Supplementary Materials for

Opt-out choice framing attenuates gender differences in the decision to compete in the lab and in the field

Joyce C. He, Sonia K. Kang, Nicola Lacetera

Correspondence to: nicola.lacetera@utoronto.ca

This PDF file includes:

Detailed Methods from Study 1

Participants

Participants were 482 undergraduate students from a large Canadian university (55.4% women; M_{age} = 19, SD = 1.64; ethnicity: 67.4% Asian, 19.9% Caucasian, 1.45% Hispanic/Latino, 1.24% African American, 9.96% indicated "other"). Participants received one course credit for participating and earned financial compensation from one randomly selected stage of the task.

Procedure

Our experiment largely follows the paradigm in Niederle and Versterlund[1], which examined gender differences in overconfidence and preference for competition. We used zTree (version 4.0) to program the experiment[2]. The experimental task consisted of adding five two-digit numbers[1]. Participants could not use a calculator, but were provided scrap paper and had up to 5 minutes to complete as many questions as they could. At the end of each stage, participants saw their own final score (i.e., the number of correct answers) for that stage. In each stage, participants saw only their own absolute score and did not see their relative performance until the end of the experiment. Participants completed three stages of this same task. Following Niederle and Vesterlund[1] as well as many other similar experiments, participants were informed that their monetary rewards would depend on their performance in one of the four stages, randomly determined; this would ensure that participants had high powered incentives in all stages. The specific compensation scheme, however, was different for each stage as described below.

Stage 1 – Piece Rate: If stage 1 was selected for compensation, participants would receive $0.50 for each correct answer.

Stage 2 – Tournament: If stage 2 was selected for payment, the focal participant's score for that stage would be compared to three other randomly chosen competitors' scores. If the participant

held the highest score in that stage compared to the other three competitors, they would receive $2 per correct answer. If they did not hold the highest score, they would receive $0. In the case of a tie, the winner was chosen randomly among the high scorers. Note that the expected payoffs for the tournament and piece-rate compensation are equal. Note also that the participants for a given group of four were randomly selected with replacement. Therefore, a given individual could be in the comparison group for more than one focal participant.

Stage 3 – Choice: Before proceeding to the task, participants were asked to choose their compensation scheme for the addition task. In the original design of the experiment[1], participants were told that they could choose piece-rate (50 cents per correct answer) or tournament ($2 per correct answer if the focal participant's score exceeded that of the other group members in the stage two tournament; winners are chosen randomly in the case of a tie). Here, we administered our central manipulation – whether the choice to enter the competitive environment (the tournament payment scheme) was framed using opt-in or opt-out framing. Participants were randomly assigned to either an opt-in or opt-out condition.

Opt In Framing Condition

In the opt-in framing condition, participants were told that by default, if stage 3 was randomly selected for payment, they would receive $0.50 per correct answer. In other words, the default was the non-competitive, piece-rate compensation. Participants could choose instead to opt in to the competitive, tournament compensation scheme. Further, they were told that if they chose to compete, their stage 3 performance would be compared with the stage 2 performance of the three other participants. We chose to compare against stage 2 performance to avoid instances where not all competitors chose to compete for stage 3. Niederle and Vesterlund[1] also note that by comparing the performance of a focal participant to the correct answers of three participants in the

previous stage, one can identify preferences for being compared separately from preferences for direct competition. If a participant wanted to be compensated according to the tournament scheme, they were asked to check a box to indicate this. Otherwise, they just had to press the next button to proceed to the next page.

Opt Out Framing Condition

In the opt-out framing condition, participants were told that by default, if stage 3 was randomly selected for payment, their performance would be compared to the same three participants' from the previous stage, and if they received the highest score they would receive $2 per correct answer and $0 if they did not receive the highest score. In other words, the default was the competitive, tournament compensation. Participants could choose to opt out of the tournament and return to the non-competitive, piece-rate compensation scheme. If they wanted to opt out of the tournament compensation scheme, they had to check a box to indicate this. Otherwise, they could press the next button to proceed to the next page.

Stage 4 – Choice: In the final stage of the experiment, participants were told that they could re-submit their stage 1 performance for compensation. They were given the choice to submit their stage 1 performance to either the piece-rate compensation or a tournament compensation, where their performance would be compared to three other participants' stage 1 performance. Note therefore that there was not an actual additional task in stage 4.

After the 4 stages, participants were asked to guess their rank (for stages 1 and 2) compared to others against whom they were competing. We asked for this information so that we could compare their guessed rank to their actual rank to obtain a measure of (over)confidence. Participants guessed their rank (1=best, 2=second best, 3=third best, or 4=fourth best) in the stage 1 piece rate scheme as well as the stage 2 tournament scheme.

Finally, participants completed a six-item version of the State Anxiety Inventory (SAI) [3,4]. The six-item version of the SAI has shown to be highly correlated with the full version of the scale, and has been shown to have high internal consistency (alphas above .90) [3,5]. We asked participants to read the statements and indicate how they felt during the experiment, on a scale from 1 (*not at all*) to 4 (*very much*). Sample items include "During the experiment, I felt calm" (reverse scored), "During the experiment, I was tense", and "During the experiment, I felt upset". In our sample, the six-item scale had an internal reliability of a = .68. Upon closer examination, it appeared that the item "I felt content" had poor item-total correlation compared to the rest of the other items. As such, we removed the item from the scale[6]. The five-item scale had good reliability (a = .73).

Table S1. Number (and share) of payoff-maximizing choices by gender, choice and condition

	Number of participants		Percent of payoff maximizing choices	
	By condition-choice-gender	By condition-gender	By condition-choice-gender	By condition-gender
Opt in, piece rate: Men	30	109	76.7%	56.0%
Opt in, tournament: Men	79		48.1%	
Opt out, piece rate: Men	25	106	92.0%	53.8%
Opt out, tournament: Men	81		42.0%	
Opt in, piece rate: Women	73	137	79.5%	59.1%
Opt in, tournament: Women	64		35.9%	
Opt out, piece rate: Women	32	130	78.1%	56.2%
Opt out, tournament: Women	98		49.0%	

Notes: The table reports the number and percentage, of subjects who made the payoff maximizing choice for them in Stage 3 by experimental condition, gender, and choice of compensation scheme. We established the payoff maximizing choice as follows. We used the estimated coefficients from a probit regression of whether a participant won their tournament in Stage 2 on the number of correct responses in that stage, to predict the likelihood of winning a tournament in Stage 3 given their performance in Stage 3 (recall that individuals who chose the tournament in Stage 3 had their performance compared against three participants from Stage 2). We then calculated the *expected* payoff from choosing a tournament or a piece rate compensation for each participant. For the piece rate, the expected payoff was \$0.50 X the number of correct responses in Stage 3. The expected payoff from a tournament was \$2 X the number of correct responses in Stage 3 X the predicted probability of winning the tournament for that subject in Stage 3. We classify a participant as having made their payoff-maximizing choice if they selected the compensation scheme that gave them the higher expected payoff. In cases where the expected payoffs from the two compensation scheme were close to each other (less than \$1 in absolute difference; this happened if a participant solved 12 questions correctly), we randomized the assignment to having made the payoff maximizing decision or not (we interpreted small differences as making a participant indifferent between the two schemes).

