about their willingness to contribute answers to new questions in each category to their group. Participants are given 40 new questions, 10 in each category. As in Part 1, all questions appear on the same page, in a randomized order, labeled with

their category. For each question, participants must indicate both their answer to the question and how willing they are to have it count as the group answer. Both partners earn 1 point if the group answer submitted is correct and lose 1/4 point if the group answer submitted is incorrect.

To measure willingness to contribute, we use the "choose a place in line" procedure introduced in Coffman (2014). For each question, participants are asked to choose a place in line between 1 and 4. The participant who submits the lower place in line for that question has her answer submitted as the group answer. To break ties, the computer flips a coin to determine whose answer is submitted. Choosing a lower place in line weakly increases the probability that one's answer is submitted for the group. We interpret place in line as willingness to contribute one's answer for the group.

Part 3 We collect data on question-specific beliefs from participants. Participants revisit the same 40 questions from Part 2. For each question, they are asked to estimate (a) the probability of their own answer being correct and/or (b) the probability of their partner's answer being correct. Keep in mind that participants are not aware of what answers their partner has chosen. Depending on their treatment, some participants know their partner's gender and others do not.

We apply the incentive-compatible belief elicitation procedure used by Mobius, Niederle, Niehaus, and Rosenblatt (2014), implemented exactly as in Coffman (2014). Participants see each of the 40 questions they saw in Part 2 again. At Ohio State, for every question they are asked to provide their believed probability of answering correctly, and their believed probability their partner answered correctly. At Harvard, we split these belief elicitations into two parts. For 20 of the 40 questions, (5 in each category faced by the participant), they provide their own believed probability of answering correctly. For the remaining 20 questions, they provide their believed probability of their partner answering correctly. This is done as a separate section of the experiment. For each mode of belief solicitation, truth-telling is profit-maximizing regardless of the participant's risk preferences (details in Appendix A).

Following the completion of the beliefs sections, participants answer demographic questions about themselves. Participants receive no feedback throughout the course of the experiment. When they have completed the study, they receive an answer key that reveals the answers to each question they saw. They also see what answers were submitted for the group in Part 2 and which group member submitted them. Participation lasted approximately 90 minutes and average earnings were approximately $30 per participant.

3. A Look at the Data

As a first step, we present some raw data from the experiment, focusing on ability and beliefs. We explore how these measures vary by gender, category, and question difficulty. Table 1 presents summary statistics on our participants.

Table I. Summary Statistics

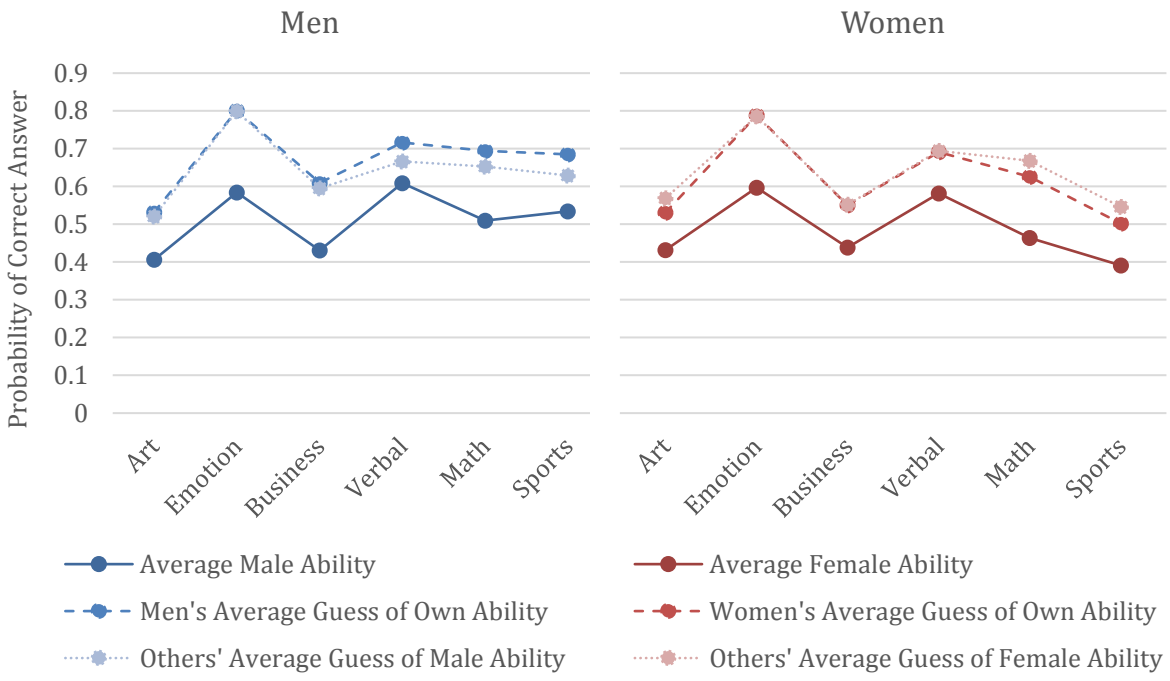
	Men	**Women**	**p value**
Proportion Harvard Participants	0.37	0.42	0.23
Current Undergraduate	0.86	0.84	0.56
Attended US High School	0.84	0.75	0.003
Ethnicity:			
Caucasian	0.60	0.43	0.00
East Asian	0.18	0.35	0.00
Black or African American	0.08	0.07	0.95
Asian Indian	0.05	0.03	0.25
Treatment:			
Proportion Known Male Partner	0.27	0.31	0.26
Proportion Known Female Partner	0.30	0.31	0.82
Proportion Unknown Partner	0.43	0.38	0.21
N	344	296	

Notes: Two participants at Ohio State dropped out when photographs were taken. One participant at Ohio State was caught cheating (looking up answers on the internet); she was dismissed. One participant at Ohio State was unable to complete the experiment due to a computer failure. All observations from these participants and their randomly-assigned partners are excluded from the analysis. P-value is given for the null hypothesis of no difference between genders using a two-tailed test of proportions.

In our sample, men are significantly more likely to have attended a U.S. high school, more likely to be white, and less likely to be East Asian. Appendix D shows that our results are similar in a more ethnically-balanced sample of men and women who attended high school in the U.S.

We begin by exploring average ability and average beliefs across gender and categories. We use Part 3 data, with question-specific data for each individual. In Figure I, we aggregate the data

by category, asking how stated beliefs compare with observed ability. In Panel (a), we plot men's average probability of answering correctly in each category, their average believed probability of themselves answering correctly, and the average of others' believed probability of men answering the question correctly. The others' belief measure averages across the "partner beliefs" of all individuals in the known gender treatment paired with a male partner. Panel (b) presents the corresponding data for women. Categories are ordered by the average gender gap in performance in a category over the entire experiment, from smallest to largest male advantage.



Panel (a) Panel (b)

Figure I

The first order observation from Figure I is that stated beliefs, both about self and about others, far exceed observed ability across all categories. While the average probability of answering a question correctly is approximately 0.49 (SD 0.50) in our sample, the average believed probability of answering correctly is 0.63 (SD 0.31). Surprisingly, beliefs about others also dramatically exceed observed ability, averaging 0.62 (SD 0.25). Differences between

beliefs about self and about others are quite small in comparison to the large gap between ability and beliefs in general. Participants greatly overestimate the ability of both themselves and others across categories – a central fact in these data.[5]

We also observe that categories are of similar difficulty for men and women. In particular, emotion and verbal are easier categories for both men and women, while art, math, and business are harder for both genders. There is more divergence for sports, which appears to be a relatively easier category for men than for women. Overestimation of ability, both for self and for others, is persistent across the range of category difficulties.

Finally, men's beliefs about themselves typically exceed others' beliefs about them, particularly as male advantage increases. For women, the opposite pattern emerges: women's beliefs about themselves are typically more pessimistic than others' beliefs about them.

Figure I reveals several patterns critical to the interpretation of the evidence. Most important is the substantial overconfidence of the participants not just about themselves, but also about others. Such evidence is unlikely to be explained just by self-confidence or another form of motivated beliefs. Rather, it is a general over-estimation of ability. Second, some areas are more difficult than others, and beliefs about both self and others adjust to differential difficulty. This too needs to be taken into account. Yet, even in Figure I, some differences between men and women emerge. Although all participants are overconfident about both themselves and others, men are relatively more over-confident about themselves than others are about them, and women are relatively less over-confident about themselves than others are about them.

[5] One might worry that beliefs about self anchor reported beliefs about others, leading to our findings. We attempted to address this concern in our Harvard experiment. Rather than elicit beliefs about self and other simultaneously, we asked for beliefs about themselves for a subset of 20 questions in Part 3. In a separate section of the experiment without access to their past answers, participants provided beliefs about their partner for a *separate* subset of 20 questions. Even with this design, we observe similar levels of overestimation across own and partner ability (16 pp for own ability, 14 pp for ability of others).

In Figure II, we reshape the same data, this time zooming in on differences across genders. The solid orange line represents the observed male advantage in performance in each category: male performance significantly exceeds female performance only in math and sports. Performance gaps are small and statistically insignificant in the other categories.

The male advantage in self beliefs (the difference in men's average believed probability of answering correctly and women's average believed probability of answering correctly, graphed as the dashed blue line) is directionally larger than the performance gap in every category. Relative to the difference in performance, the difference in beliefs about self is particularly exaggerated for business and sports.

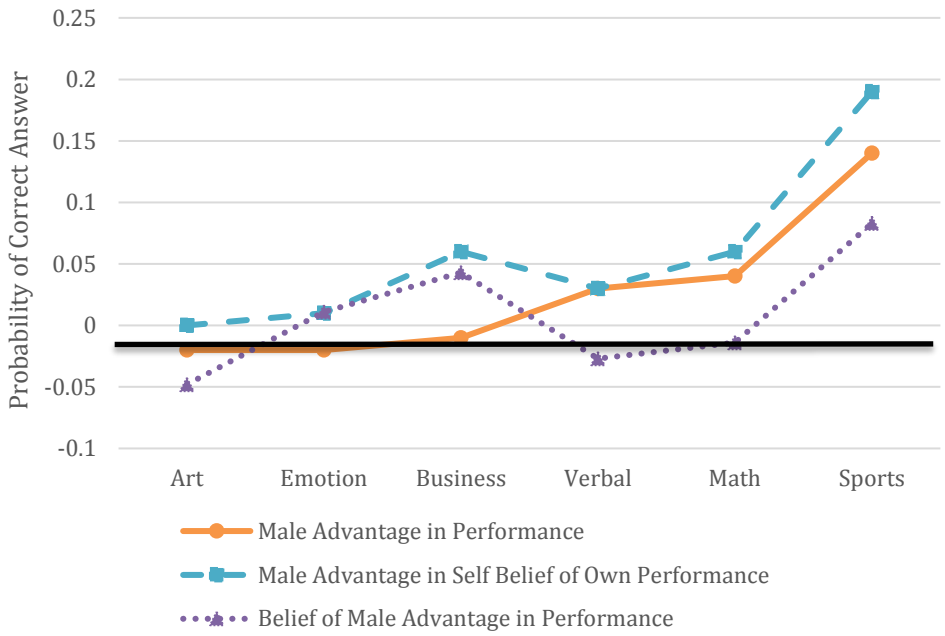


Figure II

The dotted purple line reflects differences in beliefs about men and beliefs about women (difference between the average believed probability that a male partner answers correctly and the average believed probability a female partner answers correctly). On average, participants believe women have an advantage in art, verbal, and math, and believe men have an advantage

in business and sports. In four of the six categories (art, verbal, math, and sports) participants believe the male advantage in performance is smaller than it actually is.

