Evidence on Self-Stereotyping and the Contribution of Ideas

Katherine Baldiga Coffman*

June 3, 2014

Abstract

We use a lab experiment to explore the factors that predict an individual's decision to contribute her idea to a group. We find that contribution decisions depend upon the interaction of gender and the gender stereotype associated with the decision-making domain: conditional on measured ability, individuals are less willing to contribute ideas in areas that are stereotypically outside of their gender's domain. Importantly, these decisions are largely driven by self-assessments, rather than fear of discrimination. Individuals are less confident in gender incongruent areas and are thus less willing to contribute their ideas. Because even very knowledgeable group members under-contribute in gender incongruent categories, group performance suffers and, ex post, groups have difficulty recognizing who their most talented members are. Our results show that even in an environment where other group members show no bias, women in male-typed areas and men in female-typed areas may be less influential. An intervention that provides feedback about a woman's (man's) strength in a male-typed (female-typed) area does not significantly increase the probability that she contributes her ideas to the group. A back-of-the-envelope calculation reveals that a "lean in" style policy that increases contribution by women would significantly improve group performance in male-typed domains. Classifications: J16, C92

*Thank you to the Russell Sage Foundation for their funding of this project. Excellent research assistance was provided by Erica Bailey and Mackenzie Alston. The author is also very grateful for helpful conversations about this work with Max Bazerman, Iris Bohnet, Pedro Bordalo, Lucas Coffman, Drew Fudenberg, Jerry Green, Paul Healy, John Kagel, Judd Kessler, David Laibson, Katy Milkman, Johanna Mollerstrom, Muriel Niederle, James Peck, Andrei Shleifer, the seminar audience at the Women and Public Policy Program at the Harvard Kennedy School, the Harvard Experimental and Behavioral Seminar, Cornell, and Stanford and the participants of the Experimental Methods in Policy Conference and BEAM 2014. Mailing Address: 1945 N. High St, Arps Hall 410, Columbus, OH 43210; Email: coffman.201@osu.edu; Telephone: 614-247-8718.

1 Introduction

From faculty committees and student projects to business teams and corporate boards, many decisions are made by groups rather than by individuals. In these settings, group members may bring differing levels of knowledge and expertise to the table; therefore, the quality of the group's decision depends largely upon how successfully the group can elicit and act upon the best information from the most informed members. While a large economics literature addresses the information aggregation issue at the heart of group decision-making – social choice theory, mechanism design, political economy, and other fields have offered both theoretical and empirical insights – less attention has been paid to the information contribution stage of group decision-making. The contribution stage is crucial, as the output of any aggregation procedure can only be as good as its best inputs. In this paper, we explore what determines whether an individual contributes her ideas to a group.

We will focus in particular on the role that gender plays in predicting the decision to contribute ideas to a group. It is well-known that women are under-represented in many domains that are stereotypically perceived as masculine (i.e. STEM fields, business). Explanations for these gender gaps have focused on ability, human capital, demand for flexibility and work-life balance, role models, discrimination, and, more recently, competitiveness (see, for example, Xie and Shauman [2003], Bertrand, Goldin, and Katz [2010], Buser, Niederle, and Oosterbeek [2014], and Goldin [2014]). In this paper, we explore a new and potentially important additional factor: willingness to contribute one's ideas. If women are less likely to contribute their ideas, this could hinder advancement. It is much harder to recognize that someone is talented if that person does not share her ideas.

Previous psychology and sociology research has suggested that women may be more reluctant than men to share ideas in group settings (see, for example, Crosby and Nyquist [1977], Thomas-Hunt and Phillips [2004], and Babcock and Laschever [2007]). Using a clever experiment, Thomas-Hunt and Phillips (2004) demonstrate that within a mixed-sex group, expert women have less influence than expert men. There is also emerging evidence in economics that women are less likely to give advice to fellow group members in strategic settings (Cooper and Kagel [2013]).

Our experiment builds on this existing literature in a few key ways. First, we create a carefully controlled laboratory environment that isolates the decision to contribute an idea. In existing work

in this area, individuals make the decision of whether or not to contribute an idea to the group and fellow group members decide how to act upon this contribution. An expectation that an idea may be disregarded could reduce the incentive to contribute. In our environment, described in detail below, we rule out discrimination. Each individual decides independently how willing she is to contribute her idea to the group. And, there is no group deliberation stage: there is a fixed decision rule that implements the decision suggested by the individual who was most willing to contribute her idea. Therefore, differential expectations about the probability of being listened to cannot drive any differences in willingness to contribute that we observe. Second, we explore the decision to contribute in a variety of domains, covering both areas that are generally perceived as female-typed as well as areas that are generally perceived as male-typed. We can test whether the gender differences presented in earlier work persist in more female-typed domains.

In our lab experiment, participants first take a brief test of their knowledge in six different categories: Arts and Literature, Pop Culture, Environmental Science, History, Geography, and Sports and Games. They are then randomized into groups of two. These groups face a new set of questions in the same set of categories. For each question, each group member chooses an individual answer to the question and decides how willing she is to have her answer count as the "group answer". The participant who is most willing to contribute automatically has her answer submitted as the group answer. Each member of the group earns payment based upon whether the group answer is correct. By measuring willingness to contribute, we aim to capture the spirit of many decisions outside of the lab: deciding how long to wait before speaking up or raising a hand, deciding how assertively to state an opinion, or deciding whether to allow other people to speak first before offering an idea.

We find systematic inefficiencies in decisions to contribute answers to the group. In particular, we find that the decision to contribute is predicted not only by our proxies for the ability of the participant, but also by the interaction of her gender and the gender stereotype associated with the category. Both men and women are significantly more willing to contribute ideas when the stereotype associated with the category is gender congruent. Conditional on their measured ability within a category, women are significantly less likely than men to contribute their answers to the group when the question comes from a category that is perceived as male-typed. Similarly, men are significantly less likely than women to contribute answers to the group when the question

comes from a category that is perceived as female-typed. Consistent with previous literature, a participant's confidence in her ability varies significantly with how gender congruent the stereotype associated with a category is, and these differences, in part, explain the contribution patterns we observe.

The fact that participants are more reluctant to contribute in gender incongruent categories has negative consequences for group performance. First, groups miss out on correct answers: conditional on knowing the right answer, an individual is much less likely to contribute that answer in a gender incongruent category. Second, groups have difficulty recognizing who their experts are. After the group decision-making process, we ask group members who knew the most about each category. Groups are much less likely to successfully identify who is most knowledgeable when the expert's gender does not match the gender stereotype of the category. This is not driven by discrimination; female experts in male-typed domains do not get recognized simply because they contribute their ideas less often than their male counterparts. The same is true for male experts in female-typed domains.

Our attempt to improve the efficiency of group decision-making fails. We find that providing feedback to individuals informing them of their relative strengths and weaknesses prior to the group decision-making stage does not successfully solve the under-contribution problem. Receiving positive feedback does not increase the probability that a high-ability group member contributes her ideas to the group. The failure of this intervention suggests the need for additional research on policy prescriptions. In the final part of the results section, we do a calculation of the potential gains from a "lean in" policy, under which we systematically increase the probability that a woman's answer is contributed.[1] In our data set, increasing the number of answers contributed by women leads to a significant improvement in group performance in male-typed domains.

The evidence we present is consistent with the findings of Benjamin, Choi, and Strickland (2010), who demonstrate that norms related to social identity can have a significant impact on economically-important decisions. Our data also provide support for many of the theoretical predictions of the self-stereotyping model presented by Bordalo, Gennaioli, and Shleifer (2014). In particular, their model predicts that gender stereotypes shape self-assessments and that these stereotype-driven self-

[1] "Lean in" is a reference to the recent popular book by Sheryl Sandberg (2013), in which she encourages women to assert themselves in their careers.

assessments discourage women from participating in male-typed domains (but not female-typed domains). These predictions and our results are consistent with emerging evidence from the gender and competition literature: Grosse and Riender (2010), Kamas and Preston (2010), Dreber, Essen, and Ranehill (2011) and Shurchkov (2012) show that the well-documented gender gap in competitiveness (as first documented by Gneezy, Niederle, and Rustichini [2003] and Niederle and Vesterlund [2007]) is reduced when a stereotypically female task is used.

Our results have important implications for group decision-making settings. Many researchers have documented the impact of gender stereotypes on how one evaluates or perceives another person (see, for example, Eagly and Mladinic [1989], Heilman [2002], Bohnet, Van Geen, and Bazerman [2012], and Milkman, Akinola, and Chugh [2013]). We show that self-stereotyping may also play a critical role in economic environments. Individuals are less likely to contribute their ideas in areas that they perceive as outside of their gender's domain, even when discrimination, and fear of discrimination, is minimized. In our environment, we see that talented women are much less likely than talented men to be recognized as most knowledgeable in male-typed domains, simply due to the fact that they contribute their ideas less often. This finding could provide fresh insights into gender disparities in the allocation of promotions, raises, and leadership roles, particularly in male-dominated fields.

