Don't tell anyone I lost to a girl! Gender stereotypes and hiding low performance

Shuya He* Charles N. Noussair†

August 18, 2020

Abstract

It has been asserted that males incur a psychological cost when they are outperformed by a female competitor. We conduct a laboratory experiment that allows us to measure this cost, for performance in a mathematical task. The experiment is conducted in both the US and China. We find that in our Chinese sample, males are willing to pay more to hide the fact that they have performed worse than another individual than women are, while there is no gender difference in the US. In China, females are willing to pay more to hide poor performance when losing to another female than to a male. In the US, the opposite pattern is observed; women have a greater cost of revealing that they have lost to a man than to another woman. The gender of the counterpart is not a determinant of males' willingness to hide poor performance. An incentivized questionnaire reveals that a stereotype that males would outperform females exists in the Chinese sample, but not among our American participants.

JEL classification: C91, D13, D91, J16, Z13

Keywords: Gender stereotype, Cross-cultural, Stigma

*Department of Economics, The University of Arizona, 1130 E. Helen St., Tucson AZ, 85721. Email: shuyahe@email.arizona.edu

†Department of Economics and Economic Science Laboratory, The University of Arizona, 1130 E. Helen St., Tucson AZ, 85721. Email: cnoussair@email.arizona.edu

‡We thank the Economic Science Laboratory of the University of Arizona for financial support and the Behavioral Economics Laboratory at the University of Electronic Science and Technology for permitting us to run sessions there. We are grateful to Mitch Addler, Xiduo Chen, Martin Dufwenberg, Nickolas Gagnon, Senran Lin, Takeshi Murooka, Ronald L. Oaxaca, Juan Pantano, Liang Qiao, Elaine Park Rhee, Julian Romero, Evan J. Taylor, Xiaoyuan Wang, Bohan Ye, and Xiaojian Zhao for their helpful comments.

1 Introduction

It is widely claimed in the popular press that some men view it with disfavor when a female partner, co-worker, or peer is more successful or more able than they are. This disutility can express itself in a number of ways, for example in a preference for female partners who are less successful and intelligent than the man is, or in a reluctance to be employed in a subordinate position to a female[1]. A number of studies in psychology and economics support this notion.

For example, Ratliff and Oishi (2013) find that men have lower self-esteem when outperformed by a female romantic partner on either an academic or a social task, while women do not exhibit a similar effect when outperformed by male partners. Park et al. (2015) find that men indicate that they prefer women who are more intelligent than they are as partners when the potential partner is psychologically distant, but that they reverse their preference when the decision becomes proximate. Karbowski et al. (2016) report that the women that are most attractive to men are those who rate a 7 out of 10 on intelligence. Fisman et al. (2006) find, in a speed dating experiment, that men do not value women's intelligence or ambition when it exceeds their own, while women do value intelligence on the part of men. Park et al. (2011) observe that women subject to romantic priming exhibit less of a preference for STEM topics than they do otherwise. The authors conjecture that women may distance themselves from STEM when the goal is to be romantically desirable because they are aware of males' preference. Syrda (2020), studying a large sample of over 6000 households, observes that male psychological distress reaches a minimum at a point where wives earn 40% of total household income and reaches a maximum when men are entirely economically dependent on their wives. Husain et al. (2018) find that male teachers are more likely to quit their position if the school principal is female than male, while there is no difference for female teachers.

The dominant explanation for this behavior on the part of males is that it stems from cultural beliefs about gender, with their specific expectations of gender roles.

[1] See for example, Burriss (May 31, 2016), Barth (Jan 23, 2016), or Fottrell (Dec 7, 2018).

There exists a belief in many cultures that men ought to be more successful and earn more than women, and the strength of this belief varies among cultures (Van de Vijver (2007); Hofstede (1998); Hofstede (2001)). Such beliefs shape people's sense of what others expect of them and, in turn, their behavior and judgments (Ridgeway (1997); Foschi (2000); Correll (2001)). As a consequence, it may be socially costly for a man to have others know that he has performed poorly at a task when society's expectations are that he would perform well, or to be in an inferior position when society expects otherwise. Such stigmas are problematic for gender equality (Ridgeway and Correll (2000)), as well as for economic efficiency. For example, if men find it socially costly to be outperformed by women professionally, they might seek to prevent more capable women from gaining equal or superior positions to theirs within an organization, or to hide or downplay strong performance by women.

In this paper, we report the results of an experiment in which we measure the cost of being outperformed by a peer. The setup is designed to be extendable to other settings and populations. We elicit a willingness-to-accept to have one's inferior performance against a competitor on a cognitive task made public. We compare the behavior of women and men, and further contrast their behavior based on the gender of the competitor. This allows us to consider whether the social cost of being outperformed is greater for men than for women, and then to ask whether the cost depends on the gender of the person who is outperforming the other.

Specifically, in the main portion of our experiment, participants are asked to perform a series of additions of five randomly-chosen, two-digit numbers in a four-minute period. Their performance is measured as the number of the correct sums calculated, and this score is compared with that of a randomly matched partner. In some trials, the partner is of the other gender, and in others, the partner is of the same gender. The member of the pair with more correct totals is designated as the *Better performer*, and the other person is dubbed the *Worse performer*. The Worse Performer is eligible to claim a monetary payment as a consolation prize. However, to claim it, he must make his Worse Performer status publicly known. Therefore, participants in our experiment face a trade off between obtaining extra monetary

compensation and being able to hide their relatively inferior performance. Thus, we elicit participants' monetary values for hiding their poor performance. This value can be interpreted as the cost of the stigma of performing worse than the competitor. We are able to measure how this cost varies by own gender and the gender of the competitor.

The experiment is conducted in both China and the U.S., with participant pools of similar profiles. The experiment is not designed to directly compare effect sizes in the two samples, but we do compare the qualitative patterns between them. This cross-cultural approach allows us to investigate how the cultural background influences individuals' stigma costs. The two countries have very different cultures, histories, and societal norms. As we report later, we also obtain evidence that different stereotypes about the relative performance of women and men on our task are present in the two groups. We recognize, of course, that any differences between our two samples could be attributed to factors other than cultural background that might differ between our samples, or that cultural differences interact with other demographic factors so that the same patterns would not appear with other paired samples. Nevertheless, we believe that our results make a contribution to understanding differences between gender stereotypes and interaction between the two countries.

The hypotheses for the experiment emerge from a simple theoretical model, which is presented in Section 4. The model assumes that players are competing in a task in which males are believed to be better than females on average, though it could be readily translated to assume the opposite stereotype. The model predicts that more females than males would claim a given consolation prize when they are competing with a partner of a different gender. It also predicts that females' are less likely to claim the consolation prize if their performance is compared with another female than with a male. In contrast, males are less likely to claim the prize when facing a female than a male partner.

Our data show very different patterns in two countries. In China, females are more willing to accept the consolation prize than men, and are more willing to do

so when competing with a partner of a different, than of the same, gender. In the US, there is no overall gender difference in the likelihood of accepting the prize, and females are *less* likely to accept the consolation prize when paired with a male than a female. Questionnaire data shows that members of our two samples hold different stereotypes. The belief that men would outperform women on average in the mathematics addition task is held by a majority in China, though not in the US. Thus, the results are generally consistent with our model, as is discussed in Section 5.

Two remarks are in order at this point. The first is that by measuring the willingness-to-pay to hide one's Worse Performer status, we are measuring the cost of having peers find out about one's poor performance, beyond the cost of merely learning yourself that you did not perform well. This difference seems to us to be the most relevant effect to measure. In dating and employment relationships, the relative status of the two parties (profession, position in company hierarchy, educational level), is typically known to others. One typically also is aware of one's own ability before entering the relationship. Therefore, beginning a relationship implies making one's ability or status public rather than private. For this reason, we chose to study the cost of making one's status public. The second remark is that we do not try to *simulate* an employment or a romantic relationship in our experiment. Our pairings are anonymous and fleeting. This means that any cost of inferior performance in a field setting such as those described at the beginning of this section would likely be much greater than those we observe, and our experiment should therefore be viewed as a minimal paradigm to observe gender differences. In our view, this makes the gender and cultural differences that we do observe all the more striking.

The remainder of the paper is organized as follows. In the next section, we briefly discuss related literature. In Section 3, we describe the experimental design and procedures. A simple theoretical framework and hypotheses are presented in Section 4. The results are analyzed and discussed in Section 5. We close, in Section 6, with a brief summary of our findings and some concluding thoughts.

2 Related literature

There are a number of lines of literature in economics that relate to ours. The first is a very large literature in experimental economics documenting differences in the economic behavior of females and males. Gender differences in preferences for risk (Eckel and Grossman (2008), Croson and Gneezy (2009)), competition (Niederle and Vesterlund (2007), Gneezy and Rustichini (2004)), in bargaining behavior (Babcock et al. (2003)), and in other-regarding preferences (Eckel and Grossman (2001); Andreoni and Vesterlund (2001); Eckel and Grossman (1998)) such as altruism, fairness or envy are all well-documented. .

A smaller strand of studies focuses on how decisions are affected by the gender of the counterpart in a strategic interaction. Most of this work concerns individuals' performance and willingness to compete. Babcock et al. (2017) manipulate the gender composition of groups and find that individuals' behavior in volunteering to work on a low-promotability task depends on the gender of the other group members. Men and women are equally likely to volunteer when all group members are of the same gender, while women are more likely to volunteer in mixed-gender pairings. Gneezy et al. (2003) show that a woman's competitive performance is sensitive to the gender of her competitors. Women's performance does not increase in a competitive environment compared to a non-competitive one, while it does for men. These effects are more pronounced when a woman is competing against a man than against another woman. Mago and Razzolini (2019) design a best-of-five probabilistic contest and document that women exert significantly greater effort only when competing against other women, while for men the gender of the opponent is of no consequence.

There has been work on gender stereotypes in mathematics in both the US and in China. Some research suggests that in the US, a stereotype that men are better at math than women creates a gender disparity (Spencer et al. (1999), Aronson et al. (1999)). There is evidence that parents may shape their children's expectations, and in turn their performance, by communicating their beliefs about how girls and boys

should perform in math. For example, Tiedemann (2000) has found that parents with stronger stereotypical beliefs that boys are better at math than girls and find math more useful and more important than girls, had higher (lower) perceptions of the math ability of their sons (daughters). These parental beliefs were also correlated with the children's own beliefs about their math ability.

Such gender-based beliefs also exist in China. Several Chinese studies (Dong (2019), Tsui (2007)) in the field of education find that Chinese teachers have the view that even when they observe no difference in the mathematical performance of boys and girls in their classes, gender stereotypes about mathematical performance, such as "boys do better than girls in mathematics", are still present in the teachers' minds. They argue that gender stereotypes are present in Chinese textbooks, school environments, and teacher-student interactions.

In terms of actual performance, the literature does suggest that there are differences both in the relative performance of women and men in mathematics, as well as in performance stereotypes, among different countries. Liu and Wilson (2009) find that Hong Kong students showed a larger gender gap than U.S. students on the Program for International Student Assessment (PISA) 2003 mathematics test, with males scoring higher in both locations. In addition, several cross-national studies indicate that greater cultural inequities between males and females are associated with larger gaps in mathematical performance favoring males (Else-Quest et al. (2010)). Nations with higher proportions of women enrolled in post-secondary science courses and employed in scientific careers are less likely to explicitly endorse the stereotype that science is a masculine profession (Miller et al. (2015)). Females' gross enrollment ratio (GER) of tertiary education is about 56% in China, while the number is 102% in the US [2]. It thus appears that there is a stronger gender stereotype about mathematical ability in China than in the U.S.

Our study is related to a small but growing literature on differing social norms

[2]The data are from UNESCO, and GER can exceed 100% due to the inclusion of over-age and under-age students who began their studies early or late, and those who repeated one or more grades. A high GER generally indicates a high degree of participation, whether the pupils belong to the official age group or not. A GER value approaching or exceeding 100% indicates that a country is, in principle, able to accommodate all of its school-age population in the educational system.

regarding men and women and how they affect social and economic outcomes (e.g., Bursztyn et al. (2017)[3]. The norm that husbands should earn more than their wives, in particular, is now drawing more and more attention in explaining couples' behavior. First, studies show that couples actually do react to this norm. For example, a study using U.S Census Bureau data (Murray-Close and Heggeness (2018)) compares the earnings reported for husbands and wives in the CPS with their actual earnings from administrative income-tax records. They find that survey respondents inflate their reports of husbands' earnings and deflate the reports of wives' earnings. Bertrand et al. (2015) show that the distribution of the share of income earned by the wife exhibits a sharp drop to the right of 1/2, the point at which the wife's income exceeds the husband's.