Table S2. Average monetary gains (losses) from choosing a compensation scheme in Stage 3, compared to "counterfactual" choice

	Sum of net gains	Avg. net gains	N
Gender			
Men	$790.5	$3.68	215
Women	$746.0	$2.79	267
Choice architecture			
Opt in	$767.5	$3.12	246
Opt out	$769.0	$3.26	236
Conditions in stage 3, by gender			
Opt in: Men	$465.0	$4.27	109
Opt in: Women	$302.5	$2.21	137
Opt out: Men	$325.5	$3.07	106
Opt out: Women	$443.5	$3.41	130
Conditions and compensation choices in stage 3, by gender			
Opt in, piece rate: Men	-$20.0	-$0.67	30
Opt out, piece rate: Men	$85.0	$3.40	25
Opt in, piece rate: Women	$68.5	$0.94	73
Opt out, piece rate: Women	$34.0	$1.06	32
Opt in, tournament: Men	$485.0	$6.14	79
Opt out, tournament: Men	$240.5	$2.97	81
Opt in, tournament: Women	$234.0	$3.66	64
Opt out, tournament: Women	$409.5	$4.18	98

Notes: this table reports both the total gains (over all participants) and the average gains per participants from choosing a compensation scheme over the alternative one. For the participants who selected a tournament-based compensation scheme, the "counterfactual" payoff is the number of correct responses that they gave in Stage 3 multiplied by $0.5. For the participants who selected piece rate, the counterfactual payoff is zero if they would have not won the tournament in the group to which they would be assigned, and equal to the number of their correct answers multiplied by $2 had they been the winners of the groups to which they were assigned.

Table S3. Additional specification for compensation choice regressions: interactions (1).

	(1)	(2)
Outcome variable:	Choice of tournament in Stage 3	
Estimation:	Probit	
Sample:	Opt-in condition	Opt-out condition
Woman	-0.114	0.025
	(0.236)	(0.159)
# correct answ. in stage 2	0.031*	0.018*
	(0.018)	(0.011)
Woman X # correct answ. in stage 2	-0.014	-0.004
	(0.020)	(0.014)
Observations	246	236
Pseudo R2	0.074	0.017

Notes: The outcome variable is a binary indicator equal to 1 if a participant selected tournament-based compensation in stage 3, and 0 if they selected piece rate compensation. Regressions are separated by experimental condition (opt-in vs. opt-out frame). Regressors include participant gender (the omitted category is men), the number of correct responses in stage 2, and the interaction between the gender indicator and the number of correct responses. Estimated standard errors, clustered at the session level (there were 36 sessions) are in parentheses. * p<0.1, ** p<05, *** p<01.

Table S4. Additional specification for compensation choice regressions: interactions (2).

	(1)
Outcome variable:	Choice of tournament in Stage 3
Estimation:	Probit
Opt in: Woman	-0.380
	(0.283)
Opt out: Man	0.023
	(0.320)
Opt out: Woman	-0.314
	(0.312)
# correct answ. in stage 2	0.001
	(0.021)
# correct in stage 2 - # correct in stage 1	-0.013
	(0.020)
Guessed rank in stage 2 tournament	-0.199***
	(0.061)
# correct answ. in stage 2	0.006
x Opt in: Woman	(0.023)
# correct answ. in stage 2	-0.003
x Opt out: Man	(0.028)
# correct answ. in stage 2	0.006
x Opt out: Woman	(0.025)
# correct in stage 2 - # correct in stage 1	-0.006
x Opt in: Woman	(0.030)
# correct in stage 2 - # correct in stage 1	-0.005
x Opt out: Man	(0.033)
# correct in stage 2 - # correct in stage 1	0.037
x Opt out: Woman	(0.027)
Guessed rank in stage 2 tournament	0.063
x Opt in: Woman	(0.064)
Guessed rank in stage 2 tournament	0.030
x Opt out: Man	(0.076)
Guessed rank in stage 2 tournament	0.176**
x Opt out: Woman	(0.086)
Observations	482
Pseudo R2	0.129

Notes: The outcome variable is a binary indicator equal to 1 if a participant selected tournament-based compensation in stage 3, and 0 if they selected piece rate compensation. Regressors include the experimental conditions (opt-in vs. opt-out frame) interacted with the gender of the participant (the omitted category is men in the opt-in condition); the number of correct responses in stage 2, and the difference between the correct answers in stage 2 and those in stage 1; the positions that each participant guessed to have achieved in the tournament in stage 2 (out of four position, rank 1 being the winner); and interaction of the gender of the participant with correct responses in Stage 2, and the difference between the correct answers in stage 1 and those in Stage 1; the positions that each participant guessed to have achieved in the tournament in stage 2. The estimates indicate marginal effects from probit regressions, where the baseline is a male participant in the opt-in condition, with 8.52 correct answers in Stage 1, 10.24 correct answers in Stage 2, a guess for their rank in stage 2 of 24, and a confidence (actual rank – guessed rank) of 0.31. Estimated standard errors, clustered at the session level (there were 36 sessions) are in parentheses. * p<0.1, ** p<05, *** p<01.

Table S5. Correct responses in stage 3 and anxiety levels: Regression estimates

Outcome variable:	# of correct answers in stage 3						Average anxiety					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Woman	-0.137						0.100**					
	(0.341)						(0.041)					
Conditions in stage 3, by gender												
Opt in: Woman		0.187						0.074				
		(0.293)						(0.052)				
Opt out: Man		-0.303						-0.030				
		(0.282)						(0.064)				
Opt out: Woman		-0.190						0.083				
		(0.322)						(0.062)				
Conditions and compensation choices in stage 3, by gender												
Opt in, tournament: Man			1.639*	-0.038	-0.168	-0.084			-0.094	-0.056	-0.054	-0.061
			(0.904)	(0.450)	(0.442)	(0.440)			(0.103)	(0.107)	(0.109)	(0.107)
Opt out, piece rate: Man			-1.573*	-1.456**	-1.508**	-1.496**			-0.151	-0.153	-0.152	-0.157
			(0.841)	(0.589)	(0.592)	(0.589)			(0.115)	(0.118)	(0.118)	(0.119)
Opt out, tournament: Man			0.854	-0.110	-0.252	-0.183			-0.065	-0.042	-0.040	-0.050
			(0.779)	(0.439)	(0.448)	(0.440)			(0.126)	(0.125)	(0.127)	(0.128)
Opt in, piece rate: Woman			-0.133	-0.120	-0.166	-0.150			0.029	0.029	0.030	0.026
			(0.722)	(0.535)	(0.539)	(0.531)			(0.105)	(0.108)	(0.109)	(0.109)
Opt in, tournament: Woman			1.367	0.495	0.380	0.451			0.017	0.037	0.039	0.032
			(0.851)	(0.543)	(0.544)	(0.543)			(0.091)	(0.094)	(0.095)	(0.096)
Opt out, piece rate: Woman			-0.633	-0.843	-0.918*	-0.863*			-0.070	-0.065	-0.064	-0.067
			(0.956)	(0.503)	(0.505)	(0.505)			(0.106)	(0.108)	(0.107)	(0.110)
Opt out, tournament: Woman			1.061	0.033	-0.061	-0.010			0.053	0.075	0.077	0.071
			(0.916)	(0.524)	(0.519)	(0.515)			(0.125)	(0.123)	(0.122)	(0.125)
# correct answ. in stage 2		0.867***		0.910***	0.889***	0.920***		-0.019***		-0.021***	-0.021***	-0.020***
		(0.036)		(0.037)	(0.039)	(0.039)		(0.005)		(0.006)	(0.007)	(0.006)
# correct in stage 2 - # correct in stage 1				-0.194***	-0.200***	-0.198***				0.006	0.006	0.006
				(0.067)	(0.068)	(0.067)				(0.007)	(0.007)	(0.007)
Guessed rank in stage 2 tournament					-0.174						0.003	
					(0.113)						(0.028)	
(Over)confidence						0.112						0.012
						(0.081)						(0.020)
Constant	10.874***	1.984***	10.133***	1.913***	2.587***	1.830***	2.050***	2.258***	2.127***	2.312***	2.301***	2.304***
	(0.258)	(0.503)	(0.683)	(0.611)	(0.726)	(0.648)	(0.034)	(0.072)	(0.090)	(0.112)	(0.154)	(0.112)
Observations	482	482	482	482	482	482	482	482	482	482	482	482
Pseudo R2	0.000	0.682	0.049	0.704	0.705	0.705	0.011	0.033	0.018	0.042	0.042	0.042