In Figure III, we examine the relationship between question difficulty and beliefs about self. Figure III plots the believed share of correct answers in terms of self beliefs (y-axis) by the observed share of correct answers (x-axis), doing this separately for men and women for each question. Figure III leaves no doubt that men are more self-confident than women, but particularly so on the more difficult questions. In fact, this has clear implications for the gender gap in overconfidence.

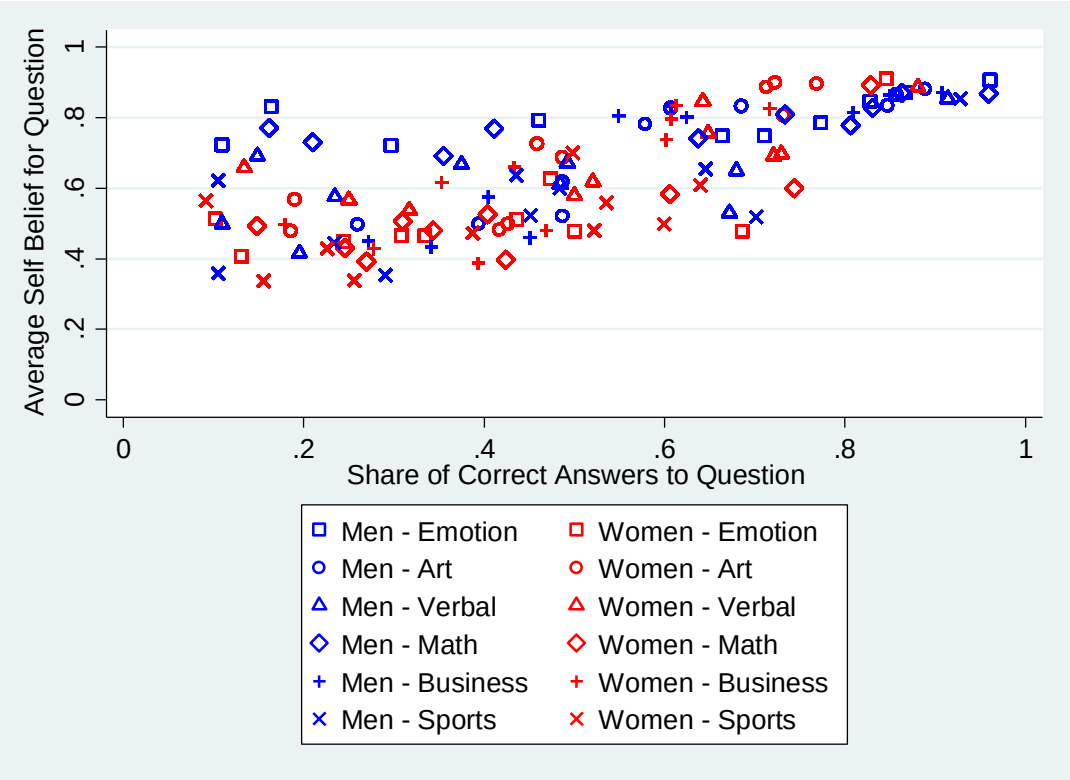


Figure III

In Figure IV, we directly explore the pivotal role that question difficulty plays in determining this gap. We create two broad buckets of questions: questions of below median difficulty (45% or more of the participants answered the question correctly) and questions of above median difficulty. For each category, we then compute the gender gap in overestimation of own ability for both these easier and harder questions. Overestimation is calculated as the average reported

self-belief for men (women) less the share of correct answers provided to the question by men (women). We then difference overestimation of men and women to compute the gender gap.

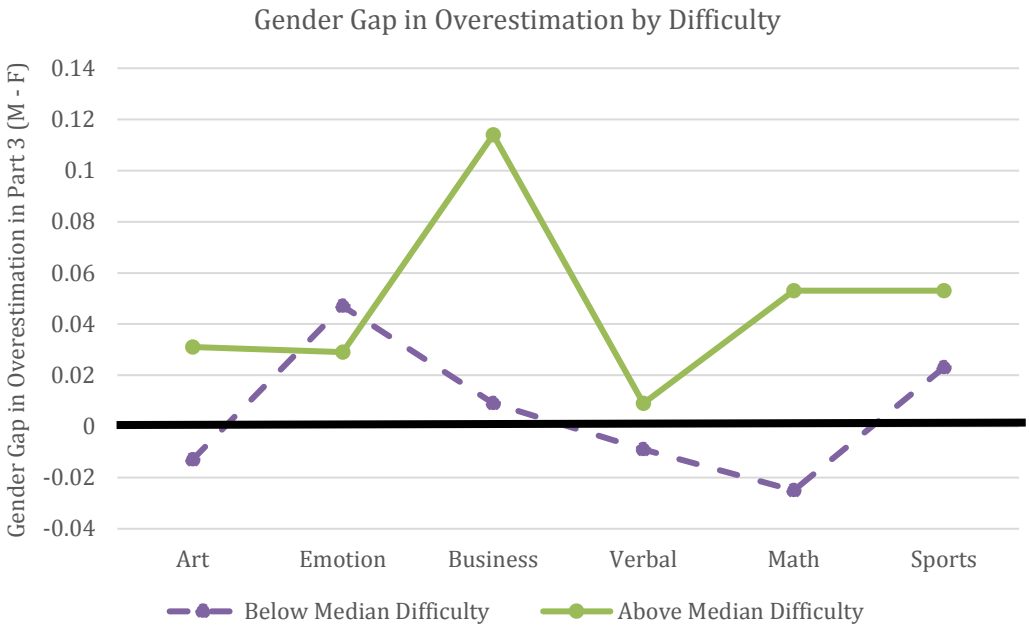


Figure IV

The gender gap in overconfidence is largely a function of question difficulty. In fact, the widely-held view that men are substantially more overconfident than women, particularly in male-typed domains, holds only for the more difficult questions. For easier questions, the gender gap in overconfidence is reduced in most categories, even pointing in favor of women in Art, Verbal, and Math. This suggests strongly that, in our analysis of beliefs, we need to take account of differences in confidence by gender that depend on the difficulty of individual questions.

Finally, we point out that, on average, women are significantly better calibrated in our dataset than men. This is true both when we consider estimates of own performance and that of others. In Part 3 data, men overestimate own performance by 15 percentage points on average (i.e. stated believed probability of answering correctly exceeds the observed one by 15 percentage points on average), while women overestimate own performance by 13 percentage points on average ($p=0.02$, from regression that clusters observations at the participant level). A

similar pattern emerges in self-beliefs of Part 1 scores: men overestimate own performance by 0.29 questions on average while women do so by only 0.06 questions on average (p=0.03).

In interpreting this evidence, it is helpful to look at men's and women's beliefs about others. We find that women are *also* better calibrated in assessing the ability of others. In Part 3 data, women overestimate partner performance by 3 percentage points less than men (11 percentage points versus 14 percentage points, p=0.07). In Part 1, women overestimate partner performance by 0.40 questions less than men (0.59 questions versus 0.98 questions, p<0.01). The finding that women are better calibrated in assessing others, not just themselves, suggests that the gender gap in beliefs is not only a function of gender differences in self-confidence, but a more general pattern of gender differences in miscalibration.[6]

This preliminary evidence tells us that we cannot merely look at the basic pictures to understand the nature of beliefs about gender. Actual ability to answer questions correctly shapes beliefs – beliefs are not entirely divorced from reality. And, both overconfidence related to question difficulty, or DIM, and gender stereotypes appear to play a role in shaping beliefs. To analyze the data, we need to write down a model that describes the various forces that might influence beliefs about gender. In section 4, we present a formal model of beliefs about gender. In Section 5, we estimate this model empirically. In Section 6, we further examine the evidence on stated beliefs and behavior shaped by assessments of relative performance.

4. The Model

In our model, reported beliefs depart from rationality due to: i) difficulty influenced miscalibration or DIM and ii) stereotypes. There are two groups of participants, $G = M, F$ (for male and female) and 6 categories of questions $J \in$

[6] One concern with this result is that the questions we ask are on average fairly hard (on average, less than half of the participants answer a given question correctly. The results might be different for very easy questions, although difficulty is likely a feature of most academic and professional settings.

$\{art, emotions, business, verbal, math, sports\}$. Denote by $p_{i,j}$ the probability that individual $i \in G$ answers the question $j \in J$ correctly. We assume that $p_{i,j}$ is given by:

$$p_{i,j} = p_{G,J} + a_{i,j}, \tag{1}$$

where $p_{G,J}$ is average performance in category J across the individual's gender group G. Component $a_{i,j}$ captures individual-specific ability and question-specific difficulty. At the gender-category level, the definition $\mathbb{E}_{ij}(p_{i,j}) = p_{G,J}$ imposes $\mathbb{E}_{ij}(a_{i,j}) = 0$. Individual $i \in G$ is better than the average member of group G in category J if $\mathbb{E}_j(a_{i,j}) > 0$. Question $j \in J$ is easier than the average in category J if $\mathbb{E}_i(a_{i,j}) > 0$.

Miscalibration of ability and question difficulty. A large literature documents the fact that people hold systematically biased beliefs about their own performance. In an experimental setting similar to ours, although not focused on gender, Moore and Healy (MH 2008) show that participants robustly and significantly overestimate their own performance in trivia questions that are difficult, namely where the true share of correct answers is low.

The psychology of this phenomenon is an open question. People may overestimate their performance due to self-serving beliefs about own ability, a phenomenon often dubbed "overconfidence". Excess optimism for hard questions may also be due to overestimation of low probability events as in Kahneman and Tversky's Prospect Theory (1979).[7] In MH (2008), agents know their average ability in a category, but get a noisy signal of the difficulty of a specific question or task. Because they are Bayesian, agents anchor beliefs to their known average ability and discount the weight of the noisy signal. This effect creates overestimation for hard questions,

[7] In principle, these forces are not observationally equivalent: "overconfidence" only applies to self-beliefs, overestimation of low probability event also applies to beliefs about others.

but also underestimation for easy questions. Here we do not seek to tease out these specific mechanisms and refer to their joint operation as "Difficulty Influenced Miscalibration", or DIM.

Figure III suggests that beliefs can be well approximated by a (gender-specific) affine function of question level difficulty. Thus, to capture DIM we write the perceived probability $p_{i,j}^{DIM}$ of answering correctly as an affine transformation of true ability $p_{i,j}$:

$$p_{i,j}^{DIM} = c + \omega p_{i,j}, \tag{2}$$

where c and ω are such that the entailed beliefs across all questions lie in [0,1]. When $c > 0$ and $\omega \in (0,1)$ participants overestimate ability in hard questions where $p_{i,j}$ is low, and may underestimate it when $p_{i,j}$ is high. Perfect calibration in easy questions occurs if $c = 1 - \omega > 0$.

Stereotypes. We model stereotypes following BCGS (2016). Beliefs about a group G overweight its more representative types, defined as the types that are most likely to occur in G relative to a comparison group $-G$. Under this approach, stereotypes contain a "kernel of truth": they exaggerate true group differences by focusing on the -- often unlikely -- features that distinguish one group from the other. For example, BCGS (2016) show that beliefs about Republicans and Democrats in the US exhibit this kernel of truth by overweighting the extreme elements in each party to the measure of their representativeness relative to the other party.