Recently, Patnaik (2019) investigates the role of non-monetary "stigma costs" in explaining the difference in parental leave participation rates between the genders. By conducting a causal analysis of the Quebec Parental Insurance Program, in which fathers enjoy an individual and non-transferable right to parental leave, the study finds that fathers respond not only to the higher benefits but also to the "daddy-only" label itself. That is, even though the quota did not alter a binding constraint for many families, reserving some weeks as "daddy-only" appears to alter the distribution of leave toward fathers. The mechanism underlying the finding that the author proposes is that the "daddy-only" label helps to reduce the stigma costs that fathers face. These stigma costs include multiple components, such as peer pressure and workplace hostility against leave-takers, and the data do not allow the author to disentangle and quantify the effect of each factor.

Our research is also related to studies investigating the role of social image concerns on behavior. People care about how others perceive who and what kind of person they are. The psychological costs of being observed and judged can have important consequences for behavior (e.g. Masclet et al. (2003)). When an audi-

[3]Bursztyn et al. (2017) conduct a field experiment that includes a real-stakes placement questionnaire. They find that single female students report lower desired salaries, as well as lower willingness to travel and to work long hours, when their preferences are observable by classmates. Folke and Rickne (2020) document that job promotions increase the chance of divorce for women, but not for men. Bertrand (2011) provides a comprehensive review of economic studies of gender issues.

ence observes a person, they become more likely to exhibit qualities that are socially desirable, such as being fair (Andreoni and Bernheim (2009), smart (McManus and Rao (2015); Bursztyn et al. (2019)), and charitable (Grossman (2015)).

Another related line of research consists of experimental studies focused on gender stereotypes. Bordalo et al. (2019) investigate how gender stereotypes shape beliefs about one's own and others' ability in different categories of knowledge, using laboratory experiments. They find that when evaluating others, people tend to overestimate the performance of men in categories that are stereotypically viewed to play to male strengths, such as mathematics. Both experimental and field evidence have documented a widespread belief that women have lower ability than men in mathematics (Guiso et al. (2008); Carrell et al. (2010); Reuben et al. (2014)).

Evidence from the recent work of Cappelen et al. (2019) suggests that men suffer greater social cost than women when they perform poorly at a task. They find that the general public views male losers differently from female. People are more likely to infer that male losers, to a greater extent than female losers, have exerted low effort and therefore consider them less deserving of public assistance. Their findings imply that being less competent is especially socially undesirable for males. Their results suggest that males may anticipate these social sanctions and thus would have a higher willingness to pay to hide a low level of competence.

3 Experimental Design

3.1 General Procedures

The experiment was conducted in late 2019 and early 2020 at two universities in two different countries: China and the United States. We ran 10 sessions at the Behavioral Economics Laboratory at the School of Economics and Management at the University of Electronic Science and Technology of China[4], located in Chengdu, China. We also conducted eleven sessions, with a total of 120 participants, at the

[4]The University of Electronic Science and Technology of China is a multidisciplinary university, rather than a university that only specializes in engineering and electronic science, as the name might suggest.

Economic Science Laboratory of the University of Arizona, located in Tucson, USA.[5] Participants' average earnings were 49 RMB (about 7 USD) in China and 17 USD in the United States[6].

Sessions had between 8 and 12 participants, divided equally between males and females. All participants were undergraduate students registered in the local subject pool. The experiment was computerized and the interface was programmed in ztree (Fischbacher (2007)). Sessions lasted approximately 70 minutes on average.

To control for any nuances in language that may cause the results to differ between countries, the instructions for the experiment in China were translated directly into Chinese from English by a native speaker of Chinese who also speaks fluent English. The sessions in the two locations were conducted by the same (female) experimenter, who is highly proficient in English and a native speaker of Chinese. The English instructions were composed by a native speaker of English.

Each session of the experiment consisted of six periods, a brief questionnaire, and a consolation prize ceremony. In the first four periods, we elicited participants' willingness to pay to hide the fact that they were a relatively poor performer on a cognitive task. In these periods, low performance can be attributed to low ability. In the last two periods, 5 and 6, we elicited their willingness to exert higher effort to hide poor performer status, in a setting where low performance is due to low effort. Afterwards, we elicited participants' beliefs regarding gender stereotypes. Finally, there was a consolation prize ceremony. The following four subsections describe the activity that took place in each session.

3.2 Periods 1 to 4: Measuring the value of hiding worse performer status

In this initial part of the session, we measured the participants' willingness-to-pay to avoid having others know that they were a worse performer than a competitor.

[5]In each of the two locations, we only permitted domestic students, defined as citizens of the home country, to participate. There was one exception. In one of the sessions conducted in the US, we permitted three foreign male students to participate in order to make the session viable and avoid cancellation of the session. The other participants were unaware that these three students were foreign. We only retained data from 117 participants in the US sessions, excluding the data from the three foreign students from our analysis.

[6]To render the stakes in the two locations comparable, we set the anticipated payment levels in each location approximately equal to the price of two typical lunch meals in campus student restaurants.

In each period, each participant was randomly paired with a newly-drawn partner. The design was counterbalanced: in one half of the sessions, the partner was of a different gender in periods 1 and 2, and of the same gender in periods 3 and 4. In the rest of the sessions, a participant was grouped with a same-gender partner first, and with a partner of a different gender for periods 3 and 4. The specific person the participant was paired with always changed randomly after each period. Individuals always knew the gender of the individual with whom they were paired, though they did not know precise identity the individual.

The task that all participants performed in these first four periods was a series of additions of five randomly-chosen, two-digit numbers. They had four minutes to solve as many problems correctly as possible. Calculators were not permitted, but scratch paper was available.

At the beginning of each period, participants were informed that their performance in the mathematics addition task would be compared to their partner's. The person who outperformed his/her partner was designated as the "Better Performer", while the person who answered fewer questions correctly was the "Worse Performer". The Better Performer earned a greater amount of money for the task than the Worse Performer. In the US, the payments were $10 and $5 for Better and Worse Performers, respectively. In China, the payments were 30RMB and 15RMB.

Participants were also informed that Worse Performers were eligible to claim consolation prizes. However, to claim a prize, a Worse Performer was required to participate in a "Consolation Prize Ceremony" at the end of the session, and that the ceremony required them to go to the front of the room to be recognized by all participants as a Worse Performer.

In each period, before the addition task, we asked each participant to make ten decisions. In each decision, they were asked to indicate whether they were willing to claim possible consolation prizes of different values by going to the front of the room, revealing to all that they were a Worse Performer. The decisions were presented in a price list format, as shown in Figure 1. At the time that they made their decisions,

they were aware of the procedure of the Consolation Prize Ceremony.[7]

Decision number	The value of the consolation prize you can claim if you go to the front of the room	Will you go to the front of the room to claim the prize?
1	$0.25	☐ Yes ☐ No
2	$0.50	☐ Yes ☐ No
3	$0.75	☐ Yes ☐ No
4	$1.00	☐ Yes ☐ No
5	$1.25	☐ Yes ☐ No
6	$1.50	☐ Yes ☐ No
7	$1.75	☐ Yes ☐ No
8	$2.00	☐ Yes ☐ No
9	$2.25	☐ Yes ☐ No
10	$2.5	☐ Yes ☐ No

Figure 1: Consolation Prize Form

Periods 1 - 4 enable us to measure each participant's willingness-to-pay to avoid having others to know that they lagged behind a partner of (a) a different gender and (b) of the same gender, and thus to measure the gap between the two values. If a participant chose "Yes" on the form shown in Figure 1 even when the value of the consolation prize was $0.25, it meant that this participant's cost of having others know that she was a Worse Performer was less than 25 cents. If a participant placed a greater value on others' perception of her relative ability, she would switch from "No" to "Yes" at a point in the decision table corresponding to a greater monetary value.

At the end of the session, only Worse Performers were eligible to claim consolation prizes. This means that participants' decisions on the Consolation Prize Form

[7]The only difference between the tables in the two locations was that the values of the prizes differed because of rounding. In China, the smallest amount (the value in the first line) was 1 RMB, and the largest amount (the value in the last line) was 7.5 RMB. The rule that we applied to convert the value in USD to RMB was to multiply the nominal USD amount in the table in each line by three and, because some denominations of RMB are not in common use, such as 0.75 RMB, we rounded the values to the nearest half RMB. For example, the value in the third line in the Chinese sessions was 2RMB rather than $0.75 * 3 = 2.35$). The values in the table used in China are given in Appendix C.

counted only if it turned out that they were the Worse Performer in their pair in the period that was chosen to count. If they were the Better Performer, their decisions on the form did not count. They only learned whether they were a Worse or Better Performer immediately prior to the ceremony.

After they completed the form for each period, they performed the mathematics addition task for four minutes. When performing the task, they saw the display shown in Figure 2.

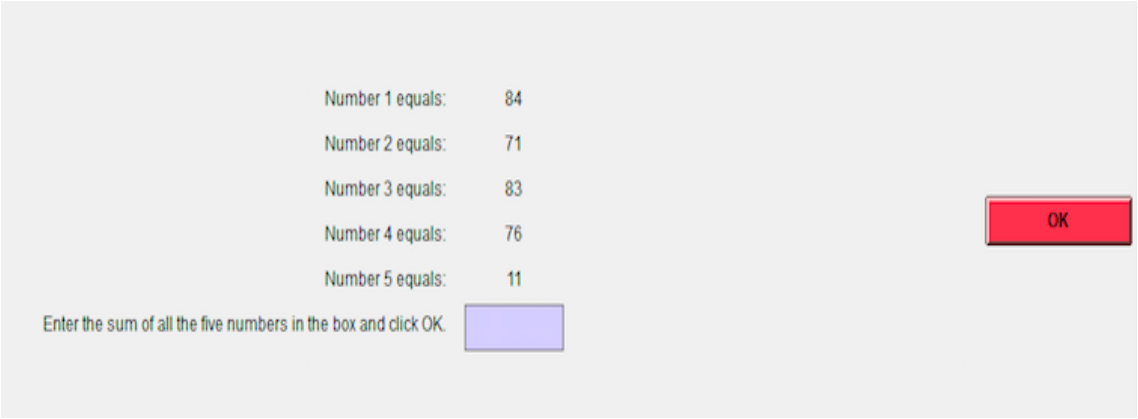


Figure 2: The display of the mathematics addition task

They had four minutes to complete as many addition problems correctly as possible. Their performance was measured as the number of correct answers in the allotted four minutes. They received no feedback on how many of their responses were correct.

After performing the task, they were asked to guess their performance in comparison with their partner's. Specifically, they guessed whether they "performed at least as well as their partner" or "performed worse than their partner". Each guess was incentivized by having $50 cents (1.5RMB) added to their final payment in the event that their answer was correct and the period was chosen to count.

If one of the periods 1 - 4 was selected to count, earnings were determined by whether an individual was a Better or a Worse Performer in the selected period. A Worse Performer's payment was equal to the earnings from the addition task, $5 (15RMB in China) plus any consolation prize received, plus the payment for correctly

guessing their own performance compared to their partner in the selected period. A Better Performer's payment was the sum of the earnings from the addition task, $10 (30RMB in China) and the payment for correctly guessing their comparative performance.

3.3 Periods 5 and 6: Avoiding revelation of Poor Performer status due to low effort

In these two periods, we investigated the willingness to avoid recognition as a Worse Performer when poor performance is due to a lack of effort. In one half of the sessions, each participant was paired with a partner of a different gender at the beginning of period 5 and with a partner of the same gender at the beginning of period 6. In the other half of the sessions, the order of the pairing was reversed.

In each of these two periods, participants performed a series of additions of five randomly-chosen two-digit numbers, as in periods 1 - 4. They were asked to calculate 10 totals correctly, with no time limit. They received a fixed 8 dollars (24 RMB in China) for the period if one of the two periods was chosen to count.

They were required to enter the 10 correct totals on their computer screens in order to proceed further in the session. All participants were able to do so. After they had done so, they could choose to quit or to continue in the task for another three minutes. The decision of only one member of a matched pair would be implemented. Those whose decisions were not implemented could not continue. Allowing implementation of only one decision controlled for selection effects, because otherwise a decision to continue could depend on beliefs about the likelihood that the other player would continue, and these beliefs could depend on the gender of the matched player. Choosing to continue increased one's chances of being a Better Performer.

Every participant knew that if her choice was implemented, then her partner's performance would be scored as exactly 10 problems solved correctly. If both players chose to stop, each player's likelihood of being designated as the Worse Performer was 50%. If she chose to continue, and her decision was implemented, she could

calculate as many totals as she wanted in three minutes and the number she solved correctly would be added to ten[8]. If the partner's choice was implemented, her own performance would be counted as exactly 10, and she could not continue the task even if she chose to continue.