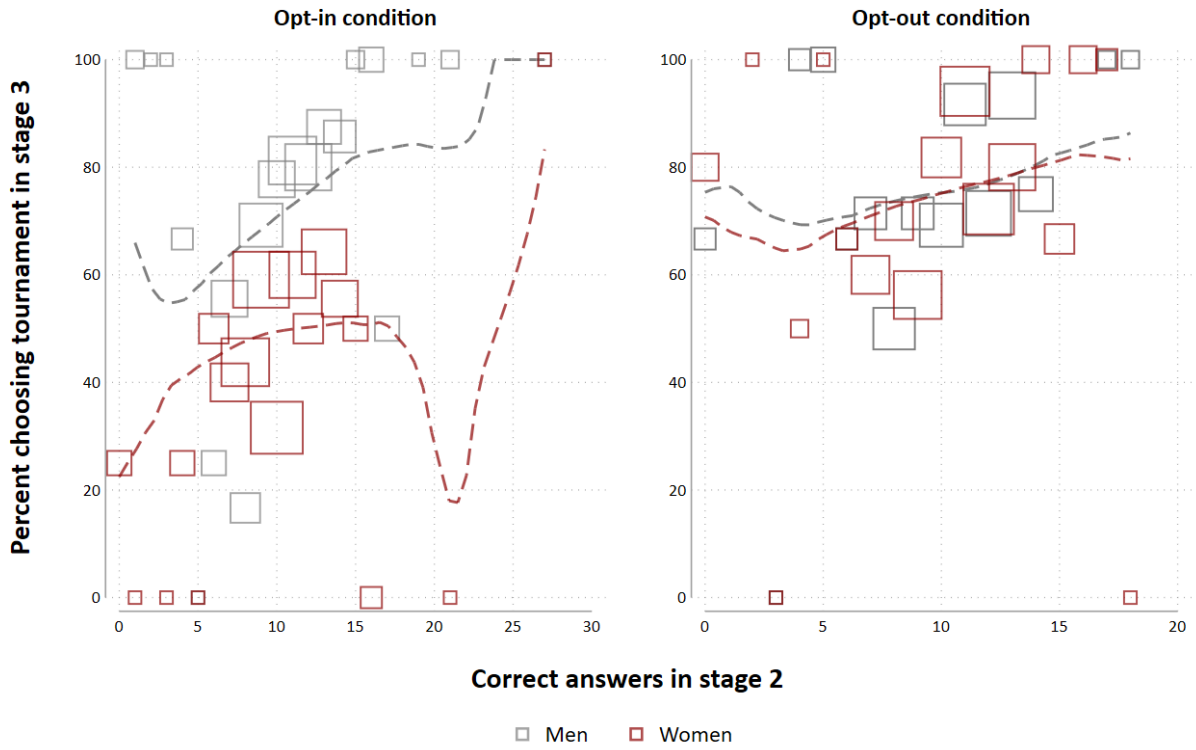
Notes: The table reports parameter estimates for linear regression models where the outcomes variables are the number of correct responses in stage 3 by each participants (columns 1 through 6) and the anxiety index (columns 7 through 12). Regressors include the experimental conditions (opt-in vs. opt-out frame) interacted with the gender of the participant (the omitted category is men in the opt-in condition); the experimental conditions (opt-in vs. opt-out frame) interacted with the gender of the participant and the choice of compensation scheme (the omitted category is men in the opt-in condition who chose piece rate); the number of correct responses in Stage 2, and the difference between the correct answers in Stage 1 and those in Stage 1; the positions that each participant guessed to have achieved in the tournament in Stage 2 (out of four position, rank 1 being the winner); and the difference between the actual position and the guessed position, as a measure of (over) confidence. The average number of correct responses and the average anxiety index for men are 10.87 and 1.56, respectively; for men in the opt in condition are 11.32 and 1.58, respectively; the average number of correct responses and the average anxiety index for men in the opt in condition who chose piece rate compensation are 10.13 and 1.97, respectively. Estimated standard errors, clustered at the session level (there were 36 sessions) are in parentheses. * p<0.1, ** p<05, *** p<01 (two-sided tests).

Table S6. Guessed rank in stage 2, and (over)confidence

Outcome variable:	Guessed rank for stage 2	(Over)Confidence
	(1)	(2)
Woman	0.215**	-0.104
	(0.085)	(0.088)
# correct answ. in stage 2	-0.182***	-0.077***
	(0.019)	(0.014)
# correct in stage 2 - # correct in stage 1	-0.038	0.029*
	(0.025)	(0.015)
Observations	482	482
Pseudo R2	0.131	0.021

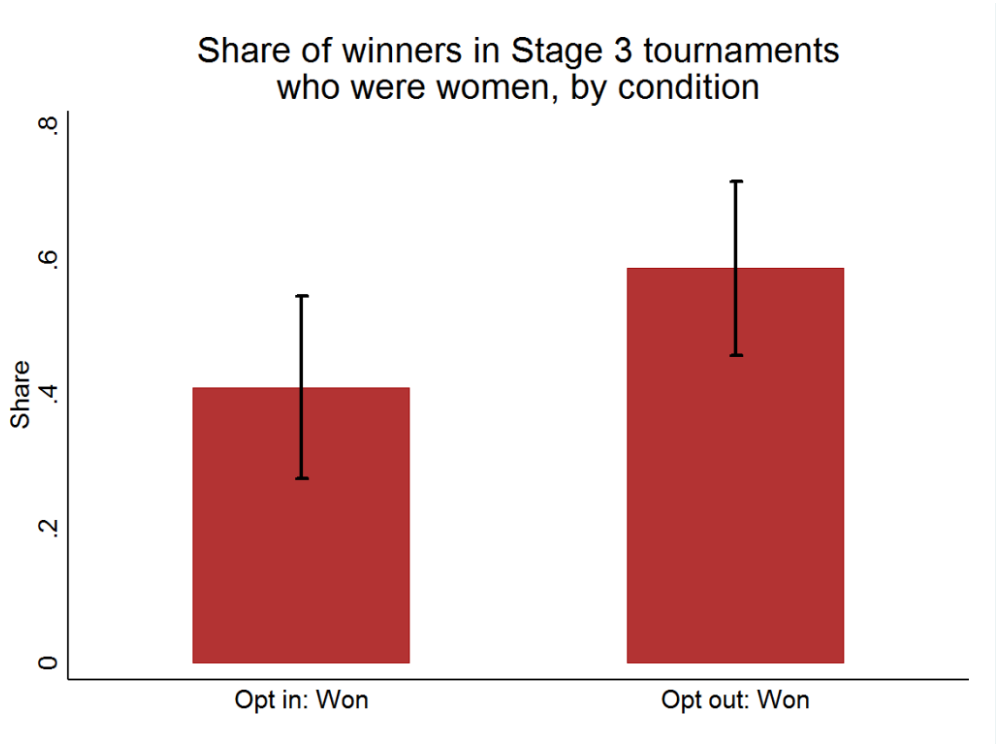
Notes: The table reports parameter estimates from ordered probit regressions where the outcome variable is the guessed tournament rank in stage 2 (column 1) and the difference between the actual rank and the guessed rank (column 2), and the regressors are the gender of the respondent, the number of correct responses in stage 2, and the difference between the number of correct responses in stage 2 and stage 1. Estimated standard errors, clustered at the session level (there were 36 sessions) are in parentheses. * p<0.1, ** p<05, *** p<01 (two-sided tests).