2 Experimental Design

The main goal of the experiment is to test how willingness to contribute ideas to a group depends upon gender and the gender stereotype associated with the decision-making domain. To better understand what drives willingness to contribute, we also collect data on ability and beliefs within each domain.

The experiment consisted of six incentivized parts and a seventh part that collected demographic information. All participants received general instructions informing them that one part of the experiment had been pre-selected for payment and would be announced at the end of the experiment. They received a $5 show-up fee and $1.50 per point earned on the pre-selected part. With the exception of the feedback treatment detailed explicitly below, participants received no information about their own or others' performance until the end of the experiment.

Participants faced multiple-choice questions from six categories: Arts and Literature (Art), Entertainment and Pop Culture (Pop), Environmental Science (Env), History (Hist), Geography (Geo), and Sports and Games (Sports). The categories were chosen to vary in how male or female-typed they were.[2] Each question had five possible answers. The general structure of the experiment is as follows: (1) participants answer multiple-choice questions in each category on their own, providing a baseline measure of individual ability in each category, (2) participants are randomized into groups and make decisions about how willing they are to contribute each of their answers to new questions in these categories to their group, and (3) participants answer additional questions about their beliefs. The heart of the paper is (2), where we collect data on willingness to contribute ideas to the group. We walk through the details of the experimental design below. We use a 2x2 across subject design in which we vary whether participants receive feedback about their individual ability prior to making contribution decisions and the publicness of the contribution. We describe both of these interventions in detail at the conclusion of the design section. Complete instructions as they appeared to participants are provided in the Online Appendix.

2.1 Parts A and B: Individual Ability

In Parts A and B of the experiment, we test participant ability in each of the six categories, observing performance on multiple-choice questions under two different protocols. In Part A, participants are allowed to skip questions. In Part B, all participants must provide answers to each question. This allows us to observe how willing individuals are to guess when unsure about the answer to a question (Part A) while still collecting an unbiased measure of their ability within the category (Part B). Because we do not use the data from Part A in the main text, we defer the full discussion of this section of the experiment to the Online Appendix.

In Part B of the experiment, participants took a 30-question test of their ability, answering five questions in each of the six categories. The order of the questions was randomized. Each multiple-choice question had five possible answers. All questions appeared on the same page, labeled with their category. Participants were forced to provide an answer to each question, earning 1 point for every correct answer provided. We will use the scores from Part B as a proxy for individual ability

[2] We collect data from the participants on their own perceptions of these categories in the final part of the experiment. Indeed, we confirm that these categories vary significantly in their perceived gender-type.

in each category.

2.2 Part C: Willingness to Contribute Answers to the Group

Part C measured willingness to contribute answers for the group. Participants were randomized into groups of two (but remained at their individual computer terminals). They were informed only that their partner was in their session. Individuals then faced five new questions from each of the same six categories. Each multiple-choice question had five possible answers. For each question, they were asked to provide an answer and their willingness to have their answer count as the group answer.

Willingness to submit one's answer as the group answer was elicited on a 1-4 scale, explained as "choosing a position in line". For each question, a participant was asked to choose a position in line, either 1, 2, 3, or 4. Of the two group members, the member who selected the lowest number, i.e. the position closest to the front of the line, would have their answer for that question submitted for the group. If both members selected the same number, the computer randomly selected one member's answer. Payment for Part C depended on the answers submitted for the group. A correct group answer earned each group member 1 point, an incorrect group answer lost each group member 1/4 point. At the time of their decision, participants were aware that the answer submitted for the group would be seen by both group members and would determine payment for Part C.[3]

The elicitation mechanism for willingness to contribute was designed to be both intuitive for participants and encourage variance. Even though there are only two members in the group, we allow for four possible position choices to give participants more flexibility in expressing their willingness to contribute, increasing our statistical power.[4] We do not write down a theoretical model

[3]Note that we study an environment in which the incentives of the group members are perfectly aligned. An individual and her group earns points if the group answers correctly. In this way, the group answering correctly benefits both an individual and her group. So, we expect other-regarding preferences will not play a role in predicting behavior. Furthermore, there are no "bonus" points allocated to members who contribute ideas more often, something which could create misalignment of the individual and group preference.

[4]The primary purpose of having four places in line is to collect finer information about preferences. It is true that if her partner chooses 1, it ultimately doesn't matter if an individual chooses 2, 3, or 4: the outcome will be the same. But an individual does not know what her partner will choose. The idea is that someone who chooses 2 may be expressing a much different preference than someone who chooses 4. A choice of 2 may say something like, "I want to answer for the group unless my partner is very sure." A choice of 4 may say something like, "I really do not want to answer for the group". While we can't be sure that participants shared exactly this type of intuition, we do know that regardless of how they view this procedure, stating a lower number generates a weakly greater probability of contributing one's answer to the group. See the Online Appendix for more on how participant beliefs mapped into places in line.

for how participants might use this mechanism. Instead, we will rely on just one important theoretical implication of this procedure: choosing a lower position in line weakly increases the probability of answering for the group. Therefore, our interpretation of this data as indicative of "willingness to contribute" seems appropriate. After reading the instructions for Part C, participants answered one quiz question aimed at evaluating their understanding of the mechanism; 98% of participants answered this question correctly. For more on how participants used this mechanism, please see the Online Appendix.

One could think of a participant's choice of position in line as paralleling how quickly she might speak up in group decision-making contexts. Of course, it isolates one important, but very specific, aspect of the decision to contribute to a group. We observe only an individual's initial willingness to contribute her answer; there is no opportunity for discussion and information about her partner is limited. The advantage is that this provides clean data that can serve as a first step toward understanding willingness to contribute.

Immediately following the instructions for Part C, participants made incentivized guesses about whether they had the highest Part B score in their group for each category. They were told that if Part C was selected for payment, they would receive an additional $0.25 for each correct guess. Participants made all of their willingness to contribute decisions on one page; group answers were not revealed until later in the experiment.

2.3 Parts D and E: Confidence and Risk Preferences

In Part D, we measured participants' confidence in their own answers and in their partners' answers. They saw each of the 30 questions from Part C again and were asked in an incentive-compatible way to estimate the probability that their own answer was correct and to estimate the probability of their partner answering the question correctly.

To incentivize these responses, we used a mechanism proposed by Karni (2009) and previously implemented experimentally by Mobius et al (2013). Participants were told there are 100 robots, each with integer probability on [1,100] of answering a question correctly. That is, there was a robot who would answer correctly 1% of the time, a robot who would answer correctly 2% of the time, ... , and a robot who would answer correctly 100% of the time. They were told that for each question, one robot would be drawn at random who could answer the question for them. They

have to decide which robots they would allow to answer the question for them; they submitted an "accuracy threshold" – a number, Y, between 1 and 100 such that if robot answered correctly less than Y percent of the time, they would prefer to have their own answer submitted instead. A correct answer submitted, whether it was theirs or the robot's, earned 1/2 a point; 0 points were earned for incorrect answers.

The same mechanism was used to elicit their believed probability that their partner's answer from Part C was correct. A separate robot was drawn for own answer and partner's answer for each question. Again, 1/2 a point was earned for a correct answer, whether it was their partner's or a robot's. Note that participants do not know their partner's answer or identity at this stage.[5] So, this elicits a more general belief about the ability of a person chosen at random from the laboratory. Following the instructions for Part D, participants answered one quiz question about the robot mechanism; 98% of participants answered this quiz question correctly.

We use the data from Part D to test for two plausible hypotheses for why we would see differences in willingness to contribute to the group: (1) individuals vary in their beliefs about their own probability of answering correctly, and (2) individuals vary in their beliefs about the probability of the other group member answering correctly. Either could, in principle, influence the decision to contribute: an individual might be less likely to answer because she believes her chances of being correct are, in absolute terms, low, or because she believes her chances of being correct are low relative to the average person. By collecting data on both types of beliefs, we can more thoroughly explore the role of confidence in predicting behavior.