After both players made their decisions, the computer randomly decided whose choice in the group was counted. Those whose choices counted and who also chose to continue, were given three more minutes to do the addition task. After three minutes, they were informed of their performance in these three minutes.

3.4 Eliciting gender stereotype

After period 6, participants answered two questions that appeared on their computer screens. First, we asked them to guess whether men or women performed better in the addition task on average in the preceding session of the study[9]. Second, they were to guess the proportion of participants in the current session who gave the same answer as they did to the first question.[10] The received $1 (3RMB) for each correct answer. All participants received this payment in addition to the earnings from the period 1 - 6 that counted.

3.5 The consolation prize ceremony

At the end of the session, if one of periods 1 - 4 was chosen to count toward the final payment, participants stood in two groups at the front of the laboratory, in full view of all participants. Those participants who were Worse Performers and chose "Yes" in the randomly selected line of the price list were required to go to the front of the room to claim a redemption code for their consolation prize in full view of other participants. The notice *"In the mathematics addition task, these participants*

[8]If a participant answered zero sums correctly in the extra three minutes, she would be randomly assigned as the Worse Performer with probability .5. This did not occur in the experiment.

[9]In each country, for the first formal session, we used the result from a previously conducted pilot session to determine the payoffs.

[10]The first question they were asked was: "We had a group of men and a group of women do the same addition task as you did in periods 1 to 4 in the preceding session. One of the two groups performed better. Please guess whether men performed better or women performed better on average." The second question was: "Please guess the proportion of people in this session agree with your choice above." They were required to choose among "Less than or equal to 25%", "26%-49%", "50%", "51%-74%", and "More than or equal to 75%".

performed worse than their partner with a different gender: Males who stand in front of the room lost to their female partner, and females who stand in front of the room lost to their male partner", or the statement *"These participants performed worse than their partner of the same gender: Males who stand in front of the room lost to their male partner, and females who who stand in front of the room lost to their female partner"*, were shown on all participants' screens while the Worse Performers were standing at the front of the room. After 30 seconds, they returned to their seats and entered their redemption code into the appropriate box on their computer screens.

The remaining participants, including those who were Worse Performers but chose "No" in the randomly selected line, as well as all Better Performers, were asked to stand for 30 seconds in the front of the room. The text *"Better Performers as well as Worse Performers who chose "No" in the selected line are standing in front of the room now"* was shown on all participants' computer screens. They then returned to their seats and the session ended. There was no way for an observer to identify which individuals were Worse Performers among this group.

If either Period 5 or 6 was chosen to be counted toward final earnings, all participants stood in the front of the room in two groups. First, those participants who were Worse Performers in the designated period stood in front of the room for 30 seconds. While they did so, the statement *"These participants are worse performers in the selected period: Males who stand in front of the room lost to their female partner, and females who stand in front of the room lost to their male partner"* or *"These participants are worse performers in the selected period: Males who stand in front of the room lost to their male partner, and females who stand in front of the room lost to their female partner"* was shown on the screen. They then returned to their seats and the participants who were Better Performers went to the front of the room and stood for 30 seconds. The statement *"These participants are better performers in the selected period"* was displayed on every participant's screen.

The total payment for the session consisted of three components, (1) the show-up fee ($5 or 15RMB), (2) earnings from one randomly chosen period (each of the six

periods was chosen to count with equal probability), and (3) the payment for any correct responses in the final questionnaire.

4 Theoretical Framework and Hypotheses

4.1 Theoretical Framework

In this subsection, we model participants' decisions in periods 1 to 4 of the experiment. We characterize the trade-off facing Worse Performers, between accepting a monetary benefit and concealing their Worse Performer status. We assume that people have heterogeneous psychological costs of having others know that they are a Worse Performer. If there are gender stereotypes present, the decision to accept the monetary benefit depends on the audience's belief that the decision maker is the Better Performer, and therefore on the decision maker's gender, the gender of the competing individual, and the audience's beliefs about the ability of each gender.

Denote individual i's performance in the task as x_i, and the opponent j's performance as x_j. An audience has beliefs about the distribution of each individual's performance based on the decision maker's gender. These beliefs may be based on stereotypes of either ability or willingness to exert effort. Each individual's performance x_i is believed to be independently drawn from a distribution that is characterized by either a function $H^M(\cdot)$ with the density function $h^M(\cdot)$ if he is a male, or $H^F(\cdot)$ with the density function $h^F(\cdot)$ if she is a female. For each decision line in the price list described in Section 3, a Worse Performer i can choose $a_i \in \{0, 1\}$. $a_i = 0$ means that he chooses not to claim the prize and to thereby hide his Worse Performer status, and $a_i = 1$ means that he chooses to claim the prize and to reveal his status. If the Worse Performer chooses to hide, observers cannot distinguish him from the winner, and he can disguise the fact that he has lost. If he chooses to reveal, he can gain a monetary benefit.

A Worse Performer values his material payoff, which is composed of the fixed payment earned from doing the task and the decision to accept the consolation prize or not, as well as others' perception of the probability that his performance x_i

is worse than x_j. For each possible prize, the utility of a Worse Performer is given by: $\mathrm{U}_i(a_i) = M_{worse} + a_i \cdot prize - \lambda_i Pr_{-i}[x_i < x_j | a_i]$, where M_{worse} is the fixed payment in the task to the worse performer, a_i denotes the action taken by the individual, $prize$ is the value of a consolation prize, and λ_i denotes how sensitive the individual is about others' knowing his performance. We assume λ_i is distributed over $[0, \bar{\lambda}]$ with cumulative distribution $F(\cdot)$ and strictly positive density $f(\cdot)$. $Pr_{-i}[x_i < x_j | a_i]$ denotes the audience's belief about i's relative performance upon observing i's action a_i.

If the Worse Performer chooses to hide his status, he does not receive the monetary value of the prize, and others do not learn his status. Then his payoff is $\mathrm{U}_i(a_i = 0) = M_{worse} - \lambda_i Pr_{-i}[x_i < x_j | a_i = 0]$. If he claims the prize and reveals that he lags behind his partner, his payoff is $\mathrm{U}_i(a_i = 1) = M_{worse} + prize - \lambda_i$

Let $\alpha(\lambda)$ be the probability that a worse performer of type λ chooses to claim the prize. First, notice that the behavior is weakly monotonic in λ: if some type λ' chooses to reveal, then all types $\lambda_0 < \lambda'$ will choose to reveal as well, and if some type λ'' chooses to hide, then all types $\lambda_0 > \lambda''$ will also hide. Second, there cannot be pooling behavior, where all types choose to hide or to reveal. To see this, notice that if all types choose $\alpha(\lambda) = 0$, then it must be that $M_{worse} + prize - \lambda_i \leq M_{worse} - \lambda_i Pr_{-i}(\mathbb{1}(x_i < x_j) | a_i = 0)$ for all λ on its support. Obviously, this inequality will be violated for individuals who don't care about others' inference at all, for whom $\lambda = 0$. Therefore, it cannot be that all the individuals choose to hide.

Similarly, it is not the case that all types choose to reveal. This would mean that all types choose $\alpha(\lambda) = 1$. Hiding one's status is beneficial for those individuals who have $\lambda > prize$. This condition would hold for at least some individuals if the value of the prize is sufficiently small (in our experiment, the smallest prizes are 25 US cents and 1 RMB). Therefore, because behavior is monotonic in λ, and $f(\lambda)$ has strictly positive density, there exists a unique cutoff type, λ^*, who is indifferent between revealing and hiding the result at a given prize (which occurs with probability 0). The point of indifference satisfies $\lambda = \frac{prize \cdot (1 - Pr[x_i < x_j] \cdot \alpha(\lambda))}{1 - Pr[x_i < x_j]}$. We can show the solution to this function is unique, as $\alpha(\lambda)$ is weakly decreasing in λ. We obtain the following

proposition.

Proposition 1. *Individual behavior is characterized by a cutoff strategy: All individuals with $\lambda < \lambda^*$ choose to reveal, ($\alpha(\lambda) = 1$), and all those with $\lambda > \lambda^*$ choose to hide, ($\alpha(\lambda) = 0$). The share of individuals choosing to reveal their status as the Worse Performer is $F(\lambda^*)$.*

Proof. To see this, we can show that $U_i(a_i = 0) \geq U_i(a_i = 1)$ if and only if $\lambda \geq \lambda^*$. This, together with the monotonicity of individuals' behavior in λ and the non-existence of pooling behavior, allows proposition 1 to be derived. ■

We now focus on possible implications of gender on behavior. We assume that the distribution of the sensitivity parameter λ is the same for women and men. However, suppose that $H^M(\cdot)$ first-order stochastically dominates (FOSD) $H^F(\cdot)$ so that there is a stereotype, which may be incorrect, that men are better performers than women on the task in question. To model this, we suppose that $H^M(x)$ first order stochastically dominates $H^F(x)$, so that $H^M(x) \leq H^F(x)$ for all x, with strict inequality at some x. This assumption guarantees that if we randomly draw a female and a male to compete in the task, an observer believes that there is a greater than 50% chance that the female would lose to the male. This assumption is motivated by existing literature (e.g., Reuben et al. (2014), Correll (2001)), as discussed in Section 2, where it is argued that the gender stereotype that males are better than females in STEM fields is widely held. Denote the cutoff type of a male in a mixed-gender pairing (where a female is matched with a male) as λ^*_{MF}, and that of the female as λ^*_{FM}. Let the cutoff type of a male in a same-gender pairing be denoted as λ^*_{MM}, and that of a female as λ^*_{FF}.

A given sensitivity parameter λ corresponds to a unique cutoff value of the consolation prize, $prize^* = \frac{\lambda(1-Pr(x_i<x_j))}{1-Pr(x_i<x_j)*\alpha(\lambda)}$, that an individual would be indifferent between accepting publicly and declining. Call the cutoff value of the prize for males and females in a mixed-gender pairing $prize^*_{MF}$ and $prize^*_{FM}$, respectively. The cutoff value for males in a same-gender pairing is $prize^*_{MM}$, and that of females is $prize^*_{FF}$. We can then state Propositions 2.1 and 2.2.

Proposition 2. *2.1.* $\lambda^*_{MF} < \lambda^*_{MM} = \lambda^*_{FF} < \lambda^*_{FM}$, *and thus* $F(\lambda^*_{MF}) < F(\lambda^*_{MM}) = F(\lambda^*_{FF}) < F(\lambda^*_{FM})$

2.2. $prize^*_{FM} < prize^*_{MM} = prize^*_{FF} < prize^*_{MF}$.

Proof. Denote the belief the audience holds about the probability that a female matched with a male competitor would be a Worse Performer as $Pr(x_F < x^C_M)$, where the superscript C indicates the performance of the competitor. Similarly, $Pr(x_M < x^C_F)$ is the probability that others believe a male would lose to his female partner. Analogously, $Pr(x_M < x^C_M)$ and $Pr(x_F < x^C_F)$ are the probabilities that a male would lose to a male partner, and that a female would lose to a female partner, respectively. We assume that $Pr(x_M < x^C_M) = Pr(x_F < x^C_F) = \frac{1}{2}$. Because $H^M(x)$ first order stochastically dominates $H^F(x)$, we have that $Pr(x_M < x^C_F) < \frac{1}{2} = Pr(x_M < x^C_M) = Pr(x_F < x^C_F) < Pr(x_F < x^C_M)$. It follows that $\frac{\partial \lambda^*}{\partial Pr(x_i < x_j)} > 0$ and $\frac{\partial prize^*}{\partial Pr(x_i < x_j)} < 0$, with $i, j \in \{M, F\}$, and thus Proposition 2 follows. ■

In words, Proposition 2 tells us that the threshold prize value, above which individuals would choose to claim the prize and below which they would not, is highest for males when they are competing with a female, and lowest for females when they are competing with a male. Consequently, the likelihood of a Worse Performer claiming a given prize is lowest for males with a female partner, and highest among females competing against a male.

4.2 Testable Hypotheses

From Propositions 1 and 2, we can extract the following Hypotheses 1 to 4, which are testable in our experiment. The first three hypotheses are implications of inequalities in Proposition 2.1 and 2.2.

Hypothesis 1 *(Existence of Public Stigma)*: A positive proportion of subjects chooses not to claim a consolation prize.

Hypothesis 2 *(Gender Difference)*: In a mixed-gender pairing, the proportion of individuals claiming the prize is greater for females than for males. The cutoff prize value for females is lower than for males.