Fig S1: Choice of compensation in stage 3 by correct answers in stage 2



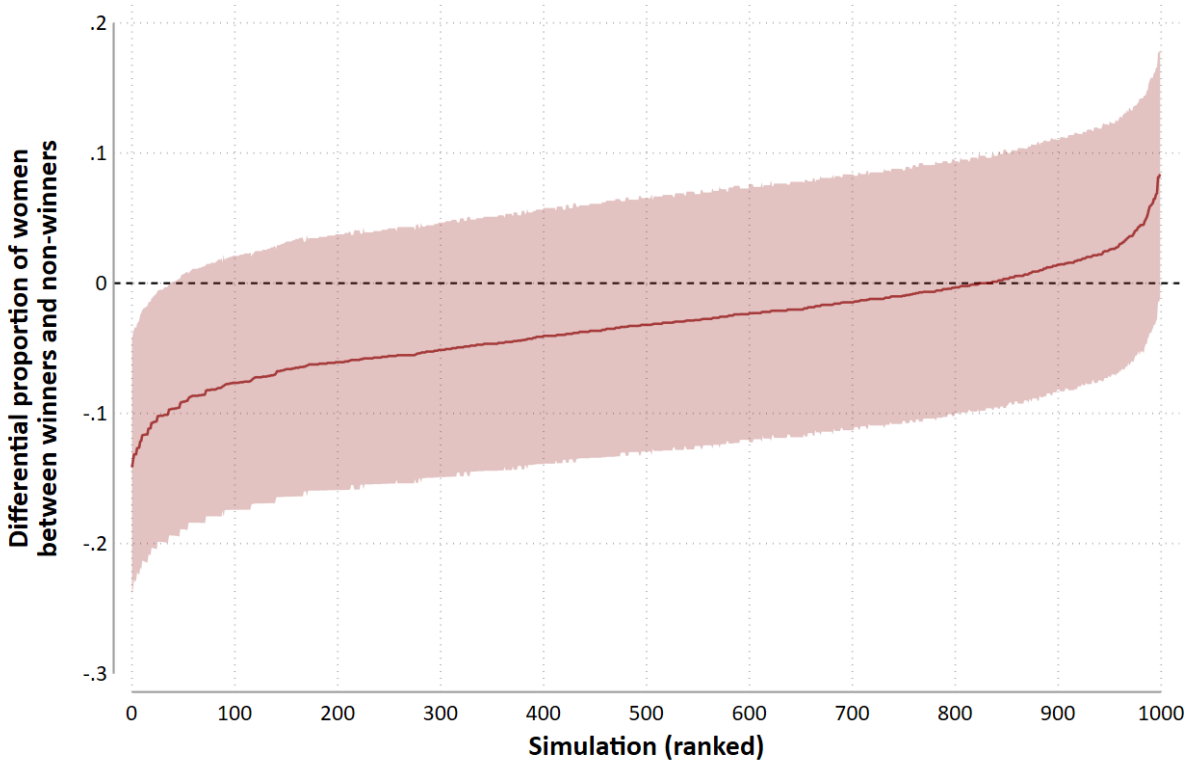
Notes: The x-axes in the graphs report the number of correct responses in stage 2. Each square represents the share of respondents (separated by gender) who chose tournament compensation in stage 3, for each number of correct responses in stage 2. The size of the circles is proportional to the number of participants, by gender and experimental condition, who attempted a given number of tasks, relative to the total number of participants of a given gender in a given condition. The dashed lines are smooth polynomial approximations of the relationship between number of correct answers in stage 2 and the likelihood of choosing tournament compensation in stage 3 (degree zero, bandwidth 3), separate by gender.

Fig. S2. Proportion of tournament winners in stage 3 who were women



Notes: The graph displays the percentage of tournament winners in Stage 3 (among all those who chose a tournament) who were women, separately for the opt-in and the opt-out condition.

Fig. S3. Simulated differences in the proportion of women between winners and non-winners of a tournament in Stage 2



Notes. The red solid line reports the estimated difference between the average proportion of tournament winners who were women, and the average proportion of tournament "non-winner" who were women, from each of 1,000 simulations where we randomly assigned each participant to a fictitious group and defined the winner as the participant(s) with the highest number of correct answers in stage 2. A negative value, for example, indicates that for a given simulation, the proportion of non-winners who were women (e.g. 53%) was higher than the proportion of winner who were women (e.g. 51%). The shaded area represents the confidence intervals around each of the 1,000 estimated average differences. Note that these randomly created groups are not the ones within which a participant actually competed in the experiment. Moreover, unlike the groups to which each participant was assigned in the experiment, in these simulations the groups are without replacement, i.e. participants in a given group cannot also be in another group. Each of the 1,000 simulations re-shuffled the groups. The purpose of this exercise is to further test for the presence of underlying (dis)advantages of women (or men) in being the winner of the tournaments in stage 2, based on their actual absolute performance in that stage.

Detailed methods from Study 1 replication

Participants

Participants were 639 undergraduate students from a large Canadian university (56.3% women; M_{age} = 19, SD = 1.48; ethnicity: 58.8% East Asian, 18% Caucasian, 14.9% South Asian, 4.5% Middle Eastern, 1.7% African American, 0.62% Hispanic/Latino, 1.41% Native American). Participants received one course credit for participating and earned financial compensation from one randomly selected stage of the task. The pre-registration is at https://aspredicted.org/blind.php?x=2ty9vt. Although we began with a total sample of 641 participants, we excluded two participants who did not understand the instructions, as pre-registered, resulting in a final sample of 639.[1]

Procedure

The procedure for this Study was identical to Study 1 in the main manuscript, with the exception of Stage 4. We removed Stage 4 of the experiment and instead replaced it with a post-experiment survey where we measured mechanisms. The post-experimental survey occurred immediately after round 3 of the math task where participants made the choice under either an opt-in or opt-out frame. Below we detail the mechanisms we measured.

Perceived norms.

We measured participants' perceptions of descriptive and injunctive norms for competition in the experiments (descriptive and injunctive competition norms), as well as perceptions of descriptive and injunctive norms for their gender (descriptive and injunctive gender norms). All scales were adapted from previous research on injunctive and descriptive gender norms [7,8].

[1] Further, for three sessions (Session 10 (n = 16), Session 15 (n = 12), Session 20 (n = 16), Z-Tree crashed in the middle of the session. We recovered what data we could with TreeRing (https://github.com/mjiangsjtu/treering), but not all data points were recovered depending on where and when the crash happened. Thus, some of our analyses will only include 623 participants (43 sessions), and some with only 595 participants and 41 sessions.

Injunctive competition norms were measured by a 1-item measure: "In general, how desirable and encouraged is it for participants to compete in this experiment?" with a scale from 1 (*highly undesirable*) to 7 (*highly desirable*). Descriptive competition norms were measured by having participants report what proportion of all the people that take part in the experiment they thought would choose to compete (as a numerical percentage from 0-100). Injunctive and descriptive gender norms were measured similarly: injunctive gender norms were captured by the same 1-item measure adapted for the participants' own gender: "Of all the women/men that take part in this survey, what proportion of women/men do you think will choose to compete" (1 = highly undesirable, 7 = highly desirable). Finally, we measured descriptive gender norms by asking participants to indicate what percentage of individuals from their own gender (women or men) they thought would choose to compete, ranging from 0-100 percent.

Agency and Communion

We asked participants to rate what kind of person would choose the tournament in Stage 3 on agency and communion. The scales were adapted from past agency and communion scales [9]. To capture perceptions of agency required to participate in the tournament, participants indicated their agreement on an 8-item scale from 1 (*highly disagree*) to 7 (*highly agree*) on what kind of person would choose the tournament: career-oriented, high in leadership ability, assertive, ambitious, competitive, intelligent, has high self-esteem, independent. Due to a bug in the programming, no responses were captured for item 3, "assertive". The scale had good internal consistency, a = .80. Perceptions of communion were similarly captured on an 8-item scale from 1 (highly disagree) to 7 (highly agree): warm, sensitive to the needs of others, cheerful, enthusiastic, cooperative, friendly, polite, humble. The scale had good internal consistency, a = .89.