Economists and psychologists have documented that individuals have difficulty calibrating beliefs about ability, both relative to actual ability and relative to others (see Moore and Healy [2007]). Existing work also highlights gender differences in confidence; importantly, the direction of the gender gap depends upon the domain, with men displaying more confidence than women particularly in male-typed domains (see Beyer [1990], Beyer and Bowden [1997], Beyer [1998], and Hong [2010]). Confidence has also been shown to be a key explanation for observed gender differences in competitive group settings; most notably, Niederle and Vesterlund (2007) show that one reason why women enter into competitive tournaments less often than men is because they are less

[5] What is important is that they have the same information about their partner in this stage as they do in Part C, when they are making decisions about willingness to contribute. Thus, any data we collect here on beliefs about their partner are likely to reflect the beliefs they had about their partner when making their contribution decisions.

confident.

Part E was designed to elicit risk preferences. Because the data from this section is not central to the results presented below, we defer discussion of this section to the Online Appendix.

2.4 Part F: Agreement with Group Answers

At the conclusion of Part E, participants received a handout from the experimenter containing the answers submitted for the group in Part C. For each question, they saw which group member was selected to answer for the group (based upon the responses provided in Part C), Member 1 or Member 2. They also saw which answer that member submitted. Then, in Part F, participants faced the same 30 questions from Part C again. They had a new opportunity to earn points based upon their answers to these questions. Participants had to provide an answer to each question, but they were free to provide any answer they wished, agreeing with the group answer or not. A correct answer was worth 1 point; there were no penalties for wrong answers. While we choose to describe Part F here for completeness, the data from this section will be part of a separate paper.

After finishing Part F, participants were asked to guess which member of their group knew the most about each category. They were told that if they and their partner both submitted the same group member, they would receive an additional $0.25 if Part F was chosen for payment.

Finally, participants were asked the following demographic questions: gender, year of birth, race, whether or not they attended high school in the United States, student/employment status, which categories they liked/disliked, and which categories they know the most/least about. The very last question asked participants to evaluate how male-typed or female-typed each category was. They were asked, "for each of the categories tested above, tell us whether you think men or women, on average, know more about it." They were asked to indicate their answer on a sliding scale, where -1 was labeled as "women know more" and 1 was labeled as "men know more", and 0 was labeled as "no gender difference."

2.5 Treatment Interventions

We use a 2x2 across subject design. First, we vary whether or not the participants received feedback about Part B performance. In the feedback treatments, each participant received a sheet of paper, immediately following the instructions for Part C, listing the categories for which they had the

highest Part B score in their group.[6] We hypothesize that receiving feedback about performance should improve the efficiency of group decision-making, with the participants who know they have the highest (lowest) Part B scores contributing more (less) often. Second, we vary the publicness of the contribution. In the public treatments, each participant had her photograph taken at the beginning of Part C. They were told that their picture would be shown to their partner, along with the answers they submitted for the group. Participants received printed copies of the pictures of their partners at the beginning of Part F.[7] The primary goal of this treatment is to explore whether agreement with group answers in Part F depends upon the gender of one's partner (as explained above - this hypothesis will be explored in a separate paper). However, it is possible that knowing that her photograph will be provided to her partner along with her answers also impacts a participant's willingness to contribute to the group. We hypothesize that the addition of photographs, with the added accountability they provide, should drive more (less) informed members to contribute more (less) often.[8]

In Figure I, we include a graphic representation of the general setup of the experiment. The flow-chart demonstrates what the participant does in each part of the experiment. We also illustrate when the treatment interventions take place. In most of the analysis that follows, we will be analyzing the data from Part C, in which participants made decisions about how willing they were to submit their answers for the group. We will use the data from Part B to proxy for individual ability and data from Part D to control for confidence.

Insert Figure I

Thirty-eight sessions were conducted at the Ohio State Experimental Economics Lab, collecting data from 460 participants. The number of participants per session varied between 4 and 30, with the median session size being 12 participants. Average earnings were $18.68 and each session lasted approximately 80 minutes. Treatments were assigned at the session level. Table **??** in the Online

[6] If there was a tie, the category was listed for both members. Participants were informed of this rule.

[7] Note that they do not have these photographs during Part C, when they are making decisions about whether to contribute answers to the group. So, no information about partner's gender is available for this key stage of data collection.

[8] It is also possible that having one's photograph taken makes identity more salient. If the decision to contribute is tied to identity, in particular gender identity, than it is possible that taking photographs might induce individuals to behave in ways that they perceive as being more consistent with their gender identity.

Appendix presents the distribution of men and women across treatments. The gender asymmetry within treatment is not intentional and reflects the underlying composition of the laboratory subject pool.

3 Results

3.1 Summary Statistics

Table I presents summary statistics for the participants. There are many significant differences between the men and women in sample, both in demographic characteristics and in performance. We take these into consideration in the analysis below.

Insert Table I

Rather than rely on an external source for labeling the categories as male- or female-typed, we collected data from our participants.[9] The final question of the experiment asked participants to rank each category according to whether, in general, men or women know more about it. They used a sliding [-1,1] scale, where -1 was labeled "women know more", 0 was labeled "no gender difference", and 1 was labeled as "men know more". In Table II, we present the data collected for this question. On average, men and women agree on the ordering of the categories; from most to least female-typed, they are Arts and Literature, Pop Culture, Environmental Science, History, Geography, and Sports and Games. In fact, 66.3% of participants provide this ordering (note: the categories do not appear in this order). We provide the average perception for each category in the third column; note that each average is significantly different from 0 ($p<0.001$). In the last column, we provide the normalized z score for "maleness" of the category: we re-scale the average perceptions to be mean 0 with a standard deviation of 1. Going forward, we will refer to Arts and Literature and Pop Culture as female-typed and Environmental Science, History, Geography, and Sports and Games as male-typed.

[9] Of course, while this data has the advantage of being generated by our sample, it also has the drawback of not being independent of the experiment itself. These participants have just answered questions in each category; their perceptions of each could in principle depend on their specific performance in this task. However, we find no evidence of this. Part B and Part C scores have no predictive power of category perceptions.

Insert Table II

In Table III, we present the raw data: average places in line chosen by men and women in Part C, broken down by category. Recall that lower numbers indicate a greater willingness to answer for the group. The gender differences in places in line closely track both the differences in ability captured in Table I and the stereotypes captured in Table II. We analyze these differences in detail in the sections that follow.

Insert Table III

3.2 Gender Stereotypes Predict Willingness to Contribute to the Group

Our first step is simply to document that a participant's willingness to contribute her answer depends upon the interaction of her gender and the gender stereotype associated with that category. In Table IV, we predict a participant's chosen position in line for Question i from her gender, whether or not she knew the correct answer to Question i (a proxy for her question-specific ability), and her Part B score in the category from which Question i was drawn (a proxy for her broader ability in that category).[10] We include what will be our standard set of controls throughout the paper: treatment dummies, race dummies, session size, fraction of women in the session, a dummy for attending high school in the United States, a dummy for being a current undergraduate at Ohio State, and the overall probability of a correct answer for that question (which could be interpreted as the difficulty of the question[11]). The first six columns break down the data by category. In the final column we pool all of the categories and add as a predictor the z score of the individual's reported "maleness" of the category from which Question i was drawn. We cluster our standard errors at the individual level. Recall that lower numbers indicate a greater willingness to contribute.[12]

It is clear that willingness to contribute depends upon the gender congruence of the category,

[10] We note that Part B Scores within a category are significantly and positively correlated with Part C Scores in that category (p-value < 0.001), suggesting that our use of Part B Score as a measure of ability within a category is reasonable.

[11] We do this to account for the fact that categories may vary not only in their gender-type but also in their difficulty.

[12] We report OLS regressions rather than ordered probits to simplify the interpretation of the coefficients. The results are not qualitatively different if we instead choose to run ordered probits.

even controlling for the proxies of ability that we have. Women are significantly more willing than men to contribute answers in Arts and Literature and Pop Culture. Men are significantly more willing than women to contribute answers in the four male-typed categories. The fact that the direction of the gender gap varies with category tells us something important about the phenomenon we are observing. This is not just about gender; it is about the interaction of gender with the gender stereotype associated with the category. The pooled specification demonstrates this result clearly: as the maleness of the category increases, men become significantly more likely to contribute answers to the group. Women become significantly less likely to contribute answers as the maleness of the category increases.