Hypothesis 3 *(Sensitivity to Gender of Counterpart)*: For females, the probability of claiming a consolation prize is greater in a mixed-gender than in a same-gender pairing. For males, the opposite relationship is present. In other words, females' cutoff value of the prize is higher in a same-gender than in a mixed-gender pairing, while males' cutoff value is higher in a mixed than in a same-gender pairing.

The next hypothesis is conditional on the questionnaire responses. It asserts that if the responses indicate that there is no gender stereotype in the task, the effects described in Hypotheses 2 and 3 would not be present.

Hypothesis 4 *(No Stereotype Case)*: If there is no stereotype, so that $Pr_{-i}[x_i < x_j] = 1/2$ when i and j are of different genders, the cutoff value is the same for men and women, and independent of the gender of the counterpart.

We now advance a hypothesis regarding gender differences in periods 5 and 6. An implication of the analysis in Section 3 is that the cost of being recognized as a Worse Performer is greater among males than females on average. If this is the case, a male would be more willing to expend extra effort, and incur the associated cost, to avoid exposure as a Worse Performer.[11] This argument is the basis for Hypothesis 5.

Hypothesis 5 *(Gender difference in cost of exposure as Worse Performer due to low effort)*: In a mixed-gender pairing, a higher fraction of males than females will choose to exert greater effort to avoid public recognition that they are a Worse Performer. Males will be more likely to choose higher effort in a mixed-gender than

[11] While we have not written the argument formally, the intuition is clear. An individual trades off the disutility of embarassment and the cost of effort. For those with a high disutility of embarassment, incurring the effort cost to reduce the probability of exposure will leave them better off in expectation.

in a same-gender pairing, while females will be more likely to choose higher effort in same-than in mixed-gender pairings. If there is no gender stereotype, both genders will be equally likely to exert greater effort, and the likelihood will not depend on the gender of the competitor.

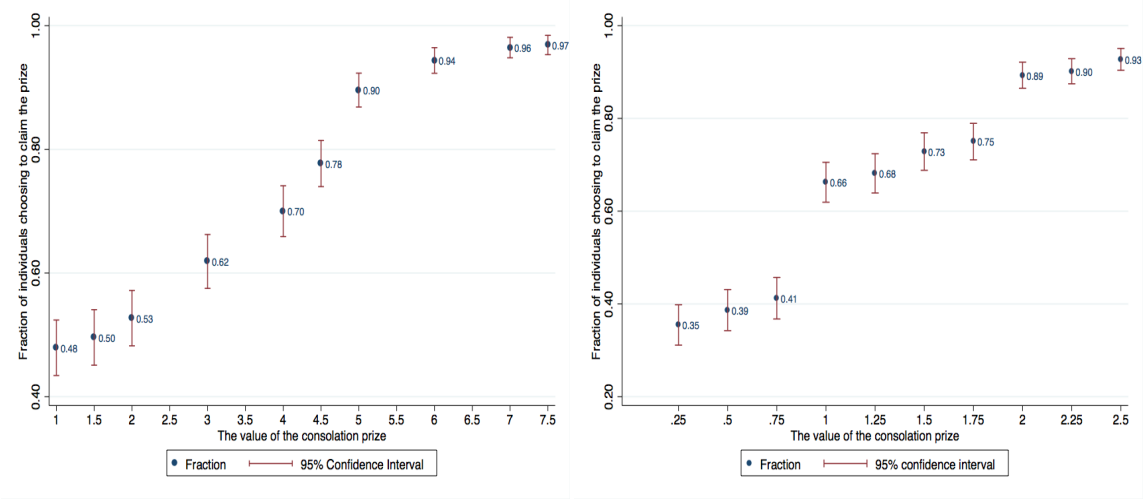
5 Results

5.1 Is there a cost of being revealed as a poor performer?

Figure 3 illustrates the percentage of participants who chose to claim the prize at each possible value, in both the Chinese and the American sessions. The figure shows that individuals' behavior was monotonic in both countries. A greater share of participants chose to reveal their Worse Performer status, the more money they received from doing so. From the left panel, we observe that approximately 50% of Chinese participants chose to claim the smallest possible prize of 1RMB. This half of subjects was not sensitive to others' perception of their relative ability. On the other hand, 3% of individuals chose to hide their Worse Performer status even if the prize was 7.5RMB, one half of the baseline fixed payment from the task. These individuals had a high cost of exposing their Worse Performer to the audience.

A similar pattern was present in the U.S. Roughly 35% of participants were willing to expose their Worse Performer status to the public for 25 cents, the minimum value of the prize. About 8% of individuals chose to hide their status at the highest prize value, $2.50. The 95% confidence intervals around the percentage of individuals accepting the consolation price does not include 1 for any prize in either location. Thus, it is clear that Hypothesis 1, asserting that some individuals are willing to forego some positive monetary payment to avoid exposing their Worse Performer status, is supported. A majority of individuals declines the smallest prize.

Figure 3: Proportion of individuals choosing to reveal Worse Performer status for each possible consolation prize.



(a) China (b) the U.S.

Note: Each dot represents the proportion of participants that choose to claim the prize. A value of 1 means 100%. The error bars are 95% confidence intervals for the proportions.

In the table below, we provide regression results that show participants' tendency to claim the prize in each line of the decision table described in Section 3. The variable "Reveal" equals 1 if the individual chooses to claim the prize, and "0" if she chooses not to. Column 1 in each panel (China, U.S) reports the result from regressing individuals' decisions to claim the prize or not on the line in the display shown in Figure 1, using a linear probability specification. Column 2 estimates a similar model using a Probit specification. The results show that people are more likely to claim the prize as the line number increases, i.e., as the value of the prize increases, as the coefficient for *Row*, the line number, is positive. *Row2*, in column 3 in each panel, is the square of the number of the line in the table. The negative coefficient on *Row2* in the US indicates that the rate of increase in switching decreases as the value of prize grows.

Table 1: The effect of consolation prize on the willingness to reveal oneself as a Worse Performer

	China			The U.S		
	(1) LPM Reveal	(2) Probit Reveal	(3) LPM Reveal	(1) LPM Reveal	(2) Probit Reveal	(3) LPM Reveal
Row	0.065***	0.235***	0.078***	0.070***	0.216***	0.108***
	(0.006)	(0.011)	(0.015)	(0.006)	(0.011)	(0.019)
Row2			-0.001			-0.003**
			(0.001)			(0.002)
Cons.	0.381***	-0.509***	0.355***	0.287***	-0.666***	0.211***
	(0.055)	(0.072)	(0.058)	(0.051)	(0.068)	(0.058)
N	480	480	480	468	468	468

Note: The standard errors are clustered at the subject level. The dependent variable is whether the individual chooses to reveal or not, and Row represents the row number in the table. A larger number corresponds to a larger monetary value of the prize. Row2 is the square of Row. Columns 1 and 3 in each panel employ a linear probability model, while column 2 uses a Probit specification. *p<0.1, **p<0.05, ***p<0.001

Figure 3 and Table 1 serve as the basis for our Result 1.

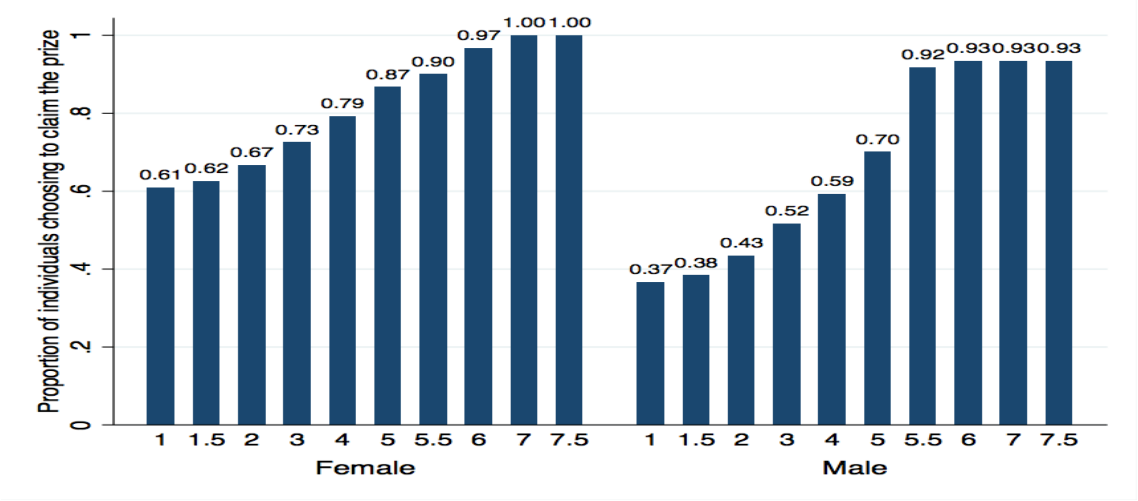
Result 1: Hypothesis 1 is supported. More that 50% of participants forego a consolation prize in order to hide their Worse Performer status. As the consolation prize becomes larger, more individuals are willing to claim it.

5.2 Is there a gender difference?

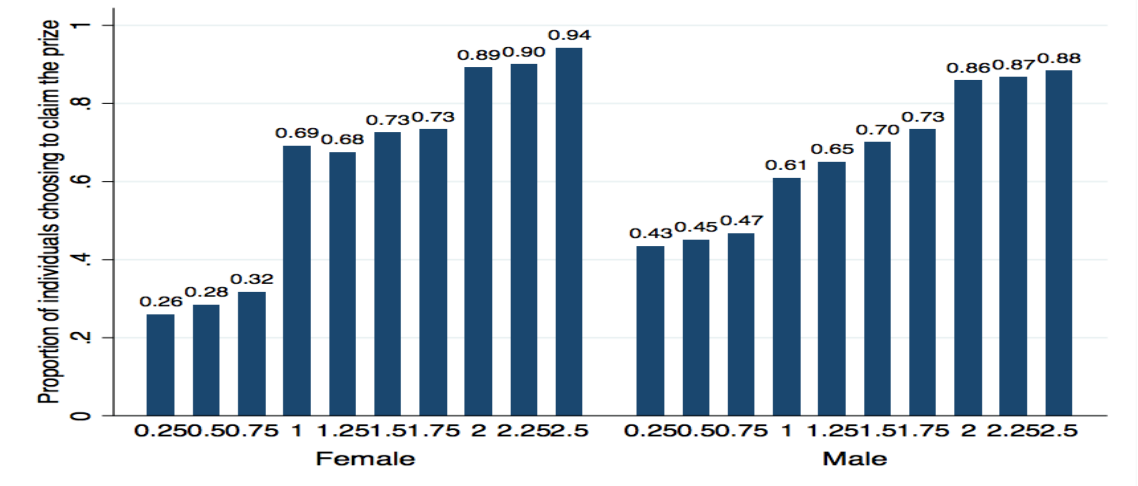
We now consider how the thresholds, at which participants are willing to claim the prize, vary by gender. Figure 4 shows the percentage of participants claiming each prize, by gender, in the mixed-gender pairings. Panels (a) and (b) in Figure 4 depict the patterns in China and in the US, respectively. The left side shows the data for

females, and the right illustrates that for males. In China, for each possible value of the prize, the proportion of males choosing to claim the prize is lower than that of females. In the US, in contrast, slightly more males than females choose to claim the prize, and thus to reveal Worse Performer status, if the prize value is relatively low. Beyond 1 US dollar, there is no obvious difference between the behavior of males and females. Figure 5 summarizes the differences in the overall average row at which switching occurs. It confirms that there is a strong gender difference in China, though not in the US.

Figure 4: Proportion of individuals choosing to reveal Worse Performer status for each consolation prize, by gender, in mixed-gender pairings



(a) China

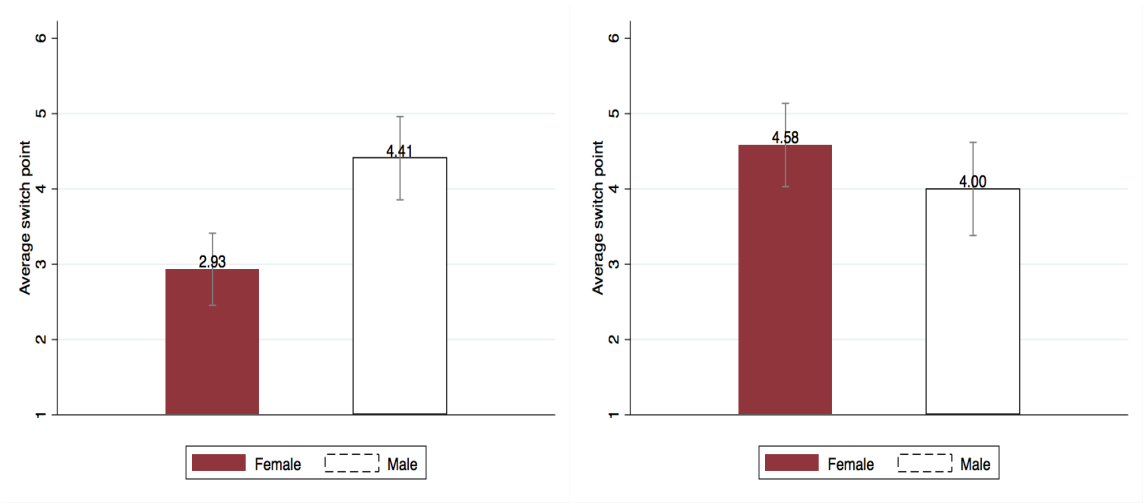


(b) the U.S.