In our setup, stereotypes distort the perceived ability $p_{G,J}$ of the average group member. In each category J there are two types: "answering correctly" and "answering incorrectly". For group G (resp. $-G$) the probability of these types is $p_{G,J}$ and $1 - p_{G,J}$ (resp. $p_{-G,J}$ and $1 - p_{-G,J}$). Following BCGS, we say that "answering correctly" is more representative for group G in category J than "answering incorrectly" when $\frac{p_{G,J}}{p_{-G,J}} > \frac{1-p_{G,J}}{1-p_{-G,J}}$, namely when $p_{G,J} > p_{-G,J}$. The stereotypical ability of the average member of G in category J is given by:

$$p_{G,J}^{st} = p_{G,J}\left(\frac{p_{G,J}}{p_{-G,J}}\right)^{\theta\sigma}\frac{1}{Z_{J,G}}, \tag{3}$$

where $\theta \geq 0$ is a measure of representativeness-driven distortions and $Z_{J,G}$ is a normalizing factor so that $p_{G,J}^{st}+\left(1-p_{G,J}\right)^{st} = 1$. Parameter σ captures the mental prominence of cross gender comparisons: the higher is σ, the more male-female gender comparisons are top of mind. The case $\theta\sigma = 0$ describes the rational agent. When $\theta\sigma > 0$, representative types are overweighted. This is different from statistical discrimination, which would suggest that individuals $i \in G$ are judged as the average member of group G (overweighting $p_{G,J}$ relative to $a_{i,j}$) but does not entail a distortion of $p_{G,J}$.

When $p_{G,J}$ is close to $p_{-G,J}$, Equation (3) can be linearly approximated as[8]

$$p_{G,J}^{st} = p_{G,J} + \theta\sigma\left(p_{G,J} - p_{-G,J}\right). \tag{4}$$

The stereotypical belief of group G in category J entails an adjustment $\theta\sigma\left(p_{G,J} - p_{-G,J}\right)$ in the direction of the *true* average gap between groups $\left(p_{G,J} - p_{-G,J}\right)$.[9] In subjects where men are on average better than women, $p_{M,J} > p_{F,J}$, the average ability of men is overestimated and that of women is underestimated. Because gaps in average ability vary across categories J, stereotypes are category-specific. The effect of the gender gap in beliefs is stronger the more gender comparisons are top of mind, namely the higher is σ. Although as we try to reduce the

[8] To see this, start from $p_{G,J}^{\theta} = p_{G,J}\left(p_{G,J} + \left(1-p_{G,J}\right)\cdot\left(\frac{1-p_{G,J}}{1-p_{-G,J}}\right)^{\theta}\cdot\left(\frac{p_{G,J}}{p_{-G,J}}\right)^{-\theta}\right)^{-1}$. Write $p_{G,J} = p_{-G,J} + \epsilon$, so that $\left(\frac{1-p_{G,J}}{1-p_{-G,J}}\right)^{\theta} \sim 1 - \frac{\theta}{1-p_{-G,J}}\epsilon$ and $\left(\frac{p_{G,J}}{p_{-G,J}}\right)^{-\theta} \sim 1 - \frac{\theta}{p_{-G,J}}\epsilon$. Then expand $p_{G,J}^{\theta}$ to first order in ϵ to get the result.

[9] This is a departure from Coffman (2014), who measured the stereotype according to self-reported perceptions of the gender-type of each category. While the two measures are highly correlated in our data (correlation of average male advantage in the category and self-reported gender-type perceptions is greater than 0.7), there are some potentially important discrepancies. In particular, while verbal is perceived as female-typed, men have an advantage in our sample on average in Part 1 and business is perceived as male-typed, but women perform better than men in business in Part 3. To the extent that our observed gaps do not coincide with participant expectations about the gaps in the population, our estimates may understate the effect of stereotypes.

prominence of gender comparisons in the experiment, different experimental treatments, in particular the assignment of a male or female partner, are expected to influence σ.

Beliefs and Empirical Strategy Denote by $p_{i,j}^b$ the probability with which person i is *believed* to have correctly answered question j. We assume that person i's belief $p_{i,j}^b$ is distorted by two separate influences: difficulty influenced miscalibration $p_{i,j}^{DIM}$ of true ability and the gender stereotype in category J. Formally, we assume that:

$$p_{i,j}^b = c + \omega\left(p_{G,J} + a_{i,j}\right) + \theta\sigma\left(p_{G,J} - p_{-G,J}\right). \tag{5}$$

This equation nests rational expectation for $c = \theta\sigma = 0$ and $\omega = 1$, in which case beliefs only depend on the objective group and individual-level abilities. If $\theta\sigma = 0$, but $c \neq 0$ or $\omega \neq 1$, then DIM is the only departure from rational expectations. If instead $\theta\sigma > 0$, but $c = 0$ and $\omega = 1$, distortions are driven only by stereotypes.

We use Equation (5) to organize our empirical investigation. It illustrates the key difference between stereotypes and miscalibration in our model and identification strategy. Miscalibration, characterized by the constant c and slope ω, can be identified by comparing beliefs to objective ability across questions within a given category J (first and second term in Equation 5). This effect is orthogonal to gender stereotypes, which are identified by comparing beliefs across categories (controlling for question difficulty).

Linking this evidence with the model raises a key issue. Given the natural variation in performance and gender gaps across samples, which gender-specific performance p_{-G} is used when forming stereotypes? For example, stereotypes about a fellow male university student may be shaped by the comparison with performance of females in the overall population, or by the performance of female university students. This concern is compounded by the fact that the point estimate of gender gaps can change signs across Parts 1 and 3. To address this concern, we show

that the overall performance of our model across all categories is similar in different parts of the experiment. This is the case because our results are driven by the categories (sports and mathematics), in which gender gaps are large and stable across different measurements. We also replicate many of our results in a broader sample of online participants where gender gaps may be more representative of the overall population (see Appendix D).

As Equation (5) makes clear, our empirical strategy remains qualitatively the same when analyzing self-beliefs, beliefs about others, and beliefs about specific gender groups.[10] Of course, the estimated coefficients $c, \omega, \theta\sigma$ can well vary across different types of beliefs. For instance, the strength of DIM as captured by c and ω may vary across men and women or across beliefs about self and others (e.g., self-serving overconfidence should only affect self-beliefs). The stereotypes coefficient $\theta\sigma$ may be higher if gender comparisons become top of mind when the partner is revealed to be of the opposite gender. By estimating Equation (5) separately for men and women, we allow parameters $c, \omega, \theta\sigma$, to vary across genders and belief types.

5. Determinants of Beliefs

As described in Section 2, we record beliefs about own and partner's performance, both at the question (Part 3) and topic (Part 1) levels. We now present our estimating equations, discuss econometric issues, and present the results.

Equation (5) describes beliefs held by $i \in G$ regarding own performance at the question level (part 3):

$$p^b_{i,j} = c + \omega\left(p_{G,J} + a_{i,j}\right) + \theta\sigma\left(p_{G,J} - p_{-G,J}\right)$$

In turn, beliefs about own level performance at the category level (Part 1) are:

$$N\mathbb{E}_{j\in J}\left(p^b_{i,j}\right) = Nc + \omega N\left(p_{G,J} + \mathbb{E}_{j\in J}\left(a_{i,j}\right)\right) + \theta\sigma N\left(p_{G,J} - p_{-G,J}\right). \quad (6)$$

[10] Equation (4) can be equivalently derived by assuming that DIM distortions apply to stereotyped beliefs, in the sense that $p^b_{i,j} = c + \omega\left(p^{st}_{G,J} + a_{i,j}\right)$. In this case, the coefficient in front of the gender gap is $\omega\theta\sigma$ and not θ.

where N is the number of questions in each category. In Equation (6) the DIM parameters c, ω are measured across individuals with different abilities, not across specific questions within an individual as in Part 3.

Beliefs about others are shaped by the same influences as self-beliefs. The belief $p^b_{i'\to i,j}$ held by individual i' about the performance of individual $i \in G$ on a given question (Part 3) is:

$$p^b_{i'\to i,j} = c + \omega\left(p_{G,J} + \mathbb{E}_i\left(a_{i,j}\right)\right) + \theta\sigma\left(p_{G,J} - p_{-G,J}\right). \tag{7}$$

The term $\mathbb{E}_i\left(a_{i,j}\right)$ reflects the fact that i' has no specific information about the ability of i in question j, so beliefs should depend on the average hit rate of group G for the same question.[11] The average believed score reported in Part 1 for a generic member of G in category J satisfies:

$$N\mathbb{E}_{j\in J}\left(p^b_{i'\to i,j}\right) = cN + \omega N p_{G,J} + \omega\theta\sigma N\left(p_{G,J} - p_{-G,J}\right). \tag{8}$$

We estimate (7) and (8) using data from participants who knew the gender of their partner.

A number of econometric issues arise from specifications (5) through (8). Estimation relies on finding proxies for two objects: i) the gender gap $\left(p_{G,J} - p_{-G,J}\right)$ in performance, and ii) individual as well as group level ability. We next discuss how we handle these explanatory variables, starting from the gender gap.

Under the assumption that $\mathbb{E}_{i\in G, j\in J}\left(a_{i,j}\right) = 0$, the gap $\left(p_{G,J} - p_{-G,J}\right)$ is directly observed in the data as the average performance gap between genders in the sample. With sufficiently large N, this measure should be reliable. Table II reports these performance gaps measured as the average number of correct questions (out of 10), separated by gender and topic, for Part 1 and Part 3 questions. In our sample, men outperform women in sports and math in both parts. Gaps in the other categories are mixed. In business and verbal skills, men outperform women by a

[11] According to Moore and Healy, beliefs should be less regressive at the question rather than at the category level because average difficulty of a category is less noisy than difficulty of a specific question. Similarly, beliefs about self should be less regressive than beliefs about others because information about others is noisier.

significant margin in Part 1, but not in Part 2. In the other stereotypically female categories (emotion recognition and art), performance gaps are small and statistically insignificant. [12]

Table II: Summary Statistics on Performance				
	Men	**Women**	**Gap (M-W)**	**p value**
Part 1 (out of 10 qns.)				
Emotion Score	7.74	7.50	0.24	0.13
Art Score	4.56	4.55	0.02	0.92
Verbal Score	6.71	6.09	0.62	0.002
Math Score	5.52	4.77	0.75	0.001
Business Score	4.18	3.39	0.79	0.000
Sports Score	4.26	2.85	1.42	0.000
Part 2 (out of 10 qns.)				
Emotion Score	5.97	5.83	-0.13	0.44
Art Score	4.06	4.32	-0.26	0.06
Verbal Score	6.09	5.80	0.29	0.16
Math Score	5.10	4.65	0.45	0.04
Business Score	4.31	4.38	-0.07	0.71
Sports Score	5.34	3.92	1.42	0.000

Notes: P-value is given for the null hypothesis of no performance difference between genders using a Fisher-Pitman permutation test for two independent samples.