Backlash

To measure perceptions of anticipated backlash from choosing to compete, we adapted scales developed to measure anticipated backlash for self-promotion and speaking up [10,11] to better suit the context of the experiment. Participants were asked to imagine that an employer knew about their choice (tournament or piece rate), and to indicate the extent of their agreement with a 5-item measure on a scale from 1 (*not at all*) to 7 (*extremely*). Sample items include: "Would you be concerned that you might be disliked?"; "Would you worry that people might think you were too confident (assertive)?"; "Would you be concerned that others might see you as too competitive?" (a = .89).

Ambivalence

We measured ambivalence about the choice by asking participants to think back to when they made their choice of tournament or piece rate. We asked them to rate the extent to which they felt conflicted, indecisive, and had mixed feelings towards the choice to compete in Stage 3 on a scale from 1 (not at all) to 7 (extremely) (a = .91).

Incongruence

Incongruence with gender was measured by adapting previous scales used to measure identity conflict [12]. As with all other scales referencing gender, women saw items that asked about their gender as a woman, whereas men saw items that were adapted to ask about men. Participants indicated their agreement with the following 3-item measure on a scale from 1 (*highly disagree*) to 7 (*highly agree*): "Choosing to compete interferes what I should do as a woman/man"; "I feel that choosing to compete is opposed to what I should to do as a woman/man"; "Choosing to compete highly conflicts with what I should choose as a woman/man" (a = .79).

Gender identity

We measured gender identity using a 4-item measure that asked participants to indicate their agreement with the following statements on a scale from 1 (*highly disagree*) to 7 (*highly agree*): "I am proud to be a woman/man"; "Being a woman/man is central to who I am"; "Being a woman/man is an important part of my self-image"; "Being a woman/man is an important reflection of who I am" [13] (a = .87).

Anxiety of choice

In this experiment, although we also measure overall state anxiety during the entire experiment, we also measured anxiety experienced specifically during the choice of piece rate or tournament. We used the same six-item version of the State Anxiety Inventory [3,4], but adapted it so that participants were prompted to think back to their experience making the choice and answer how they felt during that choice. Because we had obtained poor reliability with the item "I felt content" in the initial experiment, we dropped it in this scale. We asked participants to read the statements and indicate how they felt during the experiment, on a scale from 1 (*not at all*) to 4 (*very much*). Sample items include "While making the choice, I felt calm" (reverse scored), "While I was making the choice, I was tense", and "While I was making the choice, I felt upset". In our sample, the five-item scale had an internal reliability of a = .74.

Overall anxiety

As with our main experiment, we measured the well-being of the participants in terms of their perceived anxiety during the experiment. We relied on a six-item version of the State Anxiety Inventory (SAI) [3,4]. We asked participants to read the statements and indicate how they felt during the experiment, on a scale from 1 (*not at all*) to 4 (*very much*). In our sample, the six-item scale had an internal reliability of a = .75. Upon closer examination, it appeared that the

item “I felt content” had poor correlation compared to the rest of the other items. Thus, we removed it from the scale [6]. The five-item scale had good reliability (a = .78).

Results from Study 1 replication

Similar to Study 1, we did not find a gender difference in average performance on the task (Fig. S6). Men's average performance (M = 7.98) in Stage 1 was no different than women's (M = 8.03) (*p* from two-tailed t-test = .84), and there was no significant difference in the overall distribution (*p* from Kolmogorov-Smirnov test = .77). In Stage 2, both men and women attained significantly higher average performance. However, there were again no significant differences in score between men (M = 9.87) and women (M = 9.87; *p* from two-tailed t-test = 0.99). Graphical evidence is in Fig. S7.

When we explore the key finding of our study (i.e., the choice of compensation scheme in Stage 3), we again replicate our initial experiment. The findings from the condition where participants had to opt-in to compete are similar to our initial experiment and the evidence from previous experiments using the same paradigm, with far fewer women than men choosing tournament-based payment: 52.6% vs. 72.1% (*p* from two-tailed t-test < 001; Fig. 1 Panel A2). In contrast, the proportions of women and men choosing tournament compensation in the opt-out condition were statistically indistinguishable, and similar to the percentage of men choosing tournament in the basic opt-in scheme: 73.5% of women vs. 78% of men (*p* from two-tailed t-test = 0.38). In Table 1 model 2 we report Probit regression estimates where we also control for performance in Stages 1 and 2, as well as for overconfidence in Stage 2 (the difference between actual rank and guessed rank); the estimates from these models confirm the descriptive results of these findings.

We once again examined whether introducing an opt-out frame might lead to unintended negative consequences to performance and broader well-being, finding again no evidence of such

consequences (Table 2 and S11). Findings on payoff-maximizing choices and average gains were similar to those from the original Study 1, too (Fig. 1, Table S7 and S8).

Table S13 shows the average rating of the various measures described above by choice architecture, gender, and choice of compensation scheme in Stage 3 on our different mechanism measures. We focused on the construct that referred to beliefs and feelings about competition (injunctive and descriptive norms, ambivalence, agency, communion, backlash, incongruence).

Table S7. Number and share of payoff-maximizing choices by gender, choice and condition, in the replication of Study 1

	Number of payoff-maximizing choices		Percent of payoff maximizing choices	
	By condition-choice-gender	By condition-gender	By condition-choice-gender	By condition-gender
Opt in, piece rate: Men	27	79	65.9%	53.7%
Opt in, tournament: Men	52		49.1%	
Opt out, piece rate: Men	25	60	86.2%	45.5%
Opt out, tournament: Men	35		34.0%	
Opt in, piece rate: Women	50	92	60.2%	52.6%
Opt in, tournament: Women	42		45.7%	
Opt out, piece rate: Women	34	101	69.4%	54.6%
Opt out, tournament: Women	67		49.3%	

Notes: The table reports the number and percentage, of subjects who made the payoff maximizing choice for them in Stage 3 by experimental condition, gender, and choice of compensation scheme. We established the payoff maximizing choice as follows. We used the estimated coefficients from a Probit regression of whether a participant won their tournament in Stage 2 on the number of correct responses in that stage, to predict the likelihood of winning a tournament in Stage 3 given their performance in Stage 3 (recall that individuals who chose the tournament in Stage 3 had their performance compared against three participants from Stage 2). We then calculated the *expected* payoff from choosing a tournament or a piece rate compensation for each participant. For the piece rate, the expected payoff was \$0.50 X the number of correct responses in Stage 3. The expected payoff from a tournament was \$2 *times* the number of correct responses in Stage 3 *times* the predicted probability of winning the tournament for that subject in Stage 3. We classify a participant as having made their payoff-maximizing choice if they selected the compensation scheme that gave them the higher expected payoff. In cases where the expected payoffs from the two compensation scheme were close to each other (less than \$1 in absolute difference; this happened if a participant solved 12 questions correctly), we randomized the assignment to having made the payoff maximizing decision or not (we interpreted small differences as making a participant indifferent between the two schemes).