In China, the average cutoff value for males in mixed-gender pairings is 3.25RMB, compared with females' 2.3RMB. Both a non-parametric Mann-Whitney U test and the two-sample t-test yield $p < 0.01$, indicating that the mean and median are significantly lower for women than for men. A Kolmogorov-Smirnov test for equality of distributions results in a p-value < 0.001, amounting to overwhelming evidence of a difference between the distributions of male's and female's cutoff values.

The data from the mixed gender groups in the U.S follow a different pattern. The average cutoff value for males is lower than that for females ($1.02 vs. $1.16), though the difference is not significant. Both the non-parametric Mann-Whitney U test and the two-sample t-test yield $p > 0.1$. The p-value of the Kolmogorov-Smirnov test is 0.162. Thus, there is no evidence supporting the hypothesis that women and men have different cutoff prize levels in mixed gender pairings in the US.[12]

Figure 5: Average row at which switch occurs, by gender, in mixed pairings



(a) China

(b) the U.S.

We also examine the gender difference in the rate of accepting the smallest payment. First, we strongly reject the null hypothesis that the proportion of males and

[12]If we analyse the data from Periods 1 and 2 only, which cannot be affected by prior activity, all the conclusions above remain the same. The p-values from the Mann-Whitney U test and t-test are 0.0135 and 0.0093, respectively, in China. The corresponding p-values are 0.2908, and 0.3645 respectively in the U.S. Thus, the mean and median cutoff values are significantly lower for women than for men in China, but not in the U.S. Regarding the gender difference in the percentage of individuals choosing to reveal Worse Performer status at the smallest consolation prize, the p-values from the Fisher-exact test are 0.053 and 0.027 in China and in the US, respectively. More females are willing to accept the smallest value in China, but more males are willing to do so in the US.

females claiming the prize of the smallest value is the same in China. The p-value of the Fisher-exact test of the equality of the two proportions is 0.018. More females than males accept this lowest payment of 1RMB. The fraction of males and females that are willing to claim the prize is also different in the US. The p-value of the Fisher-exact test is 0.032. However, contrary to what we see in China, it is a higher fraction of males than females that are claiming the lowest prize.

Based on the evidence described above, we obtain our Result 2.

Result 2: Hypothesis 2 holds in China though not in the US. In China, when individuals compete with a partner of a different gender, the cutoff value for females is smaller than that for males. In the US, there is no gender difference.

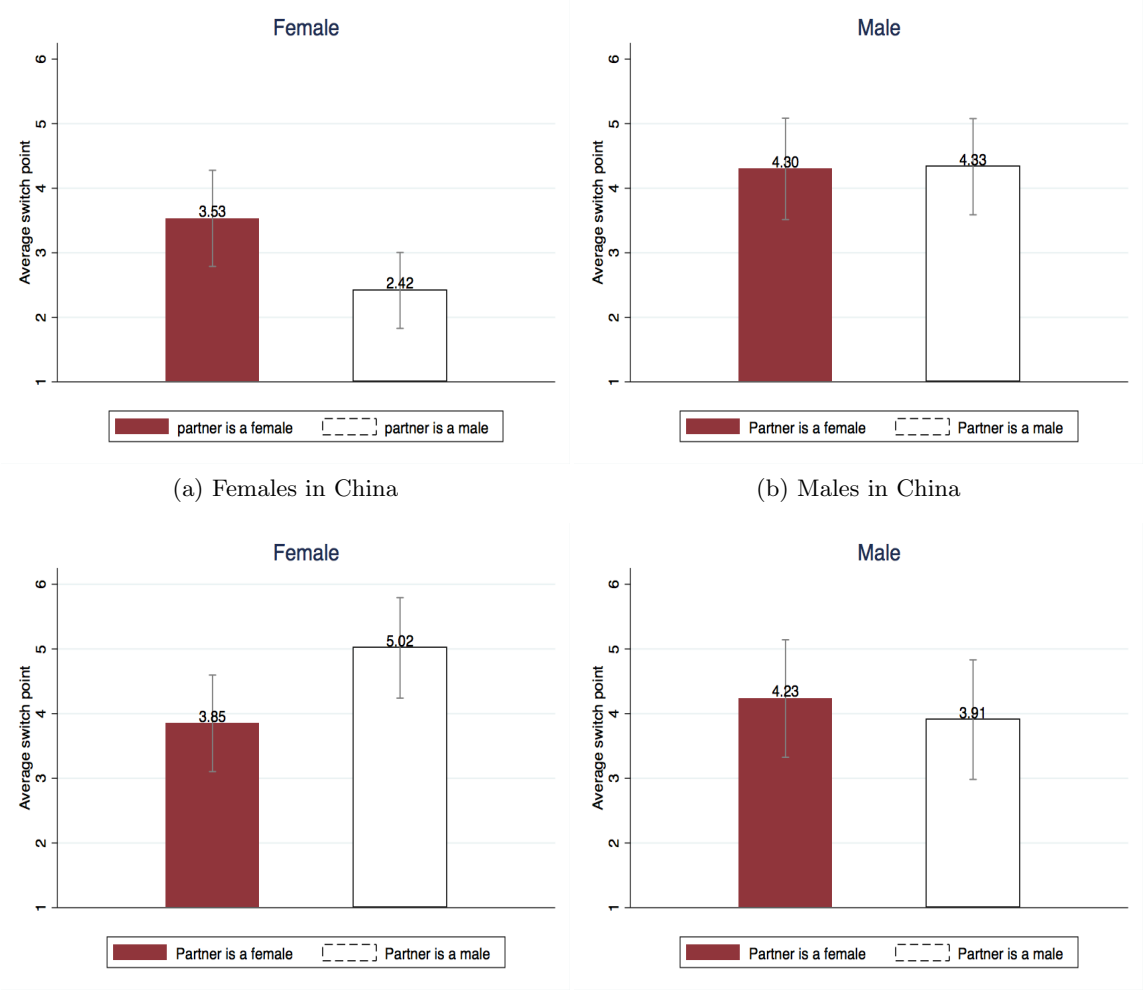
5.3 Does the gender of the counterpart matter?

We now study whether the gender of the counterpart affects individuals' willingness to reveal their worse performer status. In principle, we can conduct a within-subject comparison. However, as we show in Appendix A, there are carryover effects in our data, which means that individuals' choices in the second condition in a session, present in periods 3 and 4, are influenced by activity in the earlier condition in effect in Periods 1 and 2.

In our design, we have same-gender pairings in the first two periods in one half of the sessions, and mixed gender pairings in the first two periods in the remaining sessions. We conduct a between-subject comparison of the data from mixed-and same-gender pairings in the first two periods. Figure 6 shows that in China, females are less likely to reveal their Worse Performer status if their partner is a female than if their partner is a male. This difference is statistically significant (p-value from the one-sided t-test is 0.0108, and that of a Mann-Whitney test is 0.0290). However, the reverse relationship is present in the US, where females are less likely to reveal that they are a Worse Performer when matched with a male than a female (the p-value from the Mann-Whitney test is 0.0234, and that from the t-test is 0.0170). Males'

choices are independent of their partner's gender both in China ($p = 0.4757$, MWU $= 0.9356$) and the U.S ($p = 0.3100$, MWU $= 0.7879$). The findings are summarized as Result 3.

Figure 6: The average switch point by partner's gender, periods 1 and 2



(a) Females in China

(b) Males in China

(c) Females in the U.S

(d) Males in the U.S

Result 3: Hypothesis 3 is partially supported. Females' cutoff value for the consolation prize is higher in same gender than in mixed gender pairings in China. However, the opposite is the case for females in the US. In both countries, the behavior of males does not depend on the gender of the person they are matched with.

5.4 Differences between the two locations

The discussion above has revealed some different patterns between our two groups. We now compare the responses in the two countries to the questions regarding beliefs about the comparative performance of males and females. In the top panel of Table 2, containing the data from China, the data show that only 27 out of 120 subjects guess that females would perform better than males. This can be seen in the rightmost column of the table. Among these 27 subjects who thought females would outperform males, almost half of them (11 of 16) believe that only a minority of the cohort shares their opinion. 93 out of 120 participants believe that males perform better than females on average, and only 2 out of these 93 guess that fewer than half of participants in the same session agree with their choice. Overall, 106 of 120 participants, including 48 females and 56 males, think that the majority of the session believes that men would perform better on average. Thus, in our Chinese sample, large majorities believe that males do better than females on average and also think that a majority of others also holds that view. There is a strong gender stereotype among our Chinese sample.

In contrast to the pattern in China, the rightmost column in the lower panel of the table shows that approximately 50% of subjects in the US sessions believe that females would do better on average. The majority of individuals, regardless of their belief about which gender would perform better, think that the majority of the participants agree with their choice. Thus, there is no significant stereotype about which gender would do better in the task in the US, but there does appear to be a degree of false consensus in that most respondents believe that a majority shares the same view that they themselves have.[13]

Under this pattern of beliefs, Hypothesis 4 predicts that males would have higher cutoff prizes than females in China, and that there would be no gender difference in

[13]There is no significant gender difference in performance in our data. A t-test of the hypothesis that there is a difference between males and females yields a p-value of .178 in China and .0684 in the U.S. The average number of problems solved correctly averages approximately 11.5 in China and 7.5 in the U.S. This difference is consistent with other measures of relative performance in mathematics in the two countries, presumably reflecting the difference in emphasis between the two countries' educational systems. The lack of a gender difference in performance in our addition task is consistent with prior research in experimental economics (Niederle and Vesterlund (2007), Reuben et al. (2014)).

Table 2: Participants' beliefs about whether females or males performed better in China

	Guess the proportion of subjects in the same session agree with your choice											
Guess which group performed better		26%-49%		50%		51%-74%		=75%			
		Total	F/M	Total	F/M	Total	F/M	Total	F/M	Total	F/M
	Women performed better	12 (44%)	10/2 (45%)/(40%)	6 (22%)	5/1 (23%/)(20%)	7 (26%)	6/1 (27%)/(20%)	2 (7%)	1/1 (5%)/(20%)	27 (100%)	22/5 (81%)/(19%)
	Men performed better	1 (1.08%)	0/1 (0%)/(2%)	1 (1.08%)	0/1 (0%)/(2%)	53 (56.99%)	21/32 (55%)/(58%)	38 (40.86%)	17/21 (44%/)(38%)	93 (100%)	38/55 (41%)/(59%)

Note: This table summarizes the belief elicitation results in China at the aggregate level as well as the distribution by gender. For example, from the third column of the table, we observe that there were 12 participants who believed that women performed better in the addition task and that 26-49% of their peers in the session shared the same belief. Among these 12 people, 10 were females, and 2 were males, accounting for 45% of females and 40% of males who held the belief that women performed better

Table 3: Participants' beliefs about whether females or males performed better in the U.S

	Guess the proportion of subjects in the same session agree with your choice											
Guess which group performed better		=25%		26%49%		50%		51%-74%		=75%		
		Total	F/M	Total	F/M	Total	F/M	Total	F/M	Total	F/M	Total	F/M
	Women performed better	3 (5%)	3/0 (9%)/(0%)	9 (16%)	5/4 (15%)/(17%)	7 (12%)	7/0 (21%)/(0%)	37 (64%)	19/18 (56%)/(75%)	2 (3%)	0/2 (0%)/(8%)	58 (100%)	22/5 (81%)/(19%)
	Men performed better	0 (0%)	0/0 (0%)/(0%)	16 (27%)	4/12 (15%)/(36%)	3 (5%)	1/2 (4%)/(6%)	34 (58%)	16/18 (62%)/(58%)	6 (10%)	5/1 (8%)/(3%)	59 (100%)	38/55 (41%)/(59%)

Note: This table summarizes the belief elicitation results in the U.S at the aggregate level as well as the distribution by gender. For example, from the third column of the table, we observe that there were 3 participants who believed that women performed better in the addition task and that 25% of their peers in the session shared the same belief. All of them were females, accounting for 9% of females who held the belief that women performed better

the US. It also predicts that for each gender, there would be a gap in cutoff values between a same-and an opposite-gender pairing in China but not in the US. Specifically, the predictions of Hypothesis 4 are that (1) $prize_{FF}^{*CN} > prize_{FM}^{*CN}$, (2)prize$_{MF}^{*CN} > prize_{MM}^{*CN}$, and (3) prize$_{FF}^{*US} = prize_{FM}^{*US} = prize_{MF}^{*US} = prize_{MM}^{*US}$.