The other component of the model is individual ability, which is also measured with error. The most severe problem arises when dealing with ability in a specific question, as in Equation (5). We do not observe the objective individual- and question-specific ability $p_{i,j}$; we instead observe whether subject i answered question j correctly, denoted by a dummy $I_{i,j}$. Because $I_{i,j}$ is an imperfect measurement of $p_{i,j}$, estimating Equation (5) using $I_{i,j}$ involves well-known econometric issues. First, $I_{i,j}$ is noisier than $p_{i,j}$, which causes attenuation bias on the coefficient ω on own ability. Second, the noise in $I_{i,j}$ can also bias the gender gap coefficient $\theta\sigma$. To address this issue, we adopt a two stage approach. We first estimate $I_{i,j}$ using a set of proxies for individual-level ability: the individual's average ability in Part 3 in that category, excluding

[12] Our math questions are taken from a practice test for the GMAT Exam. In 2012 – 2013, the gender gap in mean GMAT scores in the United States was 549 vs. 504 (out of 800). See: http://www.gmac.com/~/media/Files/gmac/Research/GMAT%20Test%20Taker%20Data/2013-gmat-profile-exec-summary.pdf. Our verbal questions are taken from practice tests for the Verbal Reasoning and Writing sections of the SAT I. The relative performances we observe are broadly in line with other evidence. In SAT exams, taken by a population in many ways similar to our lab sample, men perform better than women in math (527 vs 496 out of 800) and perform equally in verbal questions (critical reading plus writing, 488 vs 492 out of 800), though these differences are not significant.

question *j*, $p_{i,J\setminus j}$, and the average frequency of a correct answer to that particular question, *j*, by all other participants, $p_{(G\cup -G)\setminus i,j}$. These two proxies do not use information about participant *i*'s performance on question *j*, but still capture her overall ability in the category *J* and the overall difficulty of the particular question *j*. We then implement the first stage regression:

$$I_{i,j} = \alpha_0 + \alpha_1\, p_{i,J\setminus j} + \alpha_2 p_{(G\cup -G)\setminus i,j} + \alpha_3\left(p_{G,J} - p_{-G,J}\right),$$

where the gender gap $p_{G,J} - p_{-G,J}$ is included from the second stage estimation. The fitted values $\hat{I}_{i,j}$ of the above regressions are then used as proxies for true individual- and question-specific ability $p_{i,j}$. This "instrumental variable" approach helps us reduce biases due to noisy ability measurement while preserving the interpretation of coefficients as distortions due to stereotypes or DIM.[13]

Individuals' ability at the category level, necessary to estimate Equations (6), (7) and (8), are proxied for with their sample counterparts. Thus $N\left(p_{G,J} + \mathbb{E}_{j\in J}\left(a_{i,j}\right)\right)$ in Equation (6) is proxied by the actual score obtained by individual i in category *J*. Similarly, the ability measures in Equations (7) and (8) are proxied by the share of correct answers by gender *G* in question *j* and in category *J*, respectively.

5.1 Beliefs about own performance

Table III reports the results from specifications (5) and (6) on self-beliefs. Columns I and II use Part 3 question-level data to estimate (5). We capture ability using the fitted values $\hat{I}_{i,j}$ described above; first stage estimates appear in Appendix C. Columns III and IV present the results for category-level performance in Part 1.

[13] In Appendix C, we perform a robustness check of the two-stage approach described above. We separately add the proxies for individual ability, $p_{i,J\setminus j}$ and $p_{(G\cup -G)\setminus i,j}$ to Equation (5). This provides a simpler method to pinning down the effect of stereotypes; however, we lose the interpretation of *c* and ω. Estimated coefficients on the gender gaps are very similar to the two-stage estimates.

Table III: Self-beliefs							
OLS Predicting Own Believed Probability of Answering Correctly in Part 3 Estimation of $p^b_{i,j} = c + \omega(p_{G,J} + a_{i,j}) + \theta\sigma(p_{G,J} - p_{-G,J})$				OLS Predicting Own Believed Part 1 Score Estimation of $N\mathbb{E}_{j\in J}(p^b_{i,j}) = Nc + \omega N\left(p_{G,J} + \mathbb{E}_{j\in J}(a_{i,j})\right) + \theta\sigma N(p_{G,J} - p_{-G,J})$			
	Para-meter	I (Men)	II (Women)		Para-meter	III (Men)	IV (Women)
Own Gender Adv. in Pt. 3	$\theta\sigma$	0.27**** (0.048)	0.21**** (0.049)	Own Gender Adv. in Pt. 1	$\theta\sigma$	0.73**** (0.092)	-0.02 (0.104)
Fitted Value of $\hat{I}_{i,j}$	ω	0.56**** (0.013)	0.63**** (0.014)	Individual's Pt. 1 Score in Category	ω	0.70**** (0.027)	0.81**** (0.029)
Constant	C	0.36**** (0.011)	0.31**** (0.012)	Constant	Nc	1.36**** (0.191)	0.90**** (0.196)
R-squared		0.21	0.22	R-squared		0.37	0.42
Clusters		344	296	Clusters		344	296
N		11,198	9,360	N		1,376	1,184

Notes: Pools observations for Ohio State and Harvard experiments. Standard errors clustered at the individual level.

There are a number of key results. First, DIM is an important determinant of self-beliefs held by both men and women. We estimate $\omega < 1$ (p<0.001) and $c > 0$ in all specifications, strongly rejecting the null of rational expectations ($c = \theta = 0, \omega = 1$). Together, the estimates for the constant and the slope imply that participants overestimate their own performance for difficult questions, where $\mathbb{E}_i p_{i,j}$ is low (around 20%, correct by chance) and underestimate their performance slightly for easy questions (where $\mathbb{E}_i p_{i,j}$ closer to 1), as $c + \omega < 1$ (in Columns III and IV the intercept must be divided by 10). DIM distortions are smaller in Part 1 than in Part 3 data (namely c is lower and ω is higher in Columns III and IV than in Columns I and II).[14]

The estimates also reveal gender differences in DIM. In both Part 3 and Part 1, we estimate lower constants and higher slopes for women than for men, suggesting that DIM is stronger for men. This echoes the findings we reported in Section 3 that, on average, men overestimate own performance more than women do, particularly for more difficult questions. This is consistent

[14] This is consistent with both the Moore and Healy mechanism (subjects perceive a more precise signal of average difficulty after observing 10 questions in a subject than after observing a single question) and with overestimation of small probabilities (which exerts a smaller distortion on the average score from several questions).

with previous studies that document that women are less overconfident and better calibrated than men (Deaux and Farris 1997, Lundeberg, Fox, and LeCount 1994).

Crucially, we find that stereotypes are a significant predictor of self-beliefs held by men in both Parts 1 and 3. The effect is large. Specification I shows that a 5 percentage point increase in male advantage in a question (roughly the size of the male advantage in math) increases beliefs about own probability of answering correctly by 0.27*5=1.35 percentage points relative to rational expectations in Part 3. Turning to Specification III on Part 1 beliefs, a 0.75 question increase in male advantage in a category, again roughly equivalent to the male advantage in math, increases beliefs about performance in that category by 0.55 questions. If we scale Part 1 results to the effects for a single question to benchmark against Part 3, we estimate that the same 5 percentage point increase in male advantage increases men's beliefs of answering a given question correctly by an estimated 3.7 percentage points.[15]

For women, stereotypes are only a significant predictor of self-beliefs in Part 3. In this case, the same 5 percentage point increase in male advantage in a category leads to approximately a 1.05 percentage point decrease in believed probability of answering correctly. However, we estimate no impact of stereotypes on women's self-beliefs in Part 1 data.

5.2 Beliefs about others' performance

Table IV reports the results from regression specifications (7) and (8) on beliefs about others' performance on individual questions (Part 3, Columns I and II) and at the category-level (Part 1, Columns III and IV). We use data from participants who knew the gender of their partner, and we

[15] To compare order of magnitudes across Parts 1 and 3, it is important to keep in mind that Part 1 measures beliefs of total score on a 10-point scale and Part 3 measures beliefs of probability of answering correctly on a 1-point scale. Thus, when making direct comparisons, we scale the Part 1 results to translate to an impact on a single question. Part 1 and Part 3 also vary in other potentially important dimensions. In particular, beliefs are elicited in two different ways across these parts. In Part 1, participants are incentivized to guess their score by receiving a fixed size bonus payment for guessing correctly and nothing otherwise. In Part 3, participants are incentivized to guess their probability of answering correctly using a procedure similar to a BDM.

pool all evaluators, without keeping track of their gender. In Appendix C we show effects separately by gender of the evaluator.

Table IV: Beliefs about Others

OLS Predicting Belief of Partner's Probability of Answering Correctly in Part 3 Estimation of $p^b_{i'\to i,j} = c + \omega\left(p_{G,J} + \mathbb{E}_i(a_{i,j})\right) + \theta\sigma(p_{G,J} - p_{-G,J})$				OLS Predicting Belief of Partner's Part 1 Score Estimation of $N\mathbb{E}_{j\in J}(p^b_{i'\to i,j}) = cN + \omega N p_{G,J} + \theta\sigma N(p_{G,J} - p_{-G,J})$			
	Para-meter	I (Beliefs About Men)	II (Beliefs About Women)		Para-meter	III (Beliefs About Men)	IV (Beliefs About Women)
Partner's Gender Adv. in Category in Pt. 3	$\theta\sigma$	0.12** (0.055)	0.21**** (0.063)	Partner's Gender Adv. in Category in Pt. 1	$\theta\sigma$	1.35**** (0.12)	-0.40**** (0.12)
Share of Partner's Gender Answering Question Correctly	ω	0.35**** (0.016)	0.37**** (0.016)	Partner's Gender Average Score in Category Pt. 1	ω	0.78**** (0.057)	0.74**** (0.054)
Constant	C	0.44**** (0.014)	0.46**** (0.013)	Constant	Nc	0.89*** (0.34)	2.09**** (0.34)
R-squared		0.11	0.11	R-squared		0.20	0.19
Clusters		196	185	Clusters		196	185
N		6,080	5,399	N		784	740

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

Table IV reveals some similarities to but also some differences from the self-beliefs estimates of Table III. First, DIM also plays a role in beliefs about others, overestimating ability on hard questions and slightly underestimating it on easy ones. These belief distortions are more severe here than in the case of self-beliefs (particularly on hard questions). This finding could be explained by the Moore Healy mechanism, because signals of difficulty for others are presumably noisier than those for self. This finding also suggests that the strong overestimation of own performance we observe in our data is driven not only by conventional self-serving biases or "overconfidence", but also by more general miscalibration in estimating task difficulty and ability.

Second, stereotypes play a consistent and significant role in shaping stated beliefs about men, in all sets of elicited beliefs. Once again, the quantitative impact of stereotypes is important, particularly in Part 1. The evidence on the role of stereotypes for beliefs about women is mixed,

just as it was for self-beliefs. In Part 3 data (Column II) stereotypes shape beliefs about women as predicted by the model (and as they did for self-beliefs). In Part 1 data, however, the effect of gender gaps goes in the opposite direction: beliefs about women are more optimistic in categories where women do worse than men (see Column IV).