Insert Table IV

Our results are robust to the inclusion of alternative measures of ability. For instance, these results are not changed if we add higher order ability terms (for instance, quadratic or cubic functions of Part B Score), if we allow for non-linearity in scores by using dummy variables for each possible Part B score, or if we proxy for ability with other measures from the experiment (for instance, number of Part C questions answered correctly or total number of questions answered correctly in the category).[13] In the Online Appendix, we report the results of a robustness test that explores whether the noisiness of our ability proxies could explain our results. In the regressions of Table IV, we control for the number of questions answered correctly in Part B, assuming that two participants with the same Part B score are of roughly equal ability. However, one could make the argument that a man with a 3/5 in, say, Sports and Games is likely to be of better ability, on average, than a woman with a 3/5 in Sports and Games, due to the fact that men on average know more about this category in our sample. To account for this, we can perform an adjustment to Part B Scores, systematically reducing the scores of individuals in gender incongruent categories by a full point. When we repeat the analyses of the paper using these adjusted Part B Scores instead of observed Part B Scores, the results remain largely unchanged (see Table **??** in the Appendix),

[13] We have also explored whether Part B Score interacts with gender. If we repeat the analysis of Table IV, predicting position in line from our proxies of ability, the maleness of the category, and gender, and add interaction terms for Part B Score x female, Part B Score x female x the maleness of category, and Part B score x the maleness of category, the results are largely unchanged. None of the new interaction terms are significant (the interaction of female and maleness of the category is still large and significant).

suggesting that ability differences (at least in terms of knowledge of the specific questions we use here) do not drive our results.

Why *does* the gender-type of the category matter for decision-making? We will show that stereotypes matter for beliefs and beliefs strongly predict decisions. Recall that, prior to Part C, participants are asked to guess whether they had the highest Part B score in their group in each category. In Table V, we show that conditional on Part B performance, men and women perceive their probability of having the highest Part B score in their group very differently. In female-typed categories, women are 21 percentage points more likely to guess that they had the highest score in their group than men with the same Part B score. On the other hand, women are 13 percentage points less likely to believe they are best at History, 25 percentage points less likely to believe they are best at Geography, and 42 percentage points less likely to believe they are best at Sports and Games (again, conditional on Part B score). There is no significant gender difference in Environmental Science. In the last column, we pool the data and we predict the probability of ranking one's self first from the maleness of the category and Part B score in that category (along with the other controls). The impact of the maleness of the category is large and significant for both men and women.[14] Individuals in gender incongruent categories greatly under-estimate their probability of having the highest score in their group. Consider the extremes: only 16% of women believe they had the highest Part B score in Sports (even though 55% actually did) and only 32% of men believe they had the highest Part B score in Art (even though 61% actually did).[15]

Insert Table V

These differences in beliefs explain a large share of the gender effects. In Table VI, we predict position in line from gender and our proxies for ability: a dummy for whether or not she has the correct answer for that question and her score in that category from Part B. We also add all of the other data we have from our participants. We add confidence measures: her believed probability of answering that question correctly, her believed probability of her partner answering that question

[14] If we look at beliefs about answering Part C questions correctly, a similar pattern emerges. Table **??** in the Appendix presents the full results.

[15] Again, these results are robust to an adjustment where we adjust down Part B Scores in all gender incongruent categories. Even assuming that observed Part B Scores overstate actual ability in each gender incongruent category by a full point does not account for the gender differences in beliefs we document here.

correctly, and whether or not she believed she was the top performer in that category in Part B. We also add our data on risk preferences (collected in Part E, described in the Online Appendix).

All three of our beliefs measures are significant predictors of willingness to contribute. When we consider the relative importance of each factor, we see that an individual's believed probability of answering question i correctly has the greatest predictive power. The final pooled specification in the last column of Table VI estimates that for a one standard deviation increase in the believed probability of answering correctly, predicted place in line falls by 0.79 places. By comparison, a one standard deviation change in our other measures of beliefs (believed probability of partner answering correctly and whether or not the individual guessed she had the highest Part B Score in that category) or in our proxies for ability (answering question i correctly and Part B Score), change predicted place in line by less than a tenth of a place.[16] Risk preferences, at least as measured in this context, have no predictive power for willingness to contribute.

Differences in confidence explain the entire gender gap observed in the female-typed categories. The gender gaps in stereotypically male categories are also greatly reduced, but we continue to estimate a significant effect of being female on position in line. That is, conditional on our measures of ability and confidence, on average women in our sample are less willing to contribute than men in Environmental Science, History, Geography, and Sports and Games. When we pool the categories in the final column, we see that even controlling for our measures of confidence, the maleness of the category predicts the decisions of both men and women (though the effect sizes are reduced by about 75%). These results suggest that while confidence seems to be a key part of the story, there may be additional factors at work. We discuss this further in Section 4.

Insert Table VI

Before moving on, we pause to think about a framework for understanding how ability, confidence, and gender stereotypes could work together to drive behavior in this setting. By having participants answer questions in each category in Part B, we attempt to measure ability within each category. Of course, our score on a five question quiz serves only as noisy proxy of an individ-

[16] If we repeat the analysis leaving believed probability of answering correctly out, the R squared in the final pooled specification falls from 0.59 to 0.36. We explore the relationship between our measures of confidence and willingness to contribute more thoroughly in the Online Appendix.

ual's true ability within the category. It important to remember that within our sample there are both real and perceived gender differences in average ability within these categories (see Table I for measures of ability and Table II for perceptions). Therefore, the behavior of our participants seems consistent with a model of statistical "self-discrimination" (see, for example, the models of Phelps [1972] and Arrow [1973] and the application to self-stereotyping by Bordalo, Gennaioli, and Shleifer [2014]). That is, in assessing her own ability (e.g., think about forecasting her Part B Score) an individual may place some weight on her noisy signal of ability (perhaps an unbiased belief about how she thinks she did in Part B) and some weight on her gender, both of which carry relevant information. This could lead to the pattern of beliefs we observe in Table V. Within this framework, the decision of a high-ability woman to not contribute her ideas in a male-typed category may be inefficient from her group's perspective (i.e. the group would benefit from hearing more of her correct answers) but not unjustified given her belief of her own ability (which takes into account that women, on average, know less about the category). In fact, in the Online Appendix, we show that given their beliefs, individuals are using the willingness to contribute mechanism quite effectively: the group member with the highest believed probability of answering the question correctly contributes her answer for the group approximately 90% of the time. An important question, then, is whether we can encourage more contributions from high-ability members in gender incongruent categories by providing a better signal of relative ability, positively impacting beliefs. We explore this next.

3.3 Can We Encourage High-Ability Members to Contribute More Often?

Given the discussion above, the obvious next question is whether providing feedback to group members about their relative strengths can encourage high-ability members to contribute more often. Our feedback treatment is designed to address exactly this question. Recall that in the feedback treatment, each group member receives a list of the categories for which she had the highest Part B score. Thus, we expect that a group member who receives this type of positive feedback within a category should contribute more often (relative to an individual of similar ability in a no feedback treatment).

Overall, we find only weak evidence that feedback increases willingness to contribute among knowledgeable group members. In Table VII, we again predict a participant's chosen position in

line for Question i from our proxies of ability (Part B score in the category, whether or not she had the correct answer to Question i), gender, and our standard set of controls. This time, however, our focus is on the coefficients on the treatment variables and their interaction with gender. To simplify our analysis, we present the data pooled for female-typed categories (Arts and Literature, Pop Culture), male-typed categories (Environmental Science, History, Geography, and Sports and Games), and all categories (with an added variable for the maleness of the category). Importantly, we look only at the behavior of those individuals who received good feedback - those that had the highest Part B score within the category. We do so because we hypothesize that feedback should only increase willingness to contribute among those individuals who receive positive feedback.[17]

In general, an individual that receives feedback that she had the highest Part B score in the category is directionally more willing to contribute than an individual with the highest Part B score that does not receive this feedback. This effect is directional in female-typed categories, marginally significant in the male-typed categories, and only directional when we pool all of the data. Importantly, the impact of receiving feedback does not vary by gender or the gender congruence of the category. If women were more sensitive to feedback regardless of the maleness of the category, we would pick this up in the interaction between gender and the feedback treatment. If, on the other hand, all individuals were more sensitive to feedback in gender incongruent categories, we would pick this up in the triple interaction of gender, the feedback treatment, and the maleness of the category in the final pooled specification. However, our model does not provide support for either of these stories: none of the interaction terms are significant.

Table VII also allows us to explore the impact of the public treatment, in which participants had their photographs taken prior to making their decisions about willingness to contribute. The public treatment had no significant impact on position chosen in line and it did not interact with gender or the gender-type of the category. We reach similar conclusions if we consider the full sample (not just those individuals who had the highest Part B score in the category).

Insert Table VII

[17] Alternatively, we could look only at the behavior of those that received bad feedback. When we do so, we derive similar results: the impact of the treatments is minimal. We could also consider all individuals and include additional interaction terms with a dummy indicating whether or not the individual had the highest Part B score in her group (again, the results would be unchanged). Category-specific results are also available in the Online Appendix.