The discussion in subsections 5.2 and 5.3 shows that inequality (1) is supported in the data. In China, women have a significantly lower cutoff value when facing a man than a woman. However, (2) is not supported, in that men's behavior does not respond to the gender of their competitor. Equality (3), the most stringent, is observed, with the exception that $prize_{FF}^{*US} < prize_{FM}^{*US}$. That is, women in our US sample have a higher cost of revealing poor performance when in a mixed-gender than a same-gender pairing. These results are summarized as our Result 4.

Result 4: The belief that males are likely to be better performers than females is present in China, but not in the U.S. As predicted by our model, males are more reluctant than females to reveal Worse Performer status in China, but not in the US. Hypothesis 4 is supported.

Table 4: Participants' choices of continuing or quitting the task in Periods 5 and 6

(a) China

Panal A: Female's Choice

	Partner's gender		
	Female	Male	Total
Quit	68	66	134
	(56.67%)	(55%)	
Continue	52	54	106
	(43.33%)	(45%)	
Total	120	120	240

Panal B: Male's Choice

	Partner's gender		
	Male	Female	Total
Quit	68	70	138
	(56.67%)	(58.33%)	
Continue	52	50	102
	(43.33%)	(41.67%)	
Total	120	120	240

(b) The U.S

Panal A: Female's Choice

	Partner's gender		
	Female	Male	Total
Quit	67	65	132
	(57.26%)	(55.56%)	
Continue	50	52	102
	(42.74%)	(44.44%)	
Total	117	117	234

Panal B: Male's Choice

	Partner's gender		
	Male	Female	Total
Quit	66	66	132
	(56.41%)	(56.41%)	
Continue	51	51	102
	(43.59%)	(43.59%)	
Total	117	117	234

Note: In each panel, the rows indicate number of participants choosing to quit and to continue, by partner's gender. The percentages of individuals of a given gender choosing to quit or to continue when facing a partner with the gender indicated in the title of the column are given in parentheses.

We now turn to the data from Periods 5 and 6. Table 4 illustrates participants' choices to continue or to quit the task. The table shows that between 55 and 60 percent of participants choose to quit, regardless of location, own gender, and the gender of their partner. We fail to reject the null hypothesis that males and females have the same likelihood of continuing the task, as a Fisher's exact test yields $p = 0.223$. Chinese and American participants also have the same probability of choosing to

continue the task ($p = 0.892$). The probability of continuing is unaffected by the gender of the competitor, as all of the relevant tests are also insignificant. The above forms the basis for our fifth result.

Result 5: Hypothesis 5 is not supported. The same proportion of females and males choose to continue the task to avoid exposure as a Worse Performer. There is no difference between Chinese and American participants.

6 Conclusion

It has been long been claimed by observers and commentators that some men experience a particular disutility if they are outperformed by women. Some economic research has documented behavior that is consistent with this notion. In this paper, we use an experimental approach to obtain a measure of the willingness-to-pay to avoid being recognized as an inferior performer relative to a peer. The measure can be compared across different gender pairings and populations, and defined over different activities. We chose to examine mathematics performance, for which gender stereotypes are often strong. We conducted the measurement exercise with two groups of university students, at an American and a Chinese university.

We observe that among our Chinese participants, there is a stereotype that men would perform better at our task. This is consistent with the findings of Chinese sociologists (Dong (2019), Tsui (2007)). Under such beliefs, we predict that men would have a greater cost of being revealed as inferior performers than women would. Indeed, this is what we observe in our data in the sessions conducted in China. However, among our American participants, the gender stereotype is not present, and there is no significant gender difference in willingness to have poor performance made public. Among both groups, most participants have a switch point within the range of consolation prize values of our experiment, indicating that our parameters

are well-calibrated to elicit participants' willingness-to-pay.

Male subjects in both participant groups behave similarly regardless of whether their competitor is male or female, suggesting that the men who do not like to be outperformed experience the same cost whether the outperforming counterpart is female or male. Women's behavior, in contrast, does respond to the counterpart's gender. The pattern is very different in the two locations. In the US, women have a stronger aversion to being recognized as performing worse than a male than another female. In China, the opposite pattern is seen, where women have less of an aversion to being outperformed by a male than a female.

While there is a stereotype that men would outperform women on our mathematical task among our Chinese participants, and this translates into a desire for men to avoid being seen as having been outperformed, we do not observe this pattern in our American data. This might reflect changing attitudes and stereotypes in the US, where the performance of women in STEM has caught up or surpassed that of men among young cohorts, such as those we have studied. Our participants are between 18 and 23 years of age. It would be interesting, in follow-up work, to conduct the experiment with older American cohorts, where gender stereotypes might be stronger, and compare their behavior with their younger counterparts.

While the first part of the experiment (periods 1 - 4) measures whether individuals have a stigma cost for exhibiting low ability, the second part focuses on a potential cost associated with exerting low effort. In the second part of the experiment, we observe that nearly half of individuals choose to continue with the task rather than be recognized as performing poorly because of low effort. This means that there is a positive cost for them to being revealed as one who was unwilling to exert more effort on the task. However, there is no difference in this tendency based on one's own or the competitor's gender. Thus, while stigma costs based on ability have gender correlates, those from low effort are similar for women and men, and also very similar between our Chinese and American participants as well.

This research can be extended in a number of directions. One avenue is to consider a task in which there is a stereotype that women are more capable than

men and likely to perform better. The research question would be to consider whether behavior in such a task would exhibit symmetric patterns to those we observe here. Another topic would be to consider the effect of making stereotypes common knowledge on performance and on the willingness to hide worse performer status. A third avenue would be to conduct a field experiment along the same lines to study the generality of the gender differences we have observed.

Appendices

We include four appendices. In Appendix A, we discuss the carryover effects that appear in our data. Appendix B contains an analysis of confidence levels and their relation to behavior. Appendix C consists of an English translation of the computerized form used in Periods 1 - 4 in China. Appendix D contains the instructions for the experiment. The Chinese version of the materials for the experiment are available from the authors upon request.

A Carryover Effects

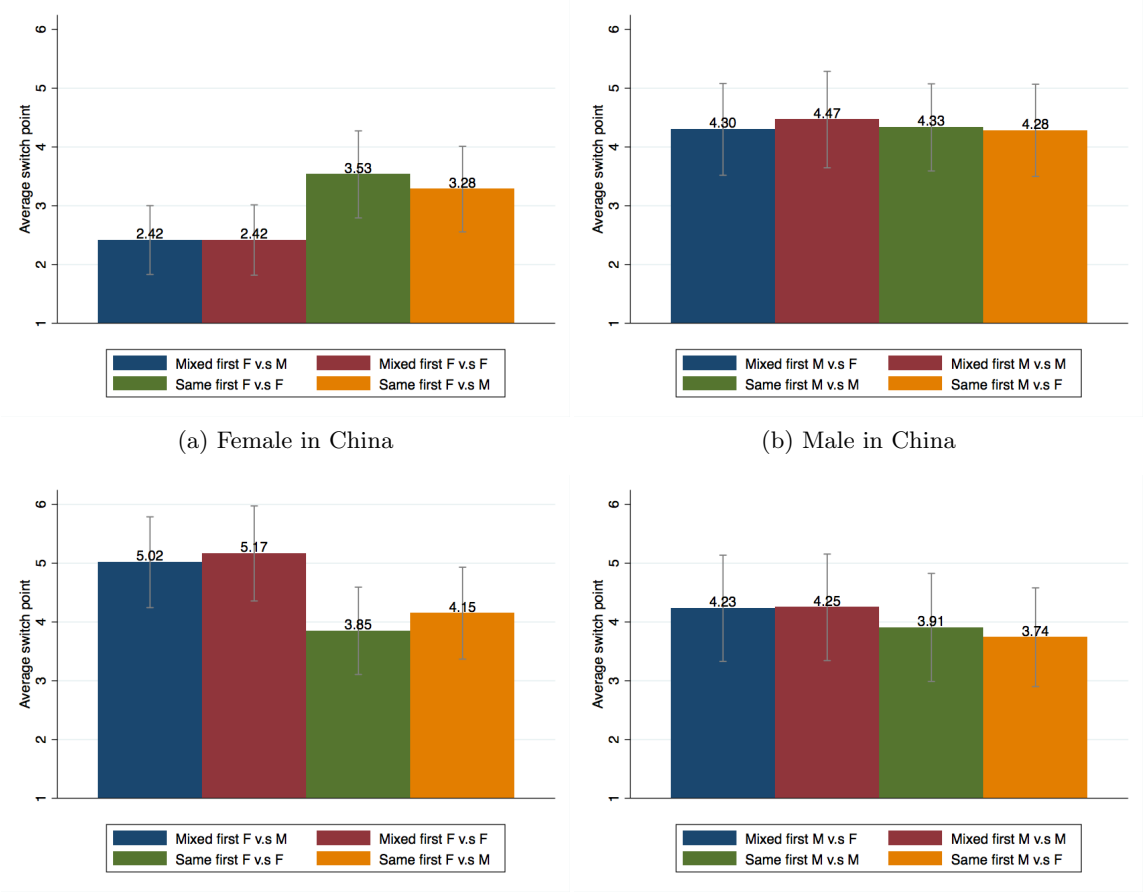
In this appendix, we show that there are strong hysteresis/carryover effects between the first two and the next two periods. In Periods 1 and 2, participants are paired with a partner of either their own or another gender. In Periods 3 and 4, those who were previously paired with a same-gender (different-gender) partner were now matched with a person of a different (same) gender.

Figure 7 displays the average cutoff prizes in the first two and next two periods of the sessions. The first two bars of data in each panel of the figure are from the same sessions. The left-most bar contains the data from the first two periods in sessions where these periods had mixed-gender pairings, and the second bar the data from Periods 3 and 4 of the same sessions, when the matching switched to same-gender. Similarly, the third and fourth bars in each panel also come from the same sessions, with the third bar containing data from the first two and the last bar from the third and fourth periods of the sessions in which the same-gender pairings occurred first.

The figure shows that the behavior in the first two periods is essentially the same as in the latter two periods, for both genders and in both countries. The first two bars in each panel are very similar to each other, and the second two are very similar to each other. However, in contrast, the two pairs of bars differ from each other. The pattern shows that the behavior in Periods 3 and 4 was affected by the earlier periods.

For this reason, we do not conduct within-subject comparisons of the first two periods with the next two to test for the effect of being matched with a partner of a same gender versus a different gender. Such within-subject comparisons are invalid because of the carryover effects. Rather, in our analysis, we rely on between-subject comparisons of behavior in Periods 1 and 2 between people who started the session paired with a competitor of the same gender and those who began paired with a person of the other gender.

Figure 7: Carryover effects between the first two and the next two periods



(a) Female in China

(b) Male in China

(c) Female in the U.S

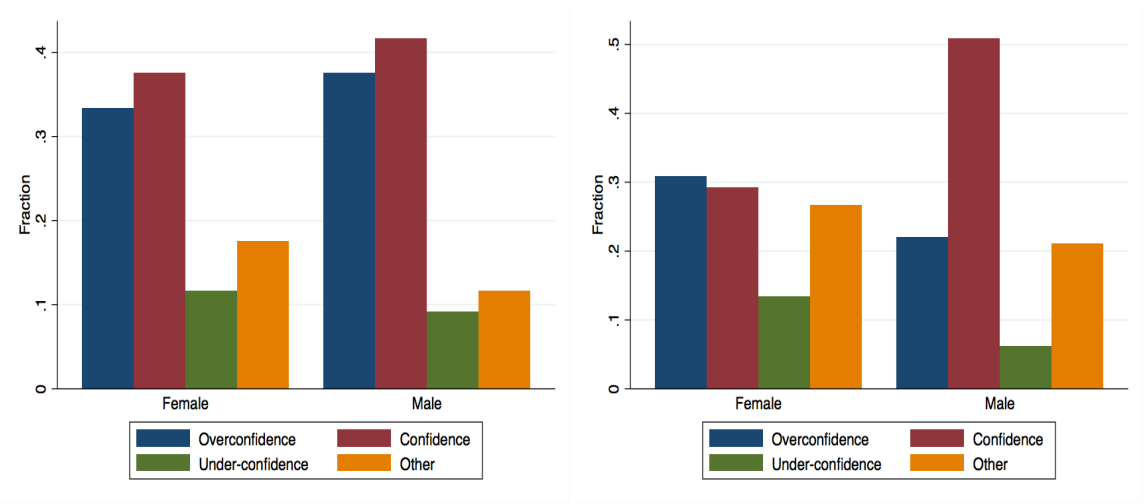
(d) Male in the U.S

Note: The average switch points in periods 1 - 4 and the corresponding 95% confidence intervals are presented in the figure. In each panel, the first two (the third and fourth) bars are the averages from sessions where participants have mixed-gender (same-gender) pairings in periods 1 to 2, and same-gender (mixed-gender) pairings in periods 3 to 4.