5.3 Assessing the Relative Importance of Stereotypes

In this section we assess model performance by comparing model-predicted beliefs to observed beliefs in our data. What is the role of stereotypes versus DIM in explaining observed beliefs? To answer, we compute prediction errors in the full model estimated in Sections 5.1 and 5.2, as well as the prediction errors obtained in a DIM-only version of the model in which we force the stereotypes parameter to zero. We perform our analysis at the category level, distinguishing beliefs about self from those about others, and beliefs about men from those about women.[16] For each comparison, we use the estimates from Tables III and IV, averaging over the Part 1 and Part 3 parameters:

$(c, \omega, \theta\sigma)$	Part 3	Part 1
Self	M: (0.36,0.56,0.27)	M: (1.36,0.70,0.73)
	F: (0.31,0.63,0.21)	F: (0.90,0.81,-0.02)
Other	M: (0.44,0.35,0.12)	M: (0.89,0.78,1.35)
	F: (0.46,0.37,0.21)	F: (2.09,0.74,-0.40)

Figure V shows the results for men. The "DIM-only" model takes the above estimates of c and ω but forces $\theta\sigma = 0$. The "Full Model" includes the estimated value of $\theta\sigma$. The key idea here is to

[16] Formally, we compute the directional prediction error in both the full model and the DIM-only model:

$$\epsilon_J^m = \frac{1}{|k|}\sum_{k,G}\frac{\mathbb{E}_{ij}p_{i,j,k}^{b,m} - \mathbb{E}_{ij}p_{i,j,k}^{b}}{\mathbb{E}_{ij}p_{i,j,k}}, \quad m = Full, DIM$$

where operator $\mathbb{E}_{ij}$ means that averages are taken over all individuals $i \in G$. The index k in the sum operator denotes the type of belief we are considering, k = self/other X Part 1/Part 3 X men/women.

ask how much incorporating stereotypes into the model improves predictions. A model is more successful when the vertical bar associated with it is closer to zero.

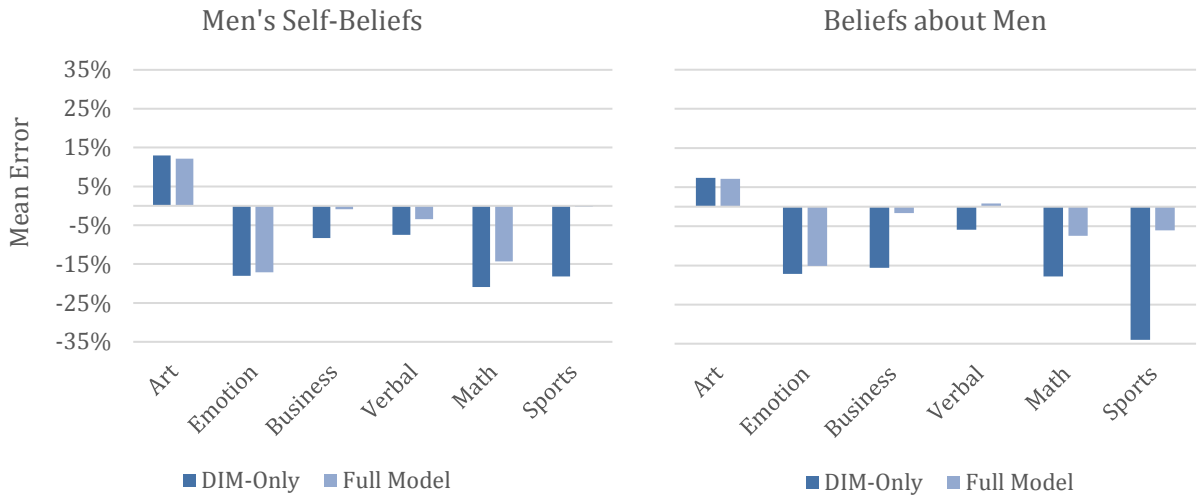


Figure V

Stereotypes are important predictors of both men's self-beliefs and beliefs about men. Beliefs tend to be fairly severely under-predicted in the DIM-only model for men (by 12% on average). This under-prediction is consistently reduced when stereotypes are added to the model, bringing the mean observed error down to 4%. The improvement from stereotypes occurs for both self-beliefs and beliefs about men, but the effect is larger for the latter. Beliefs about men are under-predicted by 14% in the DIM-only model but only by 4% when stereotypes are also accounted for. Stereotypes are an especially important determinant of beliefs in stereotypically male-typed categories, namely business, math, and sports.

Figure VI presents the results for women. Here adding stereotypes to the model does very little to reduce mean observed errors. The DIM-only model under-predicts observed beliefs for emotion, verbal, and math and over-predicts observed beliefs in art, business, and sports. Because the direction of the errors varies across categories, average errors in the DIM-only model are small, averaging only 3.2% of observed ability for self-beliefs and 0.2% of observed ability for

beliefs about women. Adding stereotypes to the model has no large impact on predicted beliefs in any category for women. For self-beliefs, average errors decrease slightly from 3.2% to 2.6% when stereotypes are added to the model; for beliefs about women, errors increase from 0.2% to 2.4% when stereotypes are added to the model. The only exception are beliefs about female ability in verbal and math, which are better explained by taking stereotypes into account.

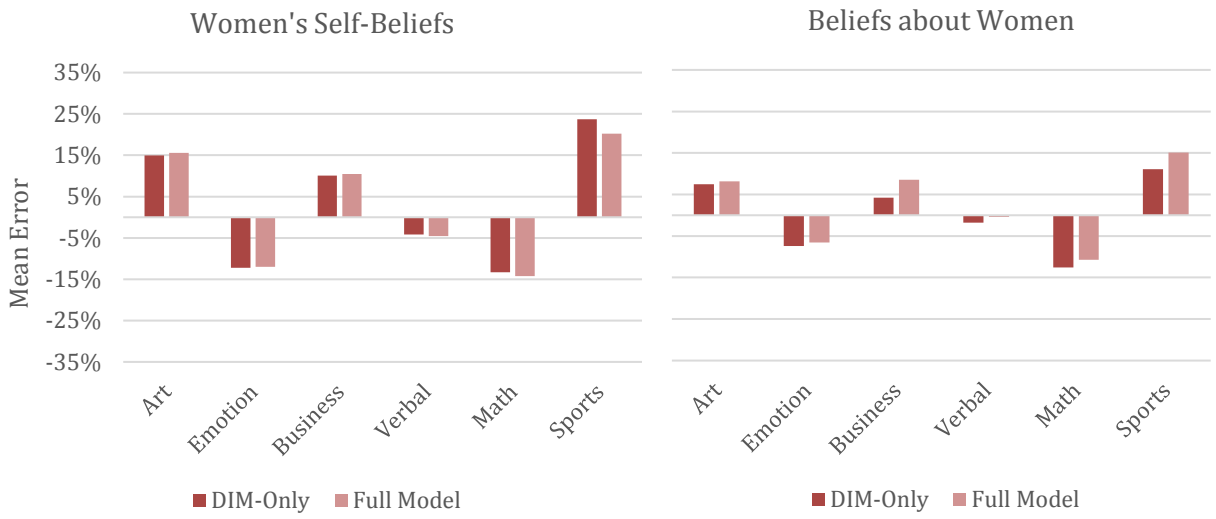


Figure VI

Our data suggest that stereotypes are not a consistent determinant of women's beliefs about themselves or of others' beliefs about women. This finding raises a concern that stated beliefs are contaminated by social desirability bias, a desire to appear as though one does not hold negative views of women. While our experiment is designed to minimize this concern (participants report beliefs about a single participant rather than a gender difference; beliefs are incentivized and collected anonymously; gender is not emphasized), we cannot fully rule out such contamination.

We explore this issue further in two ways. First, in Section 6, we link stated beliefs to strategic decisions about willingness to contribute, asking about the predictive power of these beliefs. If stated beliefs were not truthful, we might expect beliefs to have less predictive power for decision-making, and we find some suggestive evidence of this. Second, in a follow-up experiment, we collected data on social norms surrounding the appropriateness of expressing

beliefs of gender differences in different domains, and correlate this data with our findings on beliefs about women. These results are reported in Appendix B.

5.4 Self-Beliefs and Context

In this section, we extend our analysis by testing whether participants' beliefs about their own *absolute* ability are influenced by the gender of their partner. In particular, we ask whether, compared to participants who are paired with a male partner, participants paired with a female partner state more optimistic beliefs in male-typed domains (and similarly, less optimistic beliefs in female-typed domains). This type of context-dependence is a central prediction of BCGS (2016). While BCGS (2016) present evidence of context dependence in an abstract laboratory experiment (participants make guesses about the color of t-shirts worn by cartoon characters) and in field data on political beliefs, here we conduct a more demanding test of context dependence with our data. We collect data on absolute beliefs of own ability in a given question or domain, where a participant likely has much stronger priors compared to our previous experiments. If the same individual believes she is more likely to answer a particular math or sports question correctly simply because she is aware her partner is female rather than male, this is arguably quite strong evidence of the power of context dependence.

Such context dependence is a distinctive prediction of our model of stereotype distortions, relative to the DIM channel (or rational beliefs).[17] However, this test raises two issues. On the one hand, gender may already be top of mind even if the partner's gender is not known (i.e., it might be that $\sigma = 1$ already in our estimates of Equations 5 and 6). Second, and more important, bringing cross gender comparisons to the top of mind may strengthen the social desirability bias. For instance, a male participant learning that his partner is female may become reluctant to

[17] A similar notion has been explored by work on the "Stereotype threat" hypothesis (Steele and Aronson 1995, Spencer, Steele, and Quinn 1999). However, note that in our experiment, beliefs are elicited after the performance, whereas the stereotype threat operates by reducing actual performance.

report high self-confidence in stereotypically male subjects such as math, business, or sports. We keep those caveats in mind as we present our tests.

In Table V, we repeat the specifications of Table III in Section 5.1 but restrict the sample to individuals who know partner's gender. We include a dummy for a known female partner and we interact it with own gender advantage. If knowledge of the partner's gender causes gender comparison to become more top of mind (i.e., if σ increases), beliefs about self should go further in the direction of gender gaps. Men paired with women should be more optimistic about self in categories with male gender advantage (positive sign on partner female x own gender advantage in Part 3 for men), and women paired with women should be less optimistic in categories with female advantage (negative sign on partner female x own gender advantage).

Table V: Self-Beliefs with Context Dependence

OLS Predicting Own Believed Probability of Answering Correctly in Pt. 3			OLS Predicting Own Believed Pt. 1 Score		
	I (Men)	II (Women)		III (Men)	IV (Women)
Own Gender Adv. in Pt. 3	0.12 (0.085)	0.29**** (0.093)	Own Gender Adv. in Pt. 1	0.50*** (0.161)	0.18 (0.187)
Fitted Value of $\hat{I}_{i,j}$	0.54**** (0.018)	0.60**** (0.019)	Individual's Pt. 1 Score in Category	0.67**** (0.033)	0.79**** (0.038)
Partner Female	-0.02 (0.019)	-0.003 (0.017)	Partner Female	-0.05 (0.237)	-0.23 (0.273)
Partner Female x Own Gender Adv. In Part 3	0.17 (0.134)	-0.21 (0.131)	Partner Female x Own Gender Adv. In Part 1	0.23 (0.252)	-0.27 (0.229)
Constant	0.39**** (0.020)	0.33**** (0.016)	Constant	1.58**** (0.263)	1.19**** (0.285)
R-squared	0.19	0.21	R-squared	0.33	0.42
Clusters	197	184	Clusters	197	184
N	6,119	5,360	N	788	736

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

The evidence is directionally consistent with the predicted signs, but the effects are not statistically significant.[18] For both men and women in Part 3, the point estimates suggest that a given increase in own gender advantage produces more than twice as large of an increase in

[18] Reading across columns I – IV, the p-values on the interaction of interest are 0.21, 0.11, 0.36, and 0.23

beliefs about self when paired with an opposite gender partner than a same gender partner. For Part 1, we estimate that men react approximately 50% more strongly to own gender advantage when paired with a woman than paired with a man. For women in Part 1, the effect of own gender advantage is estimated to be directionally positive when paired with a male partner, but directionally negative when paired with a female partner, consistent with our hypotheses.We can also pool the data across our male and female participants to increase statistical power (recall that the above analysis restricts attention only to participants who knew the gender of their partner, reducing our statistical power relative to Table III). In Appendix C4, we present a specification that includes all participants who knew the gender of their partner, regressing self-beliefs on male advantage in the category, a dummy for partner being female, and the interaction of these two terms. Context dependence predicts a positive sign on this interaction for all participants, with self-confidence rising in male-typed domains more when paired with a female rather than male partner. The estimated interaction is unsurprisingly of similar magnitude to the estimates provided in Table V, but the increased power from the pooled sample allows for the identification of a marginally significant effect in Part 3 data (estimated interaction of 0.19, SE of 0.093, p-value of 0.05), and a marginally insignificant effect in Part 1 data (estimated interaction of 0.25, SE of 0.17, p-value of 0.14). Full regression results for this specification are reported in Table A7 in Appendix C4. Thus, while the effect sizes are modest and not tightly-estimated, there does seem to be some suggestive evidence of context dependence in our data. Appendix C4, Table A8 also shows that this effect is not limited to gender: among the sample of participants who received photographs of their partner, partner ethnicity has a large and significant impact on self-beliefs.