These results suggest that feedback, at least in the form that we provide it, is a rather ineffective policy tool in improving group decision-making in this context. Of course, we employ a rather coarse type of feedback - individuals only learn of relative ability on a five-question test in the category. In future work, it would be useful to explore whether more detailed feedback about own and others' performance would be more successful in encouraging contributions from high-ability individuals.

3.4 The Impact on Group Performance: Missed Opportunities and Unrecognized Experts

We now investigate how the contribution decisions documented above impact group performance. Groups do outperform individuals: an individual answers a Part C question correctly 51% of the time, while a group submits the correct answer 58% of the time (test of proportions, $p<0.001$). Let's put this performance into context. If for each question one group member's answer was randomly selected, the group would be expected to answer correctly only 51% of the time.[18] And, if we used a truth wins norm, where the correct answer is automatically submitted for the group whenever at least one group member has it, the group would answer correctly 71% of the time.[19] Thus, while our observed group performance is above the random selection benchmark, our groups fall short of achieving efficiency.

We start by proposing one measure of inefficiency: define a "missed opportunity" as a situation in which a group member answers question i correctly, but is not as willing to answer for the group as her partner and thus fails to answer for the group. Note, this is only a potential missed opportunity, in the sense that it is possible that the partner will submit the correct answer for the group.[20] Overall, the probability of a missed opportunity is 23%; that is, conditional on having the right answer, a participant fails to submit that answer for the group 23% of the time. Note that in approximately 68% of these cases, the group ultimately answers incorrectly (that is, the partner's submitted answer is incorrect). Thus, missed opportunities are very often costly. Overall, there is

[18] = (The Probability of Both Group Members Having the Correct Answer) + $\frac{1}{2}$(Probability of Exactly One Group Member Having the Correct Answer) = $30.8 + \frac{1}{2}(39.9)$

[19] = (The Probability of Both Group Members Having the Correct Answer) + (Probability of Exactly One Group Member Having the Correct Answer) = $30.8 + 39.9$

[20] We choose not to condition on the partner actually having the incorrect answer (i.e. a realized missed opportunity), as this information is not known at the time of the member's decision. Potential missed opportunities have the advantage of depending less on the quality of one's partner, making our analysis less dependent on the random draw of who one's partner is.

a significant gender gap in the rate of missed opportunities: 20% for men, 28% for women, test of proportions $p < 0.001$. However, the direction of the gender gap varies across category; thus, the fact that women have more missed opportunities is mostly a consequence of the fact that most of the categories are perceived as male-typed.

In Figure II, we graph the rates of missed opportunities (conditional on the individual having the right answer to Question i). The data are presented as follows: we break down the data by gender and the gender-type of the category, pooling Arts and Literature and Pop Culture as female-typed categories and Environmental Science, History, Geography, and Sports and Games as male-typed categories. To better account for the fact that participants have varying abilities, we partition the data according to Part B score. Figure II thus compares the rate of missed opportunities for men and women with the same Part B score. In the female-typed categories, there are few significant differences across gender. In the male-typed categories, however, there are sizable and significant gender gaps in the rate of missed opportunities, even among the most talented participants. Women who had a score of 4 out of 5 in Part B in a male-typed category are 12 percentage points more likely to have a missed opportunity than similarly knowledgeable men ($p<0.01$, a 55% increase). Among men and women with perfect Part B scores in a male-typed category (5 out of 5), women are 14 percentage points more likely to have a missed opportunity ($p<0.01$, a 93% increase).

Insert Figure II

In Table **??** in the Online Appendix, we use regression analysis to show that these trends are robust to the inclusion of additional controls, breaking out the data by category and presenting additional results by the maleness of the category. We again observe that men in the male-typed categories are significantly less likely to have a missed opportunity than their female counterparts in our sample. The regression specification also reveals a modest, but significant gender difference in the female-typed categories: men are estimated to be 5 percentage points more likely to have a missed opportunity in Arts and Literature ($p < 0.10$) and 4 percentage points more likely to have a missed opportunity in Pop Culture ($p < 0.05$). We also show that women are significantly more likely to have a missed opportunity as the maleness of the category increases ($p<0.01$); men are directionally less likely to have a missed opportunity as the maleness of the category increases,

though the trend is not significant.

Our results suggest that individuals who know the right answer to a question are reluctant to submit that answer in a gender incongruent category. This is particularly true for women. A woman's correct answer in a male-typed category is not submitted for the group more than a third of the time. Perhaps most surprisingly, even women who did very well in a male-typed category in Part B are reluctant to submit their answers relative to their male counterparts. Importantly, this is the case despite the fact there is no discrimination from fellow group members. In our setting, female experts are significantly less influential than male experts in male-typed domains simply because they do not contribute as often.

Outside of the laboratory, there may be an additional negative consequence of the most-informed group members not contributing answers as often as they should. Ex post, the group may fail to recognize who the most informed members are, making it more difficult to award promotions, raises, and leadership positions properly. While this question is beyond the realm of the current experiment, we can explore how well our groups are able to recognize the most knowledgeable member of their group. Following Part F (recall Figure I from the design section), participants are asked to name which member knew the most about each category. At this point, participants have seen the group answers from Part C, making them aware of how often their partner contributed and which answers were submitted for the group. Thus, we argue that an individual's perception of her partner at this stage is determined by the contribution decisions her partner made. In the analysis below, we ask how likely it is that an individual is recognized by her partner as being most knowledgeable (i.e. the individual contributed often enough to convey her expertise).

In Table VIII, we report the coefficients on gender of a probit regression predicting whether or not an individual was named most knowledgeable in a category, comparing male-typed and female-typed categories. We use the full sample (we do not condition on actually being most knowledgeable). We control for each group member's scores in the category as well as our standard set of controls.[21] Women are significantly more likely than men to be named most knowledgeable in female-typed categories, but significantly less likely to be named most knowledgeable in the male-typed categories. In the last column, we pool the categories and add a variable measuring the maleness of the category. We see that men are significantly more likely to be recognized as most

[21] We also control for the gender of one's partner, in case men and women view their partners differently.

knowledgeable as the maleness of the category increases, while women are significantly less likely to be recognized as most knowledgeable as the maleness of the category increases. These differential rates of recognition as experts are another important consequence of the patterns of willingness to contribute that we observe.

Insert Table VIII

In the Online Appendix, we provide the analysis separately for each category and we include the following addition. We report the coefficient on being in a public treatment, where participants received photographs of their partners in Part F, and the coefficient on the interaction between public treatment and gender. Recall that at this stage in the experiment, when participants are asked who they thought was most knowledgeable, participants in the public treatments have seen the group answers and the photographs of their partners. In principle, some of the gender differences we observe could be driven directly by stereotyping - if I know that my partner is a woman, I may be more likely to rank her as best in Arts and Literature (or less likely to rank her as best in Sports and Games). Our results in Table **??** show that these gender gaps are not driven by this type of bias. The coefficients on gender in the public and the private treatments are very similar, suggesting that the gender gaps in recoginition are driven by individuals speaking up less in gender incongruent categories, not discrimination by the other group member.

3.5 Lean In: Group Performance Improves if Women Contribute More Answers

We have shown that individuals are less likely to contribute in gender incongruent domains. The largest source of inefficiency stems from under-contribution by women in male-typed domains. Here, we explore a simple thought experiment. What would happen to group performance if we moved women ahead in line? We consider three possible policies: (1) move women ahead in all categories, (2) move women ahead in male-typed categories, and, for comparison, (3) move women ahead in female-typed categories. For each of these three policies, we shift all women one position closer to the front of the line. Women who selected "1" in our observed data remain at "1" to respect the constraints of the mechanism. We explore the impact that each of these policies has on the frequency of missed opportunities. Recall that a missed opportunity is a defined as a situation in

which an individual answers a question i correctly in Part C but does not contribute it. These policies will obviously eliminate at least some under-contribution by women. Of course, these gains will also come at a cost: newly contributed answers will crowd out some correct answers from men. The question is whether the benefits outweigh the costs.

We find that policies that move women ahead in male-typed categories significantly improve group performance. Consider policy (1), in which all women are moved one spot ahead in all categories. Under this policy shift, the overall frequency of missed opportunities among groups that contain at least one woman (the only groups that are potentially impacted by the change) falls from 23.8% to 21.2% (p-value from test of proportions = 0.002). The primary source of this reduction is the increased contribution from women in male-typed categories. This becomes clear if we consider policies (2) and (3). Under policy (2), women are moved ahead only in male-typed categories. Under this shift, the overall rate of missed opportunities falls from 23.8% to 21.7% (p-value = 0.016); there is no shift in the female-typed categories (as we have not implemented any changes here), while the rate of missed opportunities in male-typed categories falls from 26.8% to 23.6% (p-value = 0.004).[22] If we instead used policy (3), which does not take advantage of the fact that the primary source of inefficiency is under-contribution in the gender incongruent categories, there is no significant improvement in performance.