Table S8. Average monetary gains (losses) from choosing a compensation scheme in Stage 3, compared to "counterfactual" choice in the replication of Study 1

	Sum of net gains	Avg. net gains	N
Gender			
Men	$738.5	$2.80	260
Women	$943.5	$2.80	335
Choice architecture			
Opt in	$812.0	$2.70	298
Opt out	$870.0	$2.90	297
Conditions in stage 3, by gender			
Opt in: Men	$489.0	$3.60	137
Opt in: Women	$323.0	$2.00	161
Opt out: Men	$249.5	$2.00	123
Opt out: Women	$620.5	$3.60	174
Conditions and compensation choices in stage 3, by gender			
Opt in, piece rate: Men	-$53.0	-$1.30	40
Opt out, piece rate: Men	$39.5	$1.60	25
Opt in, piece rate: Women	-$87.0	-$1.20	73
Opt out, piece rate: Women	$15.0	$0.30	47
Opt in, tournament: Men	$542.0	$5.60	97
Opt out, tournament: Men	$210.0	$2.10	98
Opt in, tournament: Women	$410.0	$4.70	88
Opt out, tournament: Women	$605.5	$4.80	127

Notes: this table reports both the total gains (over all participants) and the average gains per participants from choosing a compensation scheme over the alternative one. For the participants who selected a tournament-based compensation scheme, the "counterfactual" payoff is the number of correct responses that they gave in Stage 3 multiplied by $0.5. For the participants who selected piece rate, the counterfactual payoff is zero if they would have not won the tournament in the group to which they would be assigned, and equal to the number of their correct answers multiplied by $2 had they been the winners of the groups to which they were assigned.

Table S9. Additional specification for compensation choice regressions in the replication of study 1: interactions (1).

	(1)	(2)
Outcome variable:	Choice of tournament in Stage 3	
Estimation:	Probit	
Sample:	Opt-in condition	Opt-out condition
Woman	-0.053	-0.054
	(0.147)	(0.129)
# correct answ. in stage 2	0.023***	0.024**
	(0.009)	(0.010)
Woman X # correct answ. in stage 2	-0.014	-0.001
	(0.013)	(0.013)
Observations	322	317
Pseudo R2	0.045	0.035

Notes: The table reports parameter estimates from ordered probit regressions where the outcome variable is the guessed tournament rank in stage 2 (column 1) and the difference between the actual rank and the guessed rank (column 2), and the regressors are the gender of the respondent, the number of correct responses in stage 2, and the difference between the number of correct responses in stage 2 and stage 1. Estimated standard errors, clustered at the session level (n=44) are in parentheses. * p<0.1, ** p<05, *** p<.01 (two-sided tests).

Table S10. Additional specification for compensation choice regressions in the replication of study 1: interactions (2)

	(1)
Outcome variable:	Choice of tournament in Stage 3
Estimation:	Probit
Opt in: Woman	-0.336
	(0.209)
Opt out: Man	-0.165
	(0.276)
Opt out: Woman	-0.213
	(0.251)
# correct answ. in stage 2	0.013
	(0.012)
# correct in stage 2 - # correct in stage 1	-0.023
	(0.015)
Guessed rank in stage 2 tournament	-0.088**
	(0.037)
# correct answ. in stage 2	-0.005
x Opt in: Woman	(0.017)
# correct answ. in stage 2	0.021
x Opt out: Man	(0.021)
# correct answ. in stage 2	0.009
x Opt out: Woman	(0.020)
# correct in stage 2 - # correct in stage 1	0.036
x Opt in: Woman	(0.022)
# correct in stage 2 - # correct in stage 1	-0.003
x Opt out: Man	(0.028)
# correct in stage 2 - # correct in stage 1	0.031
x Opt out: Woman	(0.022)
Guessed rank in stage 2 tournament	0.110**
x Opt in: Woman	(0.050)
Guessed rank in stage 2 tournament	0.058
x Opt out: Man	(0.058)
Guessed rank in stage 2 tournament	0.075
x Opt out: Woman	(0.052)
Observations	623
Pseudo R2	0.0653

Notes: The outcome variable is a binary indicator equal to 1 if a participant selected tournament-based compensation in stage 3, and 0 if they selected piece rate compensation. Regressors include the experimental conditions (opt-in vs. opt-out frame) interacted with the gender of the participant (the omitted category is men in the opt-in condition); the number of correct responses in stage 2, and the difference between the correct answers in stage 2 and those in stage 1; the positions that each participant guessed to have achieved in the tournament in stage 2 (out of four position, rank 1 being the winner); and interaction of the gender of the participant with correct responses in Stage 2, and the difference between the correct answers in stage 1 and those in Stage 1; the positions that each participant guessed to have achieved in the tournament in stage 2. The estimates indicate marginal effects from probit regressions. Estimated standard errors, clustered at the session level (n=43) are in parentheses. * p<0.1, ** p<.05, *** p<.01.

Table S11. Correct responses in stage 3 and anxiety levels in the replication of study 1: Regression estimates

Outcome variable:	# of correct answers in stage 3						Average anxiety					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Woman	0.264						0.122***					
	(0.289)						(0.035)					
Conditions in stage 3, by gender												
Opt in: Woman		0.191						0.128***				
		(0.335)						(0.045)				
Opt out: Man		-0.232						-0.017				
		(0.315)						(0.057)				
Opt out: Woman		0.139						0.099				
		(0.294)						(0.060)				
Conditions and compensation choices in stage 3, by gender												
Opt in, tournament: Man			2.363***	1.258***	1.316***	1.246**			0.121	0.129	0.147*	0.141*
			(0.810)	(0.462)	(0.467)	(0.480)			(0.077)	(0.078)	(0.076)	(0.078)
Opt out, piece rate: Man			-0.543	0.404	0.421	0.684			0.120	0.116	0.121	0.090
			(0.693)	(0.658)	(0.663)	(0.612)			(0.138)	(0.136)	(0.135)	(0.143)
Opt out, tournament: Man			0.758	0.565	0.618	0.548			0.065	0.064	0.080	0.070
			(0.636)	(0.381)	(0.384)	(0.390)			(0.089)	(0.091)	(0.088)	(0.091)
Opt in, piece rate: Woman			1.038	0.806	0.830	0.709			0.208*	0.202*	0.209*	0.216*
			(0.672)	(0.506)	(0.514)	(0.534)			(0.105)	(0.107)	(0.106)	(0.114)
Opt in, tournament: Woman			1.712***	1.050**	1.062**	1.070**			0.230***	0.246***	0.249***	0.241***
			(0.622)	(0.426)	(0.432)	(0.443)			(0.082)	(0.083)	(0.081)	(0.084)
Opt out, piece rate: Woman			0.616	0.889	0.908	0.846			0.199*	0.194*	0.200*	0.190*
			(0.650)	(0.540)	(0.548)	(0.565)			(0.102)	(0.104)	(0.103)	(0.104)
Opt out, tournament: Woman			1.654**	0.839*	0.862*	0.819*			0.187**	0.179*	0.186*	0.177*
			(0.628)	(0.424)	(0.431)	(0.440)			(0.091)	(0.092)	(0.093)	(0.095)
# correct answ. in stage 2		0.824***		0.936***	0.959***	0.918***		-0.006		-0.004	0.003	-0.006
		(0.066)		(0.027)	(0.028)	(0.027)		(0.005)		(0.006)	(0.006)	(0.007)
# correct in stage 2 - # correct in stage 1				-0.199***	-0.193***	-0.174***				-0.005	-0.003	-0.005
				(0.044)	(0.044)	(0.040)				(0.010)	(0.010)	(0.010)
Guessed rank in stage 2 tournament					0.172**						0.054*	
					(0.074)						(0.028)	
(Over)confidence						-0.085						-0.025
						(0.065)						(0.023)
Constant	9.975***	1.939***	8.854***	0.654	0.010	0.811*	1.990***	2.060***	1.907***	1.953***	1.749***	1.980***
	(0.264)	(0.384)	(0.512)	(0.436)	(0.478)	(0.450)	(0.026)	(0.068)	(0.075)	(0.091)	(0.107)	(0.103)
Observations	639	639	639	623	623	595	639	639	639	623	623	595
Pseudo R2	0.001	0.610	0.037	0.745	0.747	0.746	0.014	0.016	0.017	0.020	0.028	0.023

Notes: The table reports parameter estimates for linear regression models where the outcomes variables are the number of correct responses in stage 3 by each participants (columns 1 through 6) and the anxiety index (columns 7 through 12). Regressors include the experimental conditions (opt-in vs. opt-out frame) interacted with the gender of the participant (the omitted category is men in the opt-in condition); the experimental conditions (opt-in vs. opt-out frame) interacted with the gender of the participant and the choice of compensation scheme (the omitted category is men in the opt-in condition who chose piece rate); the number of correct responses in Stage 2, and the difference between the correct answers in Stage 1 and those in Stage 1; the positions that each participant guessed to have achieved in the tournament in Stage 2 (out of four position, rank 1 being the winner); and the difference between the actual position and the guessed position, as a measure of (over) confidence. Estimated standard errors, clustered at the session level (n=44 in columns 1-3 and 7-9; 43 in coluMns 4-5 and 10-11; 41 in columns 6 and 12) are in parentheses. * p<0.1, ** p<.05, *** p<.01 (two-sided tests).