6. Beliefs of Relative Ability and the Consequences for Decision-Making

Our model and our data deal with beliefs of absolute ability: one's own and partner's believed probability of answering correctly. However, in many decision-making contexts, it is beliefs of relative ability that may be most predictive. In deciding whether to compete in a tournament, decisions are a function of whether an individual believes she can outperform her partner; in deciding whether or not to apply for a job or promotion, decisions are likely a function of believed rank within the pool of potential candidates. It is then important to check how the patterns in believed absolute ability we identify translate into beliefs of relative ability. In this section, we first document how the determinants of beliefs explored in Section 5 – DIM, stereotypes, and context dependence -- combine to produce gender differences in beliefs of *relative* ability. We then take this analysis from beliefs to strategic decisions within a group, and examine how our participants make decisions about when to contribute ideas.

In Figure VII, we present data on beliefs of relative ability, focusing on both partner gender and category. For each participant who knew the gender of their partner, we construct the gap in average beliefs about own ability and average beliefs about partner's ability at the category level, using the Part 3 data. We ask how this believed ability gap between self and partner varies with partner gender and category. Panel (a) presents believed relative ability for men with male partners in blue and female partners in red. Panel (b) presents the same measures for women.

Figure VII clearly shows that the patterns of beliefs documented in Section 5 have important implications for beliefs about relative ability. For men paired with male partners, believed relative ability is quite constant across the categories. In every category, men believe they are roughly 4 percentage points more likely to answer the question correctly than their male partner is. For men paired with female partners, relative beliefs vary sharply with category. In emotion and art, men believe they are less likely to answer correctly than their female partner; in business,

verbal, and math, the believed ability gap is small but positive; in sports, men believe they are more than 10 percentage points more likely to answer correctly than their female partner.

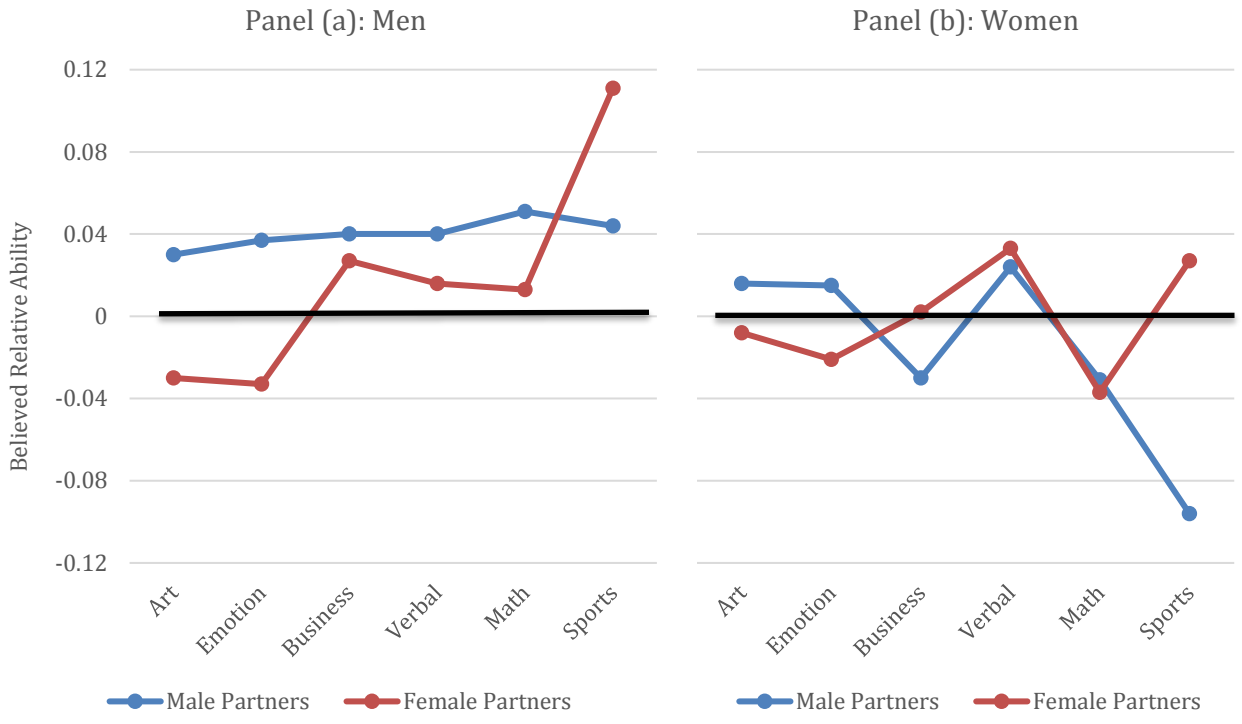


Figure VII

Compared to the men, women generally have lower beliefs of relative ability – they do not believe their own performance exceeds their partner's performance by as much on average. Their beliefs vary sharply with gender of partner and category. When paired with male partners, women believe they outperform their partner in the female-typed categories - emotion, art, and verbal - but believe they are outperformed by their partner in the male-typed categories - math, business, and sports. But, when women are paired with other women, relative beliefs vary less consistently with the gender-type of category. Unlike men, who are always more optimistic about own ability relative to a same-gendered partner, women's beliefs are mixed. They believe their female partner outperforms them in some female-typed categories, like emotion and art, but also in math. Women believe they outperform their female partner on average in verbal and sports.

These patterns suggest that decisions are likely to be a function of gender stereotypes, reflected in responsiveness to both the domain, the gender of one's partner, and the interaction of the two. We can test this using our data from the place in line game, asking 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, their beliefs about their partner, and on the observed gender gap. We predict that place in line should decrease in belief of own ability and increase in beliefs about partner's ability. We restrict attention to participants who knew the gender of their partner at Ohio State, where we have a question-level self and partner belief for each question in Part 3.[19] We first regress place in line on a set of ability proxies: own performance – proxied by the fitted value of $\hat{I}_{i,j}$ 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 IV (women) in Table VI. We then add reported beliefs in Columns II and V, where we add self-beliefs, and in Columns III and VI, where we also add beliefs about the partner.

The first specification (columns I and IV) shows that ability proxies are highly predictive of place in line in the expected direction. Both men and women move forward by 2 places in line when they answer correctly. When a man is paired with a woman, the man moves forward as male advantage increases while the woman moves back in line. 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

[19] Recall that at Harvard 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 that sub-sample. In Appendix C, we explore the robustness of the results in this section to specifications that allow for inclusion of the Harvard participants and for other approaches to controlling for ability.

forward in line by approximately 0.25 positions. Controlling for beliefs of own ability reduces the effect of the gender gap but it remains predictive.[20] Finally, adding beliefs about partner ability (specifications III and VI) shows these beliefs have a modest impact on willingness to contribute in the expected direction. But, again, male advantage remains predictive.

Table VI: Place in Line

OLS Predicting Place in Line (Lower Numbers Indicate Greater Willingness to Contribute)						
	Men			Women		
	I – No Beliefs	II – With Self Beliefs	III – Both Beliefs	IV – No Beliefs	V – With Self Beliefs	VI – Both Beliefs
Male Advantage in Part 3	-0.18 (0.38)	0.11 (0.22)	-0.01 (0.23)	1.48**** (0.45)	0.85*** (0.26)	0.72** (0.28)
Partner Female	0.08 (0.10)	0.03 (0.10)	0.01 (0.09)	-0.02 (0.11)	0.01 (0.09)	0.01 (0.09)
Partner Female x Male Advantage in Part 3	-1.86*** (0.67)	-1.10*** (0.40)	-0.75* (0.38)	-1.38** (0.60)	-1.04** (0.41)	-0.87** (0.42)
Own Performance (Fitted Value of $\hat{I}_{i,j}$)	-1.97**** (0.10)	-0.55**** (0.09)	-0.60**** (0.10)	-2.23**** (0.10)	-0.53**** (0.08)	-0.54**** (0.08)
Believed Probability of Self Answering Correctly		-2.29**** (0.21)	-2.51**** (0.19)		-2.51**** (0.12)	-2.58**** (0.16)
Believed Probability of Partner Answering Correctly			0.43 (0.28)			0.17 (0.16)
Constant	3.12**** (0.10)	3.90**** (0.17)	3.80**** (0.20)	3.37**** (0.09)	4.06**** (0.11)	4.02**** (0.11)
R-squared	0.19	0.52	0.52	0.20	0.57	0.57
Clusters	109	109	109	84	84	84
N	4,359	4,359	4,358	3,359	3,359	3,359

Notes: Includes laboratory data from OSU, including only those individuals who knew the gender of their partner. Standard errors are clustered at the individual level.

This evidence implies that reported beliefs are informative about choice of place in line, but fail to exhaust the effect of the gender gap. This result may reflect noisiness of reported beliefs measures, or failure to capture other factors that are predictive of place in line that correlate with the gender gap. But it is also consistent with reported beliefs being tainted by social desirability bias, whereby beliefs underreport the role of stereotypes. While one cannot draw a definitive

[20] Conditional on own believed ability, men still move forward in line as male advantage increases when paired with a female partner, though the effect is approximately 50% smaller than in Specification I. Women also continue to show a response to gender of partner and male advantage conditional on beliefs of own ability. Beliefs of own ability explain approximately 40% of the estimated effect of male advantage when paired with a male partner.

conclusion from this evidence, it highlights the importance of measuring social desirability bias in reported beliefs. We offer some preliminary work on this in Appendix B.

We have shown that beliefs about self and others are predictive of contribution decisions. The marginal effects presented in Table VI, however, fail to give a sense of the impact of these decisions: what are the consequences of these place in line decisions for the quantity and quality of answers submitted by male and female group members? We close this section by more explicitly tackling this issue. We will say that a participant "contributed" her answer if she submitted a place in line at least as close to the front as her partner. Women contribute 58% of their answers when paired with male partners and 68% of their answers when paired with female partners ($p<0.01$).[21] These differences are largely driven by the more male-typed categories – business, math, and sports – across which women contribute 67% of their answers when paired with women but only 50% of their answers when paired with men ($p<0.01$). We see a smaller but directionally similar discrepancy for men: men contribute 74% of their answers when paired with female partners but 69% of their answers when paired with male partners ($p<0.10$). Again, most of the difference stems from business, math, and sports, where the difference is 78% with female partners versus 68% with male partners ($p<0.01$).