It is worth noting that the "lean in" style policies explored above are not targeted or sophisticated interventions. We are not encouraging contributions only from talented women or only from those that lack the most confidence. Instead, we just move all women ahead in male-typed categories, and still, this unnuanced intervention leads to an improvement in performance. Importantly, we present these results only as an illustration of the potential gains to be had from increased contribution by women in our data set. Before making a broader argument for any type of policy change, more research would be needed on the equilibrium effects and welfare consequences of encouraging contributions in gender incongruent categories.

[22] Perhaps a valuable comparison would be exploring what would happen if we instead moved men ahead in the male-typed categories. Under this policy change, the rate of missed opportunities in male-typed categories for groups that have at least one man would increase from 25.9% to 26.3% (n.s.).

4 Discussion

Decisions are often made in groups, with the hope that bringing together individuals with varied backgrounds and expertise may enable informed decision-making in a variety of domains. However, even groups composed of experts can falter, particularly if those experts are reluctant to contribute their ideas to the group. In our experiment, we find that individuals often fail to provide their knowledge to a group, particularly when they perceive the problem at hand as being outside of their gender's domain. Men who know a lot about Arts and Literature or Entertainment and Pop Culture do not contribute answers as often as similarly knowledgeable women, and women who know a lot about Environmental Science, History, Geography, or Sports and Games contribute much less than similarly knowledgeable men. This type of under-contribution has two negative consequences: (1) groups miss out on correct answers from individuals in gender incongruent categories, and (2) groups have a hard time recognizing who their experts are.

In our setting, individuals who are told that they are talented in gender incongruent fields do not react to this information; that is, they contribute no more often than similarly-talented individuals who do not receive this information. In future research, it would be helpful to gather data on how robust this result is to stronger or more precise signals about ability. At the very least, our evidence suggests that it may be harder to convince an individual of her talent (and, as a result, harder to convince an individual to act on this talent) in an area that does not conform with gender stereotypes.

Our results are consistent with several of the theoretical predictions put forth by Bordalo, Gennaioli, and Shleifer (2014) in their work on self-stereotyping (henceforth, BGS). BGS present a model in which individuals form stereotypes based upon the most representative features of a group. Applied to gender, their model predicts that even in areas in which there is large overlap in the distributions of ability of men and women, stereotypes that exaggerate these differences can take hold. Importantly, their model yields the prediction that women will be less likely to participate in areas that are perceived as male-typed, but that this gender gap will disappear in female-typed domains: we find exactly this pattern in our data on willingness to contribute.[23] Furthermore, they

[23] However, we should note that the BGS model predicts the strongest effects for "average individuals" - that is, the effects are predicted to be driven by average women who underestimate their ability in the male-typed domains (BGS expect minimal gender gaps at the high and low end of the ability spectrum). We do not find evidence in support this hypothesis. See Figure II for a clear depiction of this: we observe gender gaps in willingness to contribute in

argue that beliefs are the channel through which stereotypes impact behavior, as stereotypes color self-assessments. This certainly seems consistent with the data presented in Table V: given the same Part B score, individuals are much less likely to believe they had the highest Part B score in their group when the category is gender incongruent, and these beliefs predict their contribution decisions.

We can only speculate as to what additional factors may contribute to the gender differences we observe, as the experiment was not designed to test other stories. That said, it seems valuable to think about what other theories might speak to our results. In her work on social role theory, Eagly (1987) explains that gender roles are not only positive (descriptive of our perceptions of men and women), but also normative (proscriptive of how men and women should behave). With this in mind, it seems plausible that the utility an individual derives from contributing to a group may depend on whether the domain is gender congruent. An individual may prefer to contribute in a gender congruent area because it is more consistent with her own and others' expectations about how she should behave. This could potentially explain why even controlling for beliefs about own ability, women are less likely to contribute in stereotypically male domains. To better understand this channel, it would be useful to conduct additional research in which we exogenously manipulate the salience of an individual's gender identity or the gender stereotype of the category.

While the types of questions that groups in our experiment face might seem somewhat frivolous, the patterns we observe in how groups answer them have implications for economically-important settings and outcomes. Consider the workplace. If a woman with expertise in a male-dominated field is more reluctant to contribute ideas in meetings or team projects, not only will her colleagues benefit less from her expertise, but her employer may also have more difficulty recognizing her talent. It seems possible that this could impede her ability to earn leadership positions or promotions – even when there is no discrimination by her employer. It may be true that women and men who ultimately self-select into these areas will be less prone to this type of under-contribution. However, even in that case, understanding how under-contribution might impact the pipeline remains important. Access and entry may depend upon how often individuals contribute their ideas in earlier stages of their education and careers.

Author affiliation: Katherine Baldiga Coffman, The Ohio State University

male-typed domains for women of all abilities.

References

Arrow, Kenneth, "The Theory of Discrimination," in O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (Princeton, NJ: Princeton University Press 1973), 3 - 33.

Baldiga, Katherine, "Gender Differences in Willingness to Guess," *Management Science*, forthcoming, 2013.

Babcock, Linda and Sara Laschever, *Women Don't Ask: The High Cost of Avoiding Negotiation - and Positive Strategies for Change* (New York, NY: Bantam Books 2007).

Ben-Shakhar, Gershon and Yakov Sinai, "Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies," *The Journal of Educational Measurement*, 28 (1991), 23-35.

Benjamin, Daniel, James Choi, and A. Joshua Strickland, "Social Identity and Preferences," *American Economic Review*, 100 (2010), 1913–1928.

Bertrand, Marianne, Claudia Goldin, and Lawrence Katz, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics,* 2 (2010), 228 - 255.

Beyer, Sylvia, "Gender Differences in the Accuracy of Self-evaluations of Performance," *Journal of Personality and Social Psychology*, 59 (1990), 960-970.

Beyer, Sylvia and Edward Bowden, "Gender Differences in Self-perceptions: Convergent Evidence from Three Measures of Accuracy and Bias," *Personality and Social Psychology Bulletin*, 23 (1997), 157-172.

Beyer, Sylvia, "Gender Differences in Self-perception and Negative Recall Biases," *Sex Roles,* 38 (1998), 103-133.

Bohnet, Iris, Alexandra Van Geen, and Max Bazerman, "When Performance Trumps Gender Bias: Joint Versus Separate Evaluation," Working Paper, 2012.

Bordalo, Pedro, Niccola Gennaioli, and Andrei Shleifer, "Stereotypes," Working paper, 2014.

Burns, Justine, Simon Halliday, and Malcolm Keswell, "Gender and Risk Taking in the Classroom," Working paper, 2012.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek, "Gender, Competitiveness and Career Choices," *Quarterly Journal of Economics*, forthcoming, 2014.

Cooper, David and John Kagel, "A Failure to Communicate: An Experimental Investigation of the Effects of Advice on Strategic Play," Working paper, 2012.

Crosby, Faye and Linda Nyquist, "The Female Register: An Empirical Study of Lakoff's Hypothesis," *Language in Society*, 6 (1997), 313-322.

Croson, Rachel and Uri Gneezy, "Gender Differences in Preferences," *The Journal of Economic Literature*, 47 (2009), 448-474.

Dreber, Anna, Emma von Essen, and Eva Ranehill, "Outrunning the Gender Gap - Boys and Girls Compete Equally," *Experimental Economics*,14 (2011), 567 - 582.

Eagly, Alice, *Sex Differences in Social Behavior: A Social-role Interpretation* (Hillsdale, NJ: Erlbaum 1987).

Eagly, Alice and Antonio Mladinic, "Gender Stereotypes and Attitudes toward Women and Men," *Personality and Social Psychology Bulletin,* 15 (1989), 543-558.

Eckel, Catherine and Philip Grossman, "Men, Women, and Risk Aversion: Experimental Evidence," *The Handbook of Experimental Economics Results*, 1 (2008), 1061-1073.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, "Performance in Competitive Environments: Gender Differences," *The Quarterly Journal of Economics,* 118 (2003), 1049 - 1074.

Goldin, Claudia, "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, 104 (2014), 1 - 30.

Grosse, Niels and Gerhard Riender, "Explaining Gender Differences in Competitiveness: Gender-task Stereotypes, Working paper, 2010.

Heilman, Madeline, "Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder," *Journal of Social Issues,* 57 (2010), 657 - 674.

Hong, Kessely, "The Role of Personal Experience in Overcoming Gender Stereotypes," Unpublished manuscript.

Kamas, Linda. and Anne Preston, "Are Women Really Less Competitive than Men?," Working paper, 2010.