B Confidence and decisions

Many studies have investigated gender differences in overconfidence. A prevalent finding is that men are more overconfident than women in domains in which men are thought to be more able (Bordalo et al. (2019)). Figures 8 - 10 depict the distributions of confidence among males and females in the two countries. We define a participant as confident if he guesses that he would perform at least as well as his partner in the task, and does in fact perform better. An overconfident subject is a person who guesses that he would finish ahead of his partner, but it turns out that he actually loses instead. An underconfident participant guesses that he will lose, but actually wins. "Other" in these figures means that the subject guesses he loses and he does lose.

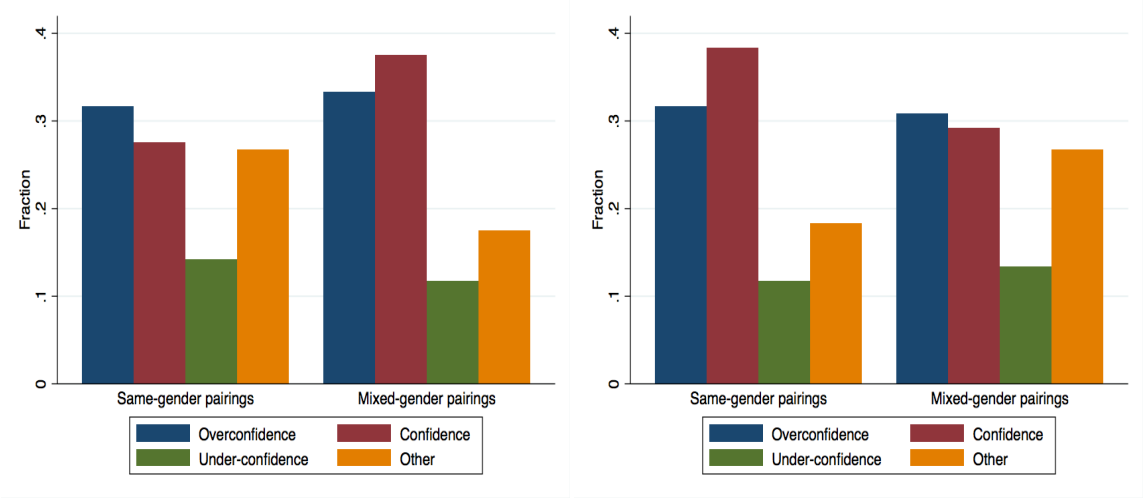
Figure 8: The distribution of confidence among males and females in mixed-gender pairings in the two countries



(a) China

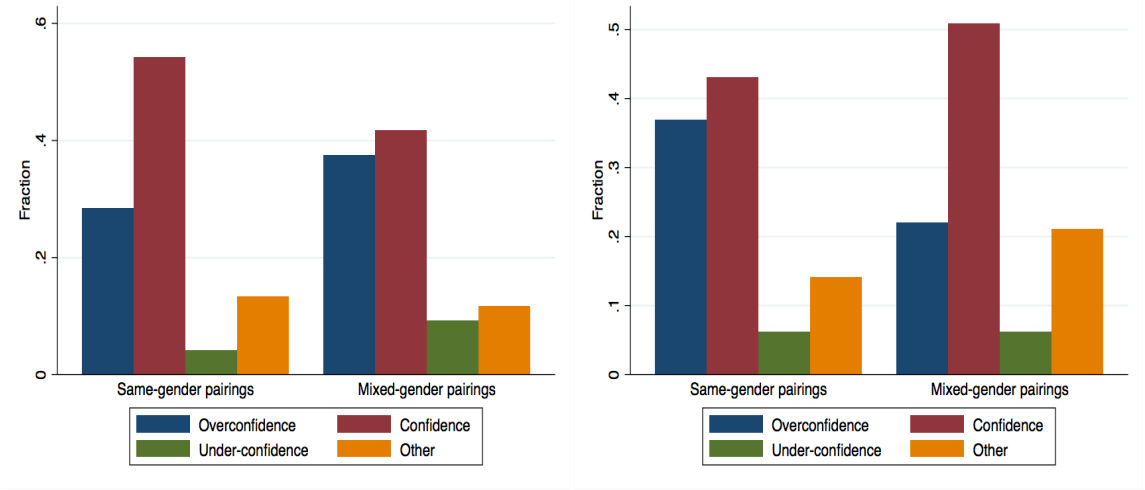
(b) the U.S.

Figure 9: Females' distribution of confidence by the gender of the partner in the two countries



(a) China (b) the U.S.

Figure 10: Males' distribution of confidence by the gender of the partner in the two countries



(a) China (b) the U.S.

When they compete with a partner of a different gender, as can been seen in Figure 8, a higher proportion of females underestimate their ability than males, both in China and in the US. Fewer females than males are overconfident in China, while a greater number of females than males are overconfident in the US. At the same time, Figure 8 also indicates that in China, a similar share of females and males holds the correct belief about their relative ability (the total of individuals in

the *Confidence* and *Other* categories). The fraction of females who hold the correct belief about their relative ability is much smaller than males in the US. According to a Kolmogorov-Smirnov test, the distributions of the confidence attitudes of males and females when they face a different gender partner are same in China ($p = 0.799$), while a gender difference exists in the US ($p = 0.008$), with males having more accurate beliefs on average.

Figures 9 and 10 illustrate females' and males' distributions of confidence by the gender of the partner, respectively. Kolmogorov-Smirnov tests fail to reject the hypotheses that the distributions of confidence are independent of the gender of the partner for both males and females. The p-values are 0.388 for females and 0.306 for males in China. The p-values in the US are 0.586 for females and 0.869 for males.

Among American men, if an individual is overconfident, he is willing to reveal Worse Performer status at a lower price. The Pearson's correlation coefficient between overconfidence and cutoff value is -0.1185 with a p-value of 0.0741. For females in the U.S, however, the Pearson's correlation coefficient is 0.0440, and the corresponding p-value 0.4980. In China, a negative correlation between overconfidence and cutoff value exists for females, but not for males. The Pearson's correlation coefficient is -0.1181 with $p = 0.0678$ for females, and -0.0358 with $p = 0.5811$ for males.

C The decision table for Periods 1 - 4 in China

决策序号	您去房间前方可以领取的安慰奖金的数额	您会去房间前方吗
1	¥1.0	□ 是 □ 否
2	¥1.5	□ 是 □ 否
3	¥2.0	□ 是 □ 否
4	¥3.0	□ 是 □ 否
5	¥4.0	□ 是 □ 否
6	¥4.5	□ 是 □ 否
7	¥5.0	□ 是 □ 否
8	¥6.0	□ 是 □ 否
9	¥7.0	□ 是 □ 否
10	¥7.5	□ 是 □ 否

Figure 11: The decision table for periods 1 - 4 in China, translation into English

D Instructions

General Instructions

Welcome to the experiment. Please read these instructions carefully. From now on, do not talk to your neighbors. Please turn off your mobile phone and keep it turned off until the end of the experiment. If you have any questions, raise your hand. We will then come to you.

The experiment is made up of two parts, Part A and B. We will hand out additional instructions at different points in the experiment.

Your earnings are the total of three payments:
(1) the show-up fee of $5;
(2) one period in Part A;
(3) and the amount you earn in part B.

Part A has six periods. We will roll a six-sided die at the end of the experiment to decide which period is counted into your final payment. Each period in Part A has an equal chance to be counted.

Instructions for Part A, Period 1 and 2

In each period, a female will be randomly matched with a male as a pair. The specific person paired with you will change randomly after each period. The pair assignment is anonymous, so you will not be told which of the participants from the other gender is matched with you in each period.

(For sessions where same gender composition in the first two periods, the description changes to: In each period, everyone will be paired with a partner of the same gender: a male will be randomly matched with a male as a pair, and a female is randomly paired with a female. The specific person paired with you will change randomly after each period. The pair assignment is anonymous, so you will not be told which of the participants from the same gender is matched with you in each period.)

The Task

In each period, all participants will perform a series of additions of five randomly-chosen two-digit numbers (e.g., 15 + 73 + 49 + 30+18). Calculators are not allowed. But you can use scratch paper for calculations. You will have four minutes to answer as many questions as possible. The computer will record the number of totals that you calculate correctly and the performance is measured by the number of totals you calculate correctly.

On the top right corner of the screen, you can see how many seconds you have left. Everyone in the room receives different randomly generated numbers to add. However, everyone faces on average the same level of difficulty.

Timing in Period 1 and 2

Stage 1

In each period, your performance in the mathematics addition task will be compared with your partner. If you outperform your partner, that is, if you have more correct answers than your partner does, you are the "Better Performer". If you answer fewer questions correctly, you are the "Worse Performer".

If you are a *"Worse Performer", you are eligible to claim consolation prizes.* To claim a prize, you must attend "The consolation prize-awarding ceremony" by going to the front of the room at the end of the experiment. The procedure of the ceremony will be described later. This is meant to make up for the fact that the *"Worse Performer"* earns less than their partner because he/she did not perform as well.

Before you do the mathematics addition task, we ask each of you to **make 10 decisions** in a table, which will be shown on your computer screen. You are asked to indicate **whether you are willing to claim possible consolation prizes of different values** by going to the front of the room *if you are a "Worse Performer"*. The decision table will look like this:

Decision number	The value of the consolation prize you can claim if you go to the front of the room	Will you go to the front of the room to claim the prize?
1	$0.25	□ Yes □ No
2	$0.50	□ Yes □ No
3	$0.75	□ Yes □ No
4	$1.00	□ Yes □ No
5	$1.25	□ Yes □ No
6	$1.50	□ Yes □ No
7	$1.75	□ Yes □ No
8	$2.00	□ Yes □ No
9	$2.25	□ Yes □ No
10	$2.5	□ Yes □ No

The first column is the decision number, the second column is a list of the value of the consolation prize you can claim if you go to the front of the room, and the third column is the choice you need to make. For example, the first line is asking you: If the prize is $0.25, and if you are a "Worse

Performer", are you willing to go to the front of the room to claim this prize? **Please make a choice in each line.**

At the end of the experiment, only Worse Performers are eligible to claim the consolation prizes. This means, your decisions on attending "The consolation prize awarding ceremony" will be counted if it turns out that you are a worse performer in your pair after the task. If you are a "Better Performer", your decisions won't be counted. And you will only know whether you are a Worse Performer or a Better Performer at the end of the experiment.

At the end of the experiment, we roll a six-sided die that decides which period is chosen to be counted towards your payment. If either period 1 or period 2 is chosen, we will roll a ten-sided die to determine which line of the decision in that period is implemented. And we will have the consolation prize awarding ceremony.

The consolation prize-awarding ceremony

At the end of the experiment, you will know whether you are a "Better Performer" or a "Worse Performer" in the selected period. Participants will stand in two groups. *Those participants who are "Worse Performers" and choose "Yes"* in the randomly selected line will be asked to go to the front of the room to claim a redemption code of the prize ***in full view of other participants***. And the information "In a mathematics addition task, these participants performed worse than their partner of a different gender: The males in the front lost to their female partner, and the females lost to their male partner" will be shown on all participants' screens. After 30 seconds, they will return to sitting and input the code into the corresponding box on computer.

The remaining participants, including *those who are "Worse Performers" but choose "No"* in the randomly selected line, as well as the *"Better Performer"* will then be asked to stand for 30 seconds. Then they will return to sitting and the experiment will end at that point.

Stage 2

After you make ten decisions, you will be asked to perform the mathematics addition task.

Stage 3

After you perform the task, we ask you to guess your performance in comparison with your partner. You are to guess whether you performed at least as well as your partner. If you guess correctly, you can earn additional $0.50.