These contribution decisions also have a modest impact on group performance. We measure group performance as the fraction of questions for which a group submits the correct answer. Our design allows us to ask how performance varies across groups that do not know each other's gender and groups in which both partners know each other's gender. While overall group performance is quite similar across these treatments (groups submit the correct answer to a question approximately 55% of the time in each), differences emerge in the cases in which we might expect gender stereotypes to be most relevant. Among mixed gender groups, group

[21] In Appendix C5, we present regressions that further explore these contribution results.

performance is 3 percentage points better when gender is unknown ($p<0.05$). This performance gap comes primarily from the male-typed categories, for which unknown gender groups perform 7 percentage points better than known gender groups ($p<0.01$).

7. Conclusion

Despite substantial evidence that, in some domains, men are more over-confident than women about their ability, sources of such overconfidence are not completely understood. In this paper, we presented evidence that overconfidence comes from (at least) two separate sources, and as such cannot be treated as one homogeneous phenomenon that differs by gender.

Our experimental design enables us to distinguish two factors shaping male and female beliefs. The first is overestimation of the ability of both self and others, which rises with the difficulty of the question, what we called difficulty-influenced miscalibration or DIM. The second is stereotyping. Because we collect data not only on beliefs about self but also on beliefs about others, and we do so for a variety of difficulties and domains, we can disentangle these hypotheses and shed new light on the drivers of overconfidence.

We found that both DIM and stereotypes shape reported beliefs, but their relative importance varies significantly between men and women. A significant part of men's greater self-confidence is due to more severe miscalibration, but miscalibration does not seem to be driven by self-serving biases. Rather, men are particularly over-confident, as compared to women, on the more difficult questions, and not differentially over-confident on easier ones. Importantly, this holds across subjects.

In contrast, stereotypes play a larger role in explaining differences in confidence in the male topics such as math and sports. In our data, this occurs primarily for reported beliefs about men. We also found that both stereotypes and overestimation are reflected in beliefs about relative and not just absolute ability, and actually influence behavior. The two factors combine to

encourage more self-confident behavior of men, and less self-confident behavior of women, particularly in male-typed fields.

Disentangling the causes of the gender gap in self-confidence may prove helpful in interpreting the existing evidence on this gap, but also understanding how this gap could be narrowed. To the extent that stereotypes shape this gap, the reality that actual performance differences between genders are narrowing, especially at the upper tail, suggests that stereotypes might become less prevalent. Moreover, factors that make gender less top of mind would diminish the effects of stereotypes on beliefs, especially in areas where actual differences remain. Although we do not understand the causes of miscalibration as well, the Moore-Healy model suggests that objective feedback about ability will diminish the influence of DIM on self-confidence. But, if DIM is driven by factors other than information, such feedback might not help.

Perhaps the central message of our research is that both stereotypes and miscalibration shape confidence. It is not one versus the other. Women may be less confident both because they are better calibrated, and because in some areas they are stereotyped as weaker by both themselves and others. Our findings can help clarify earlier research, which often treated over-confidence as a homogeneous phenomenon, but also open up avenues for a more nuanced understanding of beliefs about gender.

References

Barber, Brad M. & Terrance Odean. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly journal of Economics*, 116(1), 261-292.

Beyer, Sylvia. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology*, 59(5), 960-970.

Beyer, Sylvia. (1998). Gender differences in self-perception and negative recall biases. *Sex Roles*, 38(1-2), 103-133.

Beyer, Sylvia, & Edward M. Bowden. (1997). Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, 23(2), 157-172

Bohnet, Iris, & Bruno S. Frey. (1999). Social distance and other-regarding behavior in dictator games: Comment. *The American Economic Review*, 89(1), 335-339.

Bordalo, Pedro, Katherine B. Coffman, Nicola Gennaioli, & Andrei Shleifer. (2016). Stereotypes. Forthcoming in *The Quarterly Journal of Economics.*

Buser, Thomas, Muriel Niederle, & Hessel Oosterbeek. (2014). Gender, Competitiveness, and Career Choices. *The Quarterly Journal of Economics,* 129(3), 1409 – 1447.

Coffman, Katherine B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4), 1625-1660.

The College Board. (2013). *2013 College-Bound Seniors: Total Group Profile Report*. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/2013/TotalGroup-2013.pdf

Croson, Rachel & Uri Gneezy. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448-474.

Deaux, Kay & Elizabeth Farris. (1977). Attributing causes for one's own performance: The effects of sex, norms, and outcome. *Journal of research in Personality*, *11*(1), 59-72.

Dreber, Anna, Emma von Essen, & Eva Ranehill. (2011). Outrunning the gender gap—boys and girls compete equally. *Experimental Economics*, 14(4), 567-582.

Gneezy, Uri, Muriel Niederle, & Aldo Rustichini. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3), 1049-1074.

Grosse, Niels D., & Gerhard Riener. (2010). Explaining gender differences in competitiveness: Gender-Task Stereotypes. *Jena Economic Research Papers 2010 – 017*.

Hall, Judith A., & David Matsumoto. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion*, 4(2), 201-206.

Lundeberg, Mary A., Paul W. Fox, & Judith LeCount. (1994). Highly confident but wrong: gender differences and similarities in confidence judgments. *Journal of educational psychology*, *86*(1), 114.

Kahneman, Daniel, & Amos Tversky. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291.

Kiefer, Amy K. & Denise Sekaquaptewa. (2007). Implicit stereotypes, gender identification, and math-related outcomes a prospective study of female college students. *Psychological Science*, *18*(1), 13-18.

Mobius Markus M., Muriel Niederle, Paul Niehaus, & Tanya S. Rosenblat. Managing self-confidence. (2014). *Working Paper*.

Moore, Don A., & Paul J. Healy. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502-517.

Moran, Gwen. (2015, November 9). Women now make up 40% of students at top MBA programs. *Fortune*. Retrieved from http://fortune.com/2015/11/09/women-mba-40-percent/.

Niederle, Muriel, & Lise Vesterlund. (2007). Do women shy away from competition? Do men compete too much?. *The Quarterly Journal of Economics*, 122(3), 1067-1101.

Nosek, Brian A., Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan et al. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26), 10593-10597.

Pulford, Briony D., & Andrew M. Colman. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences*, 23(1), 125-133.

Shurchkov, Olga. (2012). Under pressure: gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, 10(5), 1189-1213.

Spencer, Steven J., Claude M. Steele, & Diane M. Quinn. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, *35*(1), 4-28.

Tversky, Amos, & Daniel Kahneman. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232.

Appendix A: Experimental Instructions and Materials (available on request)

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.

Online Experiment Design

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: 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			
	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). We had 1,019 participants in the first batch and 812 participants in the second batch (we include all participants who take the survey, even if they do not later submit it for payment, allowing us to exceed the posted HIT limit).

Figure A1 graphs the raw data collected from Amazon Mechanical Turk. 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.

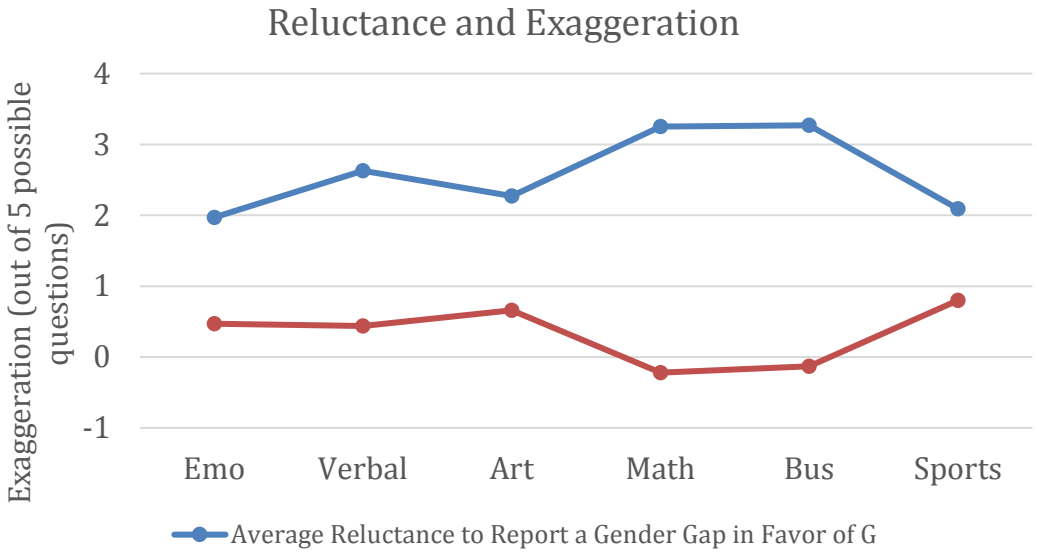


Figure A1. Exaggeration versus Reluctance to Report a Gender Difference in the Category.
Notes: Emotion, Verbal, and Art have true gaps in favor of women and we report average reluctance to report female advantages in these categories; Math, Business, and Sports have true gaps in favor of men and we 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*)	1.034**** (0.016)	0.956**** (0.018)
Share of Correct Answers in Category *J* in Part 3 by Individual *i* (Excluding question *j*)	0.343**** (0.025)	0.366**** (0.024)
Own Gender Advantage in Category	0.291**** (0.044)	0.344**** (0.045)
Constant	-0.183**** (0.012)	-0.157**** (0.012)
R-squared	0.25	0.22
Clusters	344	296
N	13,760	11,840

Notes: Pools OSU and Harvard 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 Part 3. We predict own believed probability of answering correctly from our measures of Part 3 ability: a dummy for whether the individual answered the specific question correctly, share of correct answers in category provided by individual in Part 3 on 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

OLS Predicting Own Believed Probability of Answering Correctly in Pt. 3				
	I (Men)	II (Women)	III (Men)	IV (Women)
Own Gender Adv. in Pt. 3	0.27**** (0.048)	0.21**** (0.049)	0.28**** (0.047)	0.27**** (0.051)
Fitted Value of $\hat{I}_{i,j}$	0.56**** (0.013)	0.63**** (0.014)		
Dummy for Individual Answered Qn. Correctly, $I_{i,j}$			0.20**** (0.007)	0.17**** (0.006)
Individual's Share of Correct Answers in Category in Pt. 3 excluding question *j*			0.32**** (0.025)	0.37**** (0.021)
Overall Share of Correct Answers to question *j*			0.32**** (0.014)	0.37**** (0.014)
Constant	0.36**** (0.011)	0.31**** (0.012)	0.23**** (0.017)	0.17**** (0.016)
R-squared	0.21	0.22	0.31	0.30
Clusters	344	296	344	296

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

C3. Gender of Evaluator

Appendix Tables A4 and A5 present the results on beliefs about others separated by the gender of the evaluator. For the Part 3 data, when we allow the effects of stereotypes to depend on the gender of evaluators, we find that stereotypes play a larger role when evaluating partners of the *same* gender: relative to men, women rely less on stereotypes when evaluating male partners (estimated coefficient of 0.04, p=0.59), but rely more when evaluating female partners (estimated coefficient of 0.40, p<0.001). Men, on the other hand, rely on stereotypes when evaluating men, but not when evaluating women. In Part 1, there are no significant differences in the magnitude of the stereotypes effect by gender of the evaluator.