Karni, Edi. "A Mechanism for Eliciting Probabilities," *Econometrica*, 77 (2009), 603 – 606.

Krawczyk, Michal, "Framing in the Field: A Simple Experiment on the Reflection Effect," Working paper, 2011.

Milkman, Katherine., Modupe Akinola, Dolly Chugh, "Discrimination in the Academy: A Field

Experiment," Working paper, 2014.

Mobius, Markus, Paul Niehaus, Muriel Niederle, and Tanya Rosenblat,"Managing Self-Confidence: Theory and Experimental Evidence," Working paper, 2014.

Mondak, Jeffrey and Mary Anderson, "The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge," *The Journal of Politics*, May (2004), 492-512.

Moore, Don and Paul Healy, "The Trouble with Overconfidence," *Psychological Review,* 115 (2008).

Niederle, Muriel and Lise Vesterlund, "Do Women Shy away from Competition? Do Men Compete too Much?," *The Quarterly Journal of Economics*, 122 (2007), 1067-1101.

Norton, Edward, Hua Wang, and Chunrong Ai, "Computing Interaction Effects and Standard Errors in Logit and Probit Models," *The Stata Journal*, 4 (2004), 154-167.

Phelps, Edmund, "The Statistical Theory of Racism and Sexism," *The American Economic Review*, 62 (1972), 659 - 661.

Sandberg, Sheryl, *Lean In: Women, Work, and the Will to Lead,* (New York, NY: Random House, 2013).

Shurchkov, Olga, "Under Pressure: Gender Differences in Output Quality and Quantity Under Competition and Time Constraints," *Journal of European Economic Association*, 10 (2012), 1189 - 1213.

Tannenbaum, Daniel, Do Gender Differences in Risk Aversion Explain the Gender Gap in SAT Scores? Uncovering Risk Attitudes and the Test Score Gap," Working paper, 2012.

Thomas-Hunt, Melissa and Katherine Phillips, "When What You Know is Not Enough: Expertise and Gender Dynamics in Task Groups," *Personality and Social Psychology Bulletin,* 30 (2004), 1585 - 1598.

Xie, Yu and Kimberlee Shauman, *Women in Science: Career Processes and Outcomes*, (Cambridge, MA: Harvard University Press, 2003).

Tables

Table I: Summary Statistics

	Men	Women	Total	p value [H_0: M=W]
White	75.1%	57.5%	68.3%	<0.01
East Asian	12.8%	22.3%	16.5%	<0.01
Black or African American	4.0%	7.8%	5.4%	0.07
Asian Indian	3.2%	4.5%	3.7%	0.48
Attended High School in United States	91.5%	83.2%	88.3%	<0.01
Current Ohio State Undergraduate	90.4%	95.0%	92.2%	0.08
Avg. # of Correct Part B Answers	13.6	12.6	13.2	<0.01
Arts and Literature (Art)	2.26	2.45	2.34	0.10
Pop Culture (Pop)	1.02	1.15	1.07	0.15
Environmental Science (Env)	4.05	3.78	3.95	<0.01
History (Hist)	1.64	1.17	1.46	<0.01
Geography (Geo)	1.93	1.88	1.91	0.64
Sports and Games (Sports)	2.70	2.13	2.49	<0.01
Avg. # of Correct Part C Answers	15.8	14.3	15.2	<0.01
Arts and Literature	2.06	2.25	2.13	0.05
Pop Culture	2.97	2.87	2.93	0.34
Environmental Science	2.96	2.52	2.79	<0.01
History	2.25	2.07	2.18	0.10
Geography	2.44	2.24	2.36	0.05
Sports and Games	3.14	2.31	2.82	<0.01
Totals	281	179	460	

Notes: p-values are from tests of proportions for binary variables and Fisher-Pitman permutation tests for non-binary variables, with a null of equality of distributions between men and women

Table II: Perceived Genderedness of Categories

	Avg. Maleness Given by Men	Avg. Maleness Given by Women	Overall Average	Normalized z Score
Art	-.317	-.419	-.356	-1.18
Pop	-.263	-.348	-.297	-1.01
Env	.142	.057	.109	0.13
Hist	.196	.061	.144	0.23
Geo	.215	.065	.157	0.27
Sports	.643	.571	.615	1.56

Note: Elicited on [-1,1] scale where -1 is labeled "Women know more", 1 is labeled "Men know more", 0 is "no gender difference"

Table III: Average Place in Line by Category and Gender

Average Place in Line

Lower Positions in Line Indicate Greater Willingness to Contribute

	Art	Pop	Env	Hist	Geo	Sports	Pooled
Men	2.60	2.24	2.11	2.22	1.96	2.11	2.21
	[0.62]	[0.58]	[0.71]	[0.71]	[0.65]	[0.68]	[0.51]
Women	2.36	2.20	2.39	2.48	2.31	2.69	2.40
	[0.66]	[0.60]	[0.67]	[0.71]	[0.71]	[0.67]	[0.51]
p value	<0.01	0.42	<0.01	<0.01	<0.01	<0.01	<0.01
Observations	460	460	460	460	460	460	460

Unit of observation is a mean of an individual's five places in line within each category or across all categories in the pooled column. P values derived from Fisher-Pitman permutation test for two independent samples, testing the null of equality of the two distributions using Monte Carlo method with 200,000 simulations. Standard deviations of mean place in line shown in [ ]

Table IV: Willingness to Contribute

OLS Predicting Position in Line for Question i in Part C

Lower Positions in Line Indicate Greater Willingness to Contribute

Category	Art	Pop	Env	Hist	Geo	Sports	Pooled
Maleness z score	-1.18	-1.01	0.13	0.23	0.27	1.56	
Female	-0.226****	-0.090*	0.217****	0.145**	0.296****	0.383****	0.145***
Dummy	(0.065)	(0.048)	(0.067)	(0.071)	(0.065)	(0.058)	(0.048)
Maleness							-0.120****
of Category							(0.013)
Female x							0.265****
Maleness							(0.019)
Answered Qn. i	-0.480****	-0.977****	-0.683****	-0.415****	-0.553****	-1.003****	-0.706****
Correctly	(0.053)	(0.058)	(0.040)	(0.046)	(0.048)	(0.046)	(0.023)
Part B Score	-0.081****	-0.016	-0.103***	-0.119****	-0.081***	-0.067**	-0.024***
in Category	(0.025)	(0.025)	(0.034)	(0.026)	(0.031)	(0.031)	(0.009)
Constant	3.53****	4.27****	2.26****	3.42****	2.80****	3.78****	3.33****
	(0.226)	(0.175)	(0.294)	(0.273)	(0.252)	(0.255)	(0.193)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	460	460	460	460	460	460	460
Obs.	2299	2300	2300	2298	2298	2298	13793
R^2	0.268	0.548	0.160	0.170	0.166	0.322	0.241

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level.

Maleness z score is the z score of the average perception of the gender-type of the category elicited from participants.

The unit of observation is individual j's answer to a Part C question i, Std. errors clustered at individual level.

Controls are treatment dummies, race dummies, session size, gender composition of session, U.S. H.S. dummy,

OSU undergrad dummy, and overall probability of a correct answer in our sample for that particular Question i from Part C.

Table V: Predicting Participant Beliefs about Part B Performance

Probit Predicting Pr(Guessed She Had Highest Part B Score in Group)							
Category	Art	Pop	Env	Hist	Geo	Sports	Pooled
Maleness z	-1.18	-1.01	0.13	0.23	0.27	1.56	
Female	0.211****	0.207****	-0.021	-0.132***	-0.254****	-0.416****	-0.088****
Dummy	(0.052)	(0.051)	(0.051)	(0.051)	(0.050)	(0.045)	(0.023)
Maleness							0.121****
of Category							(0.015)
Female x							-0.249****
Maleness							(0.021)
Part B Score	0.123****	0.075***	0.100****	0.125****	0.089****	0.125****	0.092****
in Category	(0.023)	(0.026)	(0.026)	(0.022)	(0.025)	(0.028)	(0.008)
Constant	0.403****	0.389****	0.614****	0.390****	0.511****	0.438****	0.457****
	(0.021)	(0.022)	(0.022)	(0.021)	(0.022)	(0.020)	(0.009)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	460	460	460	460	460	460	460
Observations	460	460	460	460	460	460	2760
Pseudo R^2	0.114	0.071	0.054	0.121	0.067	0.231	0.113

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level.

Interaction corrected using Norton et al (2004).