Table S12. Guessed rank and (over)confidence in stage 2 in the replication of study 2

Outcome variable:	Guessed rank for stage 2	(Over)Confidence
	(1)	(2)
Woman	0.208***	-0.176**
	(0.060)	(0.088)
# correct answ. in stage 2	-0.177***	-0.072***
	(0.015)	(0.016)
# correct in stage 2 - # correct in stage 1	-0.033*	0.012
	(0.020)	(0.016)
Observations	623	595
Pseudo R2	0.124	0.022

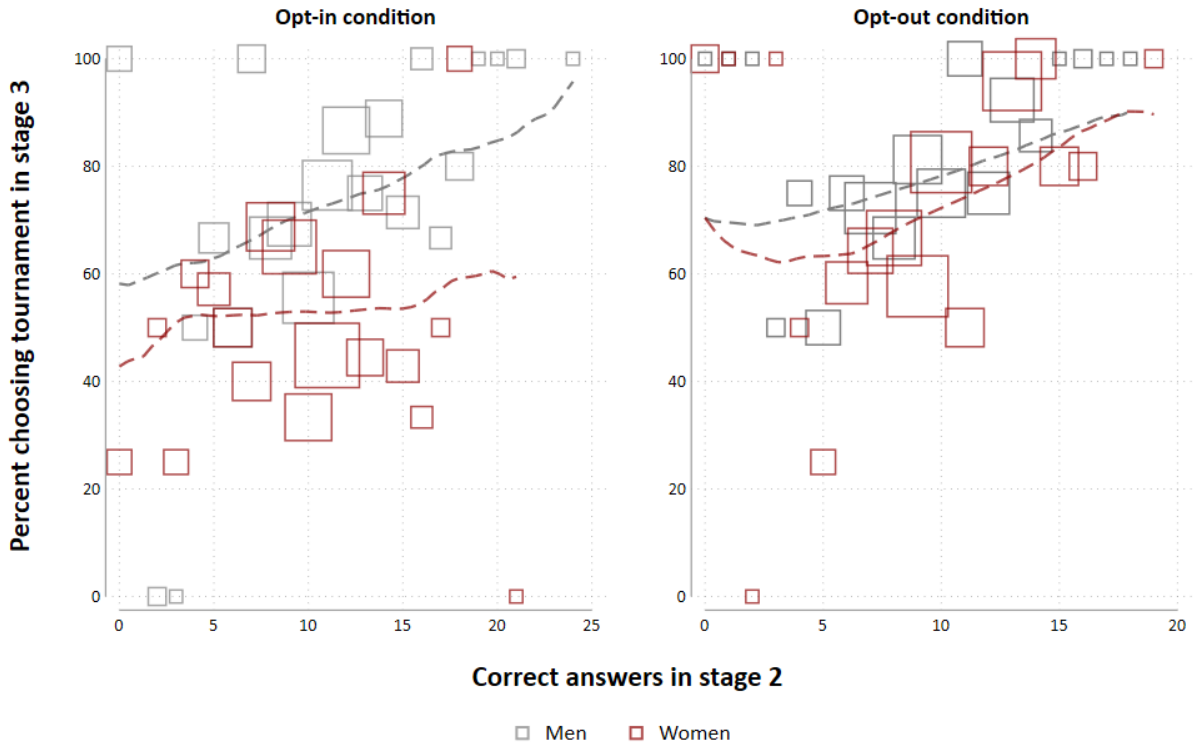
Notes: The table reports parameter estimates from ordered probit regressions where the outcome variable is the guessed tournament rank in stage 2 (column 1) and the difference between the actual rank and the guessed rank (column 2), and the regressors are the gender of the respondent, the number of correct responses in stage 2, and the difference between the number of correct responses in stage 2 and stage 1. Estimated standard errors, clustered at the session level (n=43 in column 1; 41 in column 2) are in parentheses. * p<0.1, ** p<05, *** p<01 (two-sided tests).

Table S13. Average response scores by gender, condition, and compensation scheme choice in the replication of Study 1, to survey questions about opinions on features of competition choices.

	Desirable for women (men)	Proportion of women (men)	Desirable in general	Proportion in general	Agency	Communality	Backlash	Ambivalence	Incongruence	Identity
Opt in, piece rate: Men	5	65.54	4.9	58.34	4.6	3.7	2.51	3.48	2.6	5.32
Opt in, tournament: Men	5.39	74.33	5.60	68.15	5.05	3.91	2.41	2.76	2.63	5.17
Opt out, piece rate: Men	4.90	68.59	5.03	63.90	4.74	3.81	2.17	3.52	2.44	5.3
Opt out, tournament: Men	5.29	78.97	5.29	73.21	5.03	4.14	2.45	2.54	2.42	5.31
Opt in, piece rate: Women	4.34	55.54	4.81	63.75	4.73	4.05	2.92	3.66	2.18	5.75
Opt in, tournament: Women	4.49	67.01	5.49	74.15	5.14	4.39	2.69	2.76	1.68	5.74
Opt out, piece rate: Women	4.18	51.08	4.73	62.37	4.68	3.95	2.58	3.63	1.78	5.99
Opt out, tournament: Women	4.70	72.71	5.14	78.16	4.92	4.20	2.91	2.96	2.04	5.64

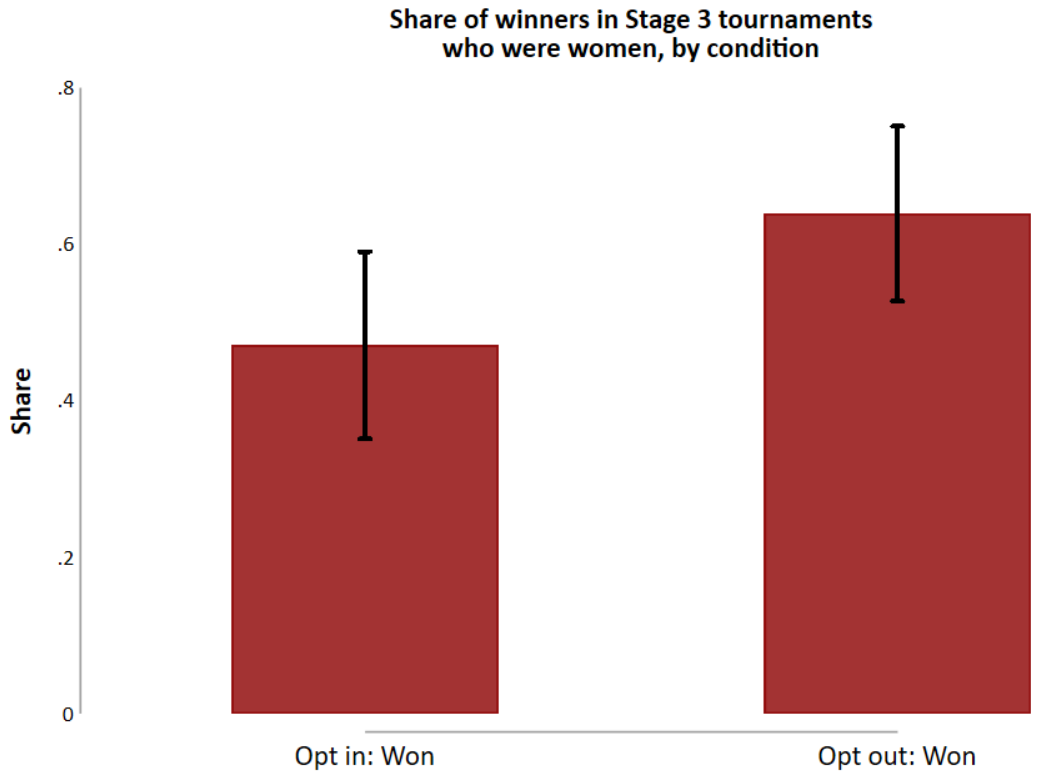
Notes: The survey was part of the replication of Study 1. "Proportion of women (men)" and "Proportion in general" refer to questions about the predicted share of participants of one's own gender, and participants in general, who would choose a tournament-based compensation in Stage 3. Participants could report any integer value between 0 and 100. "Desirable for women (men) and "Desirable in general" refer to questions about whether respondents considered competitive schemes desirable and encouraged for people of their own gender and people in general. The responses were on a scale from 1 (strongly disagree) to 7 (strongly agree). In the columns "Backlash", Ambivalence" and "Incongruence", the values are the average of averages across responses to multiple questions on three issues: whether respondents felt that choosing to compete might lead to backlash, whether they felt conflicted/ambivalent about the choice to compete, and whether they felt that the choice to compete may be incongruent with their own gender. Answers were all on a 1 to 7 scale. n = 639 for all variables.

Fig S4: Choice of compensation in stage 3 by correct answers in stage 2



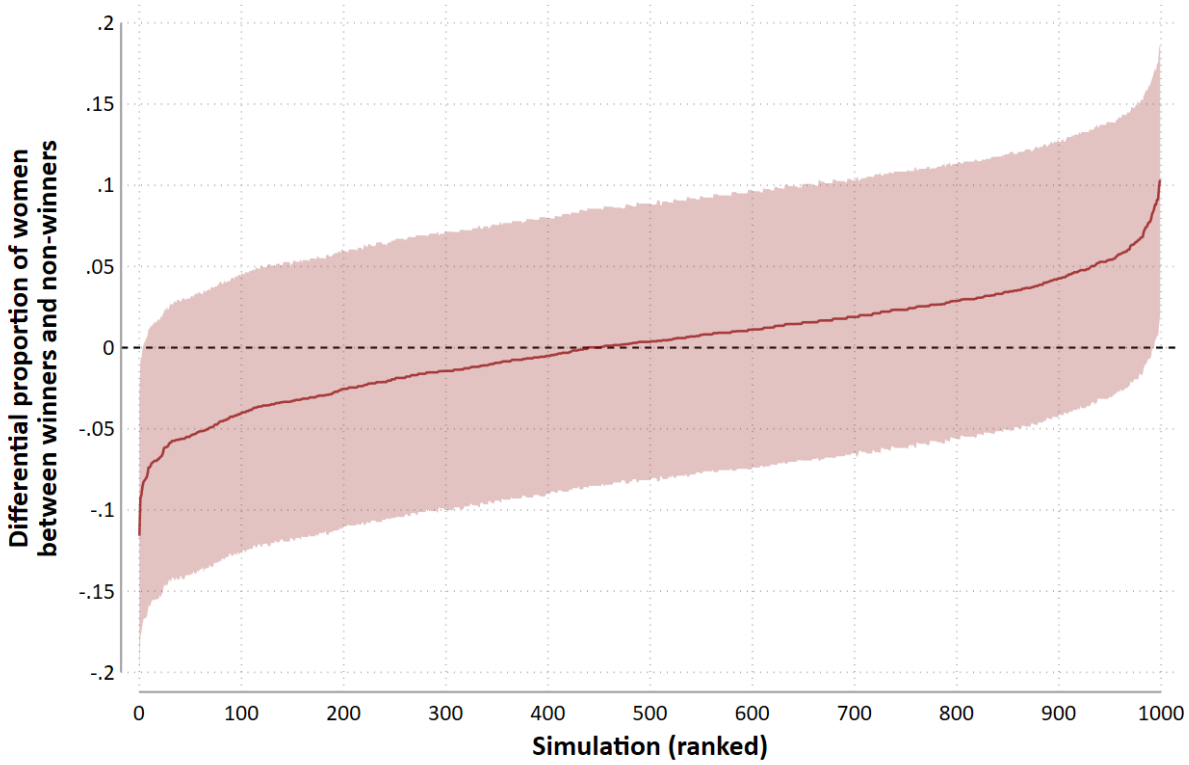
Notes: The x-axes in the graphs report the number of correct responses in stage 2. Each square represents the share of respondents (separated by gender) who chose tournament compensation in stage 3, for each number of correct responses in stage 2. The size of the circles is proportional to the number of participants, by gender and experimental condition, who attempted a given number of tasks, relative to the total number of participants of a given gender in a given condition. The dashed lines are smooth polynomial approximations of the relationship between number of correct answers in stage 2 and the likelihood of choosing tournament compensation in stage 3 (degree zero, bandwidth 3), separate by gender.

Fig. S5. Proportion of tournament winners in stage 3 who were women



Notes: The graph displays the percentage of tournament winners in Stage 3 (among all those who chose a tournament) who were women, separately for the opt-in and the opt-out condition.

Fig. S6. Simulated differences in the proportion of women between winners and non-winners of a tournament in Stage 2



Notes. The red solid line reports the estimated difference between the average proportion of tournament winners who were women, and the average proportion of tournament "non-winner" who were women, from each of 1,000 simulations where we randomly assigned each participant to a fictitious group and defined the winner as the participant(s) with the highest number of correct answers in stage 2. A negative value, for example, indicates that for a given simulation, the proportion of non-winners who were women (e.g. 53%) was higher than the proportion of winner who were women (e.g. 51%). The shaded area represents the confidence intervals around each of the 1,000 estimated average differences. Note that these randomly created groups are not the ones within which a participant actually competed in the experiment. Moreover, unlike the groups to which each participant was assigned in the experiment, in these simulations the groups are without replacement, i.e. participants in a given group cannot also be in another group. Each of the 1,000 simulations re-shuffled the groups. The purpose of this exercise is to further test for the presence of underlying (dis)advantages of women (or men) in being the winner of the tournaments in stage 2, based on their actual absolute performance in that stage.

Detailed methods from field experiment

Empirical setting

We implement a pre-registered field experiment on an online labor market – Upwork – to test the implications of our proposed treatment in a field setting. The pre-registration is available at https://aspredicted.org/blind.php?x=vf4r4n. Upwork is an online labor market for skilled freelancers who wish to obtain work, and clients who wish to outsource a job or task. Employers on Upwork can hire skilled freelancers to complete various jobs, ranging from data entry, creating mobile apps, to acting (include other examples).

Freelancers on Upwork are more educated than most, with 77% holding a college degree[14]. On Upwork, clients can post one-time (fixed amount) jobs to long-term (hourly) jobs, with typical jobs taking days or weeks to complete. Compensation ranges from tens to hundreds of dollars, and the average hourly wage on Upwork is $28/hour, which is comparable to an average annual U.S. household income[15]. In 2017, Upwork reported