Your earnings:

If either period 1 or 2 in Part A ends up counting for your final earnings, then:

1. If you are a *"Worse Performer"*, your final earnings equal:

Your show-up fee of **$5** + Earnings from doing the addition task, which are **$5** + Your consolation prize + Earnings from the guess of the performance comparison result + Your earnings in Part B (to be described later)

If period 1 or 2 is chosen to count for your payment, the ten-sided die is rolled, to determine your payment. For example, if the realization of the ten-sided die is 2, it means the decision in line 2 you made in the selected will be implemented. If you chose "Yes" in line 2, you will have $0.50 added to your payment by attending the ceremony. If you choose "No" in line 2, you don't need to go to the front as a Worse Performer, but you receive no consolation prize.

2. If you are a *"Better Performer"*, your final earnings equal:

Your show-up fee of **$5** + Earnings from doing the addition task, which are **$10** + Earnings from the guess of the performance comparison result + Your earnings in Part B (to be described later)

Practice Period

To help familiarize yourself with the interface and the task, you will do one practice round of the addition task. This period is not counted towards your final payment.

Instructions for Part A, Periods 3 and 4
(The instructions for these periods are given at the end of period 2.)

You have now completed 2 periods. In next block of 2 periods, the entire process and the payment decision rule will be the same as in Period 1 and 2, **except that now you are paired with a partner of a same gender.** (*In the same gender group first sessions, the description changes to:* **except that now you are paired with a partner of a different gender.**)

If a period in Period 3 or 4 is selected to decide your payment in Part A at the end of the experiment, and if you are a "Worse Performer" and choose "Yes" in the randomly selected line, the information on all participants' screens will change to "In a mathematics addition task, these participants performed worse than their partner of a same gender: The male in the front lost to his male partner, and the female lost to her female partner". (*In the same gender group first sessions, the description changes to:* "In a mathematics addition task, these participants performed worse than their partner of a different gender: The male in the front lost to his female partner, and the female lost to her male partner".)

Instructions for Part A, Period 5
(The instruction for this period is given at the end of period 4.)

In Period 5, one male and one female will be randomly matched to be a pair. (*In the same gender group first sessions, the description changes to:* one male and one male will be randomly matched to be a pair, one female and one female will be randomly matched to be a pair.)

The timing
Stage 1:

You will still perform a series of additions of five randomly-chosen two-digit numbers. Everyone is asked to calculate 10 totals correctly with no time limit.

Stage 2
After you have 10 correct totals, you can choose to quit or to continue the task for another three minutes. And your partner faces the same decision.

However, either your choice or your partner's choice will count and be implemented.

If your choice counts, your partner's performance will stay at 10. If you choose to quit, your likelihood of being designated as the Worse Performer is 50%. If you choose to continue, you can calculate as many totals as you want in three minutes. If you calculate 0 correct totals in three minutes, your likelihood of being designated as the Worse Performer is 50%.

If your partner's choice counts, your performance will stay at 10.

Your performance will be compared with your partner's. And the Worse Performer will be asked to stand in front of the room.

Stage 3

After you make the decision, the computer will randomly decide whether your choice or your partner's choice in Stage 2 counts.

If your choice counts, and you choose to continue, then you will be given three minutes more to do the addition task.

If your choice counts, and you choose to quit, then you will proceed to Stage 4.

If your partner's choice counts, then you will proceed to Stage 4.

Stage 4

You will be informed of the performance result in this period.

If your choice counts, and you choose to continue, your performance is: 10+number of totals you calculate correctly in the additional three minutes.

If your choice counts, and you choose to quit, your performance is 10.

If your partner's choice counts, your performance will stay at 10.

At the end of the experiment, if Period 5 is chosen to be counted into your final earnings, all participants will stand in the front of the room in two groups. First, those participants who are Worse Performers will stand in front of the room for 30 seconds. Then, they will return to seats

and all the participants who are Better Performers will go to the front of the room and stand for 30 seconds.

Your payment
Your payment in this period is fixed. You will earn $8 from this period. If this period is chosen to be counted to your earnings. Then your final earnings equal:

Your show up fee $5 + Payment in period 5 **$8** + Your payment in part B (described later)

Instructions for Part A, Period 6
(The instruction for this period are given at the end of period 5.)

You have now completed Period 5. In Period 6, the entire process and the payment decision rule will be the same as in Period 5, **except that now you are paired with a partner of the same gender.** (*In the same gender group first sessions, the description changes to:* **except that now you are paired with a partner of a different gender.**)

Instruction for Part B (Period 7)
(The instruction for this period are given at the end of period 6.)

In this part of the experiment, there are two decisions you need to make.

Please answer both questions on the screen. Please indicate a single choice for each question.

1. We had a group of men and a group of women do the same addition task as you did in periods 1 – 4, yesterday. One of the two groups performed better. Please guess which of the following is correct.

- Men performed better than women on average
- Women performed better than men on average

2. Please guess the proportion of people in this session agree with your choice above
- Less than or equal to 25%
- 26%-49%
- 50%
- 51%-74%
- More than or equal to 75%

Your payment in Part B depends on the accuracy of your response to question 1 and 2. If you answer both correctly, you will receive $2 for part B. If you answer exactly one question correctly, you will receive $1. Otherwise, you will receive nothing for this part of the experiment.

References

J. Andreoni and B. D. Bernheim. Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636, 2009.

J. Andreoni and L. Vesterlund. Which is the fair sex? gender differences in altruism. *The Quarterly Journal of Economics*, 116(1):293–312, 2001.

J. Aronson, M. J. Lustina, C. Good, K. Keough, C. M. Steele, and J. Brown. When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, 35(1):29–46, 1999.

L. Babcock, S. Laschever, M. Gelfand, and D. Small. Nice girls don't ask. *Harvard Business Review*, 81(10):14–16, 2003.

L. Babcock, M. P. Recalde, L. Vesterlund, and L. Weingart. Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3):714–47, 2017.

F. D. Barth. Do men have a problem with dating smart women? Jan 23, 2016. URL `https://www.psychologytoday.com/us/blog/the-couch/201601/do-men-have-problem-dating-smart-women?page=2`.

M. Bertrand. New perspectives on gender. In *Handbook of Labor Economics*, volume 4, pages 1543–1590. Elsevier, 2011.

M. Bertrand, E. Kamenica, and J. Pan. Gender identity and relative income within households. *The Quarterly Journal of Economics*, 130(2):571–614, 2015.

P. Bordalo, K. Coffman, N. Gennaioli, and A. Shleifer. Beliefs about gender. *American Economic Review*, 109(3):739–73, 2019.

R. P. Burriss. Men's self-esteem boosted by female pheromone. May 31, 2016. URL `https://www.psychologytoday.com/intl/blog/attraction-evolved/201605/mens-self-esteem-boosted-female-pheromone?amp`.

L. Bursztyn, T. Fujiwara, and A. Pallais. 'acting wife': Marriage market incentives and labor market investments. *American Economic Review*, 107(11):3288–3319, 2017.

L. Bursztyn, G. Egorov, and R. Jensen. Cool to be smart or smart to be cool? understanding peer pressure in education. *The Review of Economic Studies*, 86 (4):1487–1526, 2019.

A. W. Cappelen, R. Falch, and B. Tungodden. The boy crisis: Experimental evidence on the acceptance of males falling behind. *NHH Dept. of Economics Discussion Paper*, (06), 2019.

S. E. Carrell, M. E. Page, and J. E. West. Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, 125(3):1101–1144, 2010.

S. J. Correll. Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology*, 106(6):1691–1730, 2001.

R. Croson and U. Gneezy. Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74, 2009.

B. Dong. A study on the influence of gender stereotype in mathematics on mathematics achievement of junior middle school students. *East China Normal University Working Paper*, 2019.

C. C. Eckel and P. J. Grossman. Are women less selfish than men?: Evidence from dictator experiments. *The Economic Journal*, 108(448):726–735, 1998.

C. C. Eckel and P. J. Grossman. Chivalry and solidarity in ultimatum games. *Economic Inquiry*, 39(2):171–188, 2001.

C. C. Eckel and P. J. Grossman. Men, women and risk aversion: Experimental evidence. *Handbook of Experimental Economics Results*, 1:1061–1073, 2008.

N. M. Else-Quest, J. S. Hyde, and M. C. Linn. Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological Bulletin*, 136(1):103, 2010.

U. Fischbacher. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178, 2007.

R. Fisman, S. S. Iyengar, E. Kamenica, and I. Simonson. Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics*, 121(2):673–697, 2006.

O. Folke and J. Rickne. All the single ladies: Job promotions and the durability of marriage. *American Economic Journal: Applied Economics*, 12(1):260–87, 2020.

M. Foschi. Double standards for competence: Theory and research. *Annual Review of Sociology*, 26(1):21–42, 2000.

Q. Fottrell. Men are more likely to leave jobs if their boss is a woman. Dec 7, 2018. URL `https://www.marketwatch.com/story/men-are-more-likely-to-leave-jobs-if-their-boss-is-a-woman-2018-11-19`.

U. Gneezy and A. Rustichini. Gender and competition at a young age. *American Economic Review*, 94(2):377–381, 2004.

U. Gneezy, M. Niederle, and A. Rustichini. Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074, 2003.

Z. Grossman. Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization*, 117:26–39, 2015.

L. Guiso, F. Monte, P. Sapienza, and L. Zingales. Culture, gender, and math. *Science*, 320(5880):1164–1165, 2008.

G. Hofstede. *Masculinity and femininity: The taboo dimension of national cultures*, volume 3. Sage Publications, 1998.

G. Hofstede. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001.

A. N. Husain, D. A. Matsa, and A. R. Miller. Do male workers prefer male leaders? an analysis of principals' effects on teacher retention. Technical report, National Bureau of Economic Research, 2018.

A. Karbowski, D. Deja, and M. Zawisza. Perceived female intelligence as economic bad in partner choice. *Personality and Individual Differences*, 102:217–222, 2016.

O. L. Liu and M. Wilson. Gender differences and similarities in pisa 2003 mathematics: A comparison between the united states and hong kong. *International Journal of Testing*, 9(1):20–40, 2009.

S. D. Mago and L. Razzolini. Best-of-five contest: An experiment on gender differences. *Journal of Economic Behavior & Organization*, 162:164–187, 2019.

D. Masclet, C. Noussair, S. Tucker, and M.-C. Villeval. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, 93(1):366–380, 2003.

T. C. McManus and J. M. Rao. Signaling smarts? revealed preferences for self and social perceptions of intelligence. *Journal of Economic Behavior & Organization*, 110:106–118, 2015.

D. I. Miller, A. H. Eagly, and M. C. Linn. Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, 107(3):631, 2015.

M. Murray-Close and M. L. Heggeness. Manning up and womaning down: How husbands and wives report their earnings when she earns more. *US Census Bureau Social, Economic, and Housing Statistics Division Working Paper*, (2018-20), 2018.

M. Niederle and L. Vesterlund. Do women shy away from competition? do men

compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101, 2007.

L. E. Park, A. F. Young, J. D. Troisi, and R. T. Pinkus. Effects of everyday romantic goal pursuit on women's attitudes toward math and science. *Personality and Social Psychology Bulletin*, 37(9):1259–1273, 2011.

L. E. Park, A. F. Young, and P. W. Eastwick. (psychological) distance makes the heart grow fonder: Effects of psychological distance and relative intelligence on men's attraction to women. *Personality and Social Psychology Bulletin*, 41(11): 1459–1473, 2015.

A. Patnaik. Reserving time for daddy: the consequences of fathers' quotas. *Journal of Labor Economics*, 37(4):1009–1059, 2019.

K. A. Ratliff and S. Oishi. Gender differences in implicit self-esteem following a romantic partner's success or failure. *Journal of Personality and Social Psychology*, 105(4):688, 2013.

E. Reuben, P. Sapienza, and L. Zingales. How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408, 2014.

C. L. Ridgeway. Interaction and the conservation of gender inequality: Considering employment. *American Sociological Review*, pages 218–235, 1997.

C. L. Ridgeway and S. J. Correll. Limiting inequality through interaction: The end (s) of gender. *Contemporary Sociology*, 29(1):110–120, 2000.

S. J. Spencer, C. M. Steele, and D. M. Quinn. Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35(1):4–28, 1999.

J. Syrda. Spousal relative income and male psychological distress. *Personality and Social Psychology Bulletin*, 46(6):976–992, 2020.

J. Tiedemann. Gender-related beliefs of teachers in elementary school mathematics. *Educational Studies in Mathematics*, 41(2):191–207, 2000.

M. Tsui. Gender and mathematics achievement in china and the united states. *Gender Issues*, 24(3):1–11, 2007.

F. J. Van de Vijver. Cultural and gender differences in gender-role beliefs, sharing household task and child-care responsibilities, and well-being among immigrants and majority members in the netherlands. *Sex Roles*, 57(11-12):813–824, 2007.