Table A4

OLS Predicting Believed Probability of Partner Answering Correctly in Part 3				
	Partner is Male		Partner is Female	
	I	II	IIII	IV
Share of Partner's Gender Answering Question Correctly	0.35**** (0.016)	0.38**** (0.024)	0.37**** (0.016)	0.39**** (0.023)
Partner's Gender Adv. in Category in Part 3	0.12** (0.055)	0.18** (0.075)	0.21**** (0.063)	0.039 (0.093)
Female Evaluator		0.026 (0.027)		0.013 (0.027)
Female Evaluator x Share of Partner's Gender Answering Question Correctly		-0.060* (0.032)		-0.048 (0.031)
Female Evaluator x Partner's Gender Advantage in Category in Part 3		-0.13 (0.110)		0.36*** (0.120)
Constant	0.44**** (0.014)	0.43**** (0.020)	0.46**** (0.013)	0.45**** (0.018)
R-squared	0.11	0.11	0.11	0.12
Clusters	196	196	185	185
N	6,080	6,080	5,399	5,399

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

Table A5

OLS Predicting Belief of Partner's Part 1 Score				
	Male Partners		Female Partners	
	I	II	III	IV
Partner's Gender Average Score in Category	0.78**** (0.057)	0.80**** (0.076)	0.74**** (0.054)	0.71**** (0.076)
Partner's Gender Adv. in Category in Part 1	1.35**** (0.12)	1.25**** (0.17)	-0.40**** (0.12)	-0.23 (0.17)
Female Evaluator		0.001 (0.68)		-1.11 (0.68)
Female Evaluator x Partner's Gender Avg. Score in Category		-0.04 (0.116)		0.07 (0.109)
Female Evaluator x Partner's Gender Adv. In Category		0.21 (0.23)		-0.32 (0.24)
Constant	0.89*** (0.34)	0.88* (0.48)	2.09**** (0.34)	2.66**** (0.51)
R-squared	0.20	0.21	0.19	0.21
Clusters	196	196	185	185
N	784	784	740	740

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

C4. More on Context Dependence

In Section 5.4 we presented results on context dependence in beliefs of own ability. In Appendix Table A7, 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 Part 3 data, demonstrated by the significant interaction of partner female and male advantage in Part 3 in specification I. We find a directionally similar result in the Part 1 data, though it is not significant (p=0.14).

Appendix Table A7

OLS Predicting Own Believed Probability of Answering Correctly in Part 3		OLS Predicting Own Believed Part 1 Score	
	I (Pooled)		II (Pooled)
Male Adv. in Pt. 3	0.11 (0.078)	Male Adv. in Pt. 1	0.48**** (0.142)
Fitted Value of $\hat{I}_{i,j}$	0.54**** (0.018)	Part 1 Score	0.67**** (0.033)
Partner Female	-0.01 (0.013)	Partner Female	-0.14 (0.180)
Partner Female x Male Adv. in Part 3	0.19** (0.093)	Partner Female x Male Adv. in Part 1	0.25 (0.171)
Female	-0.05** (0.022)	Female	-0.47 (0.353)
Female x Male Adv. in Pt. 3	-0.39**** (0.093)	Female x Male Adv. in Pt. 1	-0.65**** (0.188)
Female x Fitted Value of $\hat{I}_{i,j}$	0.060** (0.026)	Female x Part 1 Score	0.12** (0.050)
Constant	0.39**** (0.018)	Constant	1.63**** (0.255)
R-squared	0.21	R-squared	0.39
Clusters	381	Clusters	381
N	11,479	N	1,524

Notes: Includes laboratory data from OSU and Harvard 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 Part 1 and Part 3 data (average Asian performance – average performance of all

non-Asians in sample). We proxy for ability as we did for gender: in Part 1 analysis, we simply use Part 1 score in category and in Part 3 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 non-Asian 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 A8 demonstrates that is indeed what we find, both in Part 1 data and Part 3 data.

Table A8

OLS Predicting Own Believed Probability of Answering Correctly in Part 3		OLS Predicting Own Believed Part 1 Score	
	I (Non-Asians)		II (Non-Asians)
Asian Adv. in Pt. 3	-0.30**** (0.048)	Asian Adv. in Pt. 1	0.33**** (0.071)
Fitted Value of $\hat{I}_{i,j}$	0.67**** (0.021)	Part 1 Score	0.68**** (0.041)
Partner Asian	-0.03 (0.022)	Partner Asian	-0.00 (0.215)
Partner Asian x Asian Adv. in Part 3	-0.29*** (0.108)	Partner Asian x Asian Adv. in Part 1	-0.27** (0.125)
Constant	0.32**** (0.017)	Constant	2.21**** (0.231)
R-squared	0.24	R-squared	0.33
Clusters	131	Clusters	131
N	5,240	N	524

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.

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. 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 include a proxy for individual ability, our fitted $\hat{I}_{i,j}$ term from Section 5, in order to account for any role own ability plays in driving these effects.

Appendix Table A9

OLS Predicting Participant "Contributed" Answer						
	Men			Women		
	I	II	III	IV	V	VI
Partner Female	0.047* (0.028)	0.018 (0.031)	-0.012 (0.025)	0.095*** (0.030)	0.048 (0.033)	0.024 (0.026)
Fitted Value of $\hat{I}_{i,j}$	0.16**** (0.026)	0.16**** (0.026)	0.42**** (0.035)	0.21**** (0.030)	0.19**** (0.030)	0.51**** (0.031)
Male Adv. in Part 3		-0.15 (0.15)	-0.038 (0.12)		-1.19**** (0.16)	-0.66**** (0.13)
Partner Female x Male Adv. in Part 3		0.79**** (0.20)	0.19 (0.18)		1.32**** (0.20)	0.61**** (0.17)
Partner Place in Line			0.17**** (0.016)			0.22**** (0.010)
Constant	0.60**** (0.027)	0.61**** (0.028)	0.11* (0.057)	0.48**** (0.030)	0.54**** (0.032)	-0.097** (0.038)
R-squared	0.01	0.02	0.20	0.02	0.03	0.27
Clusters	197	197	197	184	184	184
N	7,880	7,880	7,880	7,360	7,360	7,360

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

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 Part 3 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 fewer answers as male advantage increases when paired with a male partner, but 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, men do not react to the gender of their partner. Women, however, do adjust their place in line based upon the gender of their partner. They move significantly back in line as male advantage increases when paired with a male partner, but significantly less so when they are paired with a female partner.

Appendix D: Robustness Tests

First, we show that results are quite similar when restricted to the sample that attended high school in the United States. Table A10 shows the results for self-beliefs. The only difference from our main results is the estimate of the constant for women's self-beliefs in Part 1. In the full sample, we estimated a constant of 0.90 (SE 0.20); in the restricted

sample, the estimated constant increases to 1.48 (SE 0.24). This suggests that in this sub-sample, the gender differences in DIM that we observed in the full sample for Part 1 self-beliefs are reduced.

Table A10

Replication of Table II: Self-beliefs							
OLS Predicting Own Believed Probability of Answering Correctly in Pt. 3 US HS ONLY				OLS Predicting Own Believed Pt. 1 Score US HS ONLY			
	Para-meter	I (Men)	II (Women)		Para-meter	III (Men)	IV (Women)
Own Gender Adv. in Pt. 3	$\theta\sigma$	0.34**** (0.052)	0.15** (0.058)	Own Gender Adv. in Pt. 1	$\theta\sigma$	0.78**** (0.099)	0.14 (0.129)
Fitted Value of $\hat{I}_{i,j}$	ω	0.55**** (0.014)	0.59**** (0.015)	Individual's Pt. 1 Score in Category	ω	0.68**** (0.031)	0.73**** (0.034)
Constant	c	0.38**** (0.012)	0.34**** (0.012)	Constant	Nc	1.48**** (0.224)	1.48**** (0.241)
R-squared		0.22	0.21	R-squared		0.34	0.38
Clusters		289	221	Clusters		289	221
N		9,338	6,700	N		1,156	884

Notes: Pools observations for Ohio State and Harvard experiments. Standard errors clustered at the individual level.

In Table A11, 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.

Table A11

Replication of Table III: Beliefs about Others							
OLS Predicting Belief of Partner's Probability of Answering Correctly in Part 3 US HS ONLY				OLS Predicting Belief of Partner's Part 1 Score US HS ONLY			
	Para-meter	I (Men)	II (Women)		Para-meter	III (Men)	IV (Women)
Partner's Gender Adv. in Category in Pt. 3	$\theta\sigma$	0.15** (0.063)	0.21*** (0.070)	Partner's Gender Adv. in Category in Pt. 1	$\theta\sigma$	1.31**** (0.14)	-0.29** (0.13)
Share of Partner's Gender Answering Qn. Correctly	ω	0.37**** (0.019)	0.36**** (0.017)	Partner's Gender Avg, Score in Category in Pt. 1	ω	0.78**** (0.066)	0.70**** (0.059)
Constant	c	0.43**** (0.016)	0.46**** (0.014)	Constant	Nc	0.85** (0.38)	2.34**** (0.37)
R-squared		0.13	0.11	R-squared		0.22	0.18
Clusters		148	156	Clusters		148	156
N		4,480	4,479	N		592	624

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

In Appendix Tables A12 and A13, we replicate the Part 1 results on self-beliefs and beliefs about others using the MTurk data. Results on stereotypes for beliefs about women get much stronger. For men, however, the results get weaker. While the role for stereotypes in beliefs about men looks similar, the role of stereotypes in predicting men's self-beliefs has an unexpected sign, with an increase in male advantage predicted to decrease men's beliefs about own ability.

Table A12

OLS Predicting Own Believed Pt. 1 Score				
	Laboratory (out of 10 questions)		Mechanical Turk (out of 5 questions)	
	I (Men)	II (Women)	III (Men)	IV (Women)
Own Gender Adv. in Pt. 1	0.73**** (0.092)	-0.02 (0.029)	-0.47**** (0.057)	1.46**** (0.071)
Individual's Pt. 1 Score in Category	0.70**** (0.027)	0.81**** (0.029)	0.47**** (0.014)	0.46**** (0.014)
Constant	1.36**** (0.191)	0.90**** (0.196)	1.68**** (0.051)	1.58**** (0.055)
R-squared	0.37	0.42	0.27	0.36
Clusters	344	296	987	843
N	1,376	1,184	5,922	5,064

Notes: Pools observations for Ohio State and Harvard experiments. Standard errors clustered at the individual level.

Table A13

OLS Predicting Belief of Partner's Part 1 Score				
	Beliefs about Men		Beliefs about Women	
	Lab (out of 10 questions)	Mturk (out of 5 questions)	Lab (out of 10 questions)	Mturk (out of 5 questions)
	I	II	III	IV
Partner's Gender Average Score in Category	0.78**** (0.057)	0.65**** (0.020)	0.74**** (0.054)	0.03 (0.018)
Partner's Group Adv. in Category in Part 1	1.35**** (0.12)	1.07**** (0.048)	-0.40**** (0.12)	2.06**** (0.062)
Constant	0.89*** (0.34)	0.62**** (0.07)	2.09**** (0.34)	3.13**** (0.07)
R-squared	0.20	0.06	0.19	0.19
Clusters	196	1,826	185	1,826
N	784	10,986	740	10,986

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