Controls are treatment dummies, race dummies, session size, gender composition of session, U.S. high school dummy, OSU undergrad dummy, Std. errors clustered at subject level for pooled specification; marginal effects reported

Table VI: The Role of Confidence in Predicting Willingness to Contribute

OLS Predicting Position in Line for Question i in Part C

Lower Positions in Line Indicate Greater Willingness to Contribute

Category	Art	Pop	Env	Hist	Geo	Sports	Pooled
Maleness z	-1.18	-1.01	0.13	0.23	0.27	1.56	
Female	-0.011	-0.023	0.144***	0.104*	0.090*	0.099**	0.079*
Dummy	(0.054)	(0.040)	(0.052)	(0.056)	(0.048)	(0.051)	(0.040)
Maleness							-0.036****
of Category							(0.010)
Female x							0.071****
Maleness							(0.015)
Qn. i	-0.171****	-0.232****	-0.243****	-0.147****	-0.166****	-0.244****	-0.193****
Correct	(0.041)	(0.045)	(0.035)	(0.037)	(0.036)	(0.040)	(0.016)
Part B	-0.011	0.023	-0.023	-0.029	-0.007	0.016	0.010
Score	(0.022)	(0.020)	(0.027)	(0.022)	(0.023)	(0.025)	(0.007)
Pr(Qn. i	-0.026****	-0.026****	-0.025****	-0.025****	-0.026****	-0.027****	-0.027****
Correct)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Pr(Partner	0.003*	0.002**	0.005****	0.003***	0.002**	0.005****	0.003****
Qn. i	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Correct)							
Ranked	-0.106**	-0.134****	-0.061	-0.130**	-0.147****	-0.198****	-0.126****
Self First	(0.052)	(0.038)	(0.052)	(0.054)	(0.044)	(0.050)	(0.021)
Constant	4.24****	4.35****	3.68****	4.18****	4.12****	3.93****	4.14****
	(0.192)	(0.166)	(0.225)	(0.219)	(0.170)	(0.202)	(0.157)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	460	460	460	460	460	460	460
Observations	2299	2300	2300	2298	2296	2298	13791
R^2	0.578	0.744	0.495	0.476	0.558	0.632	0.593

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level.

The unit of observation is individual j's answer to a Part C question i. Std. errors clustered at individual level.

Controls are risk preferences, treatment dummies, race dummies, session size, gender composition of session, U.S. H.S. dummy, OSU undergrad dummy, and overall probability of a correct answer in our sample for that particular Question i from Part C.

Table VII: The Impact of Treatments on Position in Line

OLS Predicting Position in Line for Question i in Part C for Group Member with Best Part B Score in Category

Lower Positions in Line Indicate Greater Willingness to Contribute

	Female-Typed Categories	Male-Typed Categories	Pooled
Female Dummy	-0.112	0.361****	0.224***
	(0.091)	(0.092)	(0.082)
Maleness of Category			-0.094****
			(0.024)
Female x Maleness			0.248****
			(0.043)
Feedback Treatment	-0.079	-0.124*	-0.089
	(0.068)	(0.065)	(0.058)
Public Treatment	0.068	0.057	0.073
	(0.067)	(0.062)	(0.057)
Female x Feedback	0.019	-0.020	-0.027
	(0.113)	(0.106)	(0.096)
Female x Feedback			-0.038
x Maleness			(0.049)
Female x Public	-0.132	-0.095	-0.120
	(0.110)	(0.104)	(0.095)
Female x Public			0.081
x Maleness			(0.050)
Constant	3.574****	2.807****	3.157****
	(0.217)	(0.247)	(0.214)
Controls	Yes	Yes	Yes
Clusters	391	452	459
Observations	2914	5845	8759
R^2	0.395	0.168	0.242

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level.

Female-typed categories are Art and Pop; male-typed categories are Env, Hist, Geo, and Sports. Controls are a dummy for answering Question i correctly, Part B score, race dummies, session size, gender composition of session, U.S. H.S. dummy, OSU undergrad dummy; we also include Public x Maleness and Feedback x Maleness for completeness; marginal effects reported

Table VIII: Predicting Being Ranked Most Knowledgeable by Partner

Probit Predicting Pr(Ranked Most Knowledgeable by Partner after Part F)			
	Female-Typed Categories	Male-Typed Categories	Pooled
Female Dummy	0.181****	-0.088***	-0.012
	(0.042)	(0.030)	(0.025)
Maleness of Category			0.035***
			(0.013)
Female x Maleness			-0.103****
			(0.021)
Constant	0.481****	0.442****	0.455****
	(0.015)	(0.011)	(0.009)
Controls	Yes	Yes	Yes
Clusters	230	230	230
Observations	920	1840	2760
Pseudo R^2	0.123	0.100	0.088

Notes: * indicates significance at the 10% level, ** at the 5% level, *** at the 1% level, and **** at the 0.1% level; Controls are both members' Part B and C scores, treatment dummies, race dummies, session size, gender composition of session, U.S. H.S. dummy, OSU undergrad dummy, and partner's gender; Std. errors clustered at group level; marginal effects reported. Interaction corrected using Norton et al (2004). Female-typed categories are Art and Pop; male-typed categories are Env, Hist, Geo, and Sports.

Figures

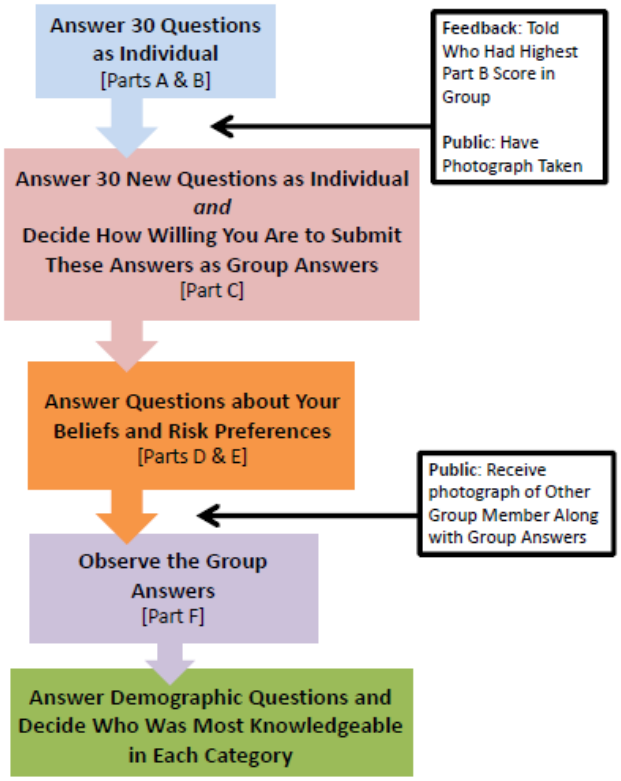

Figure I: The flow-chart demonstrates what a participant does in each part of the experiment and when the treatment interventions take place. Our central focus is the data from Part C, in which participants made decisions about how willing they were to submit their answers for the group. We will use the data from Part B to control for individual knowledge and data from Part D to control for confidence.

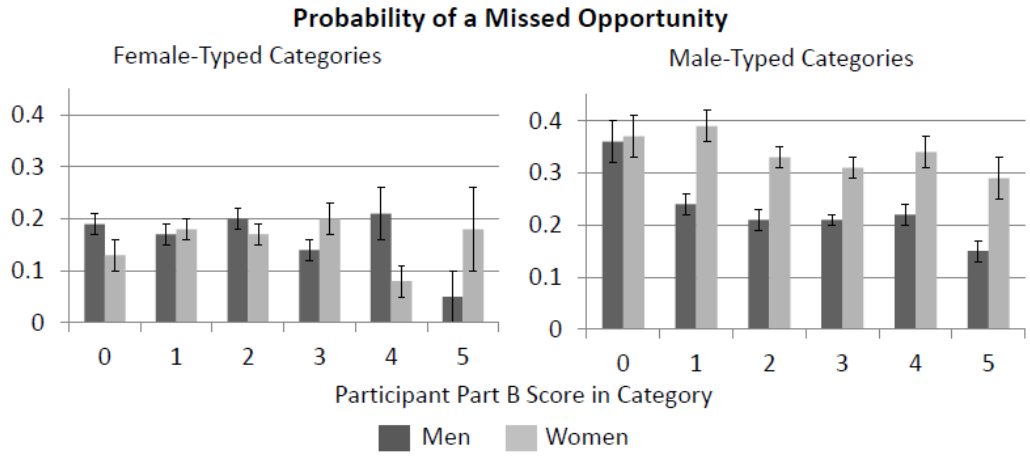

Figure II: The Probability of a Missed Opportunity, Split by Gender and Gender-Type of Category

Notes: A missed opportunity occurs when an individual answers a question correctly in Part C but submits a place in line strictly greater than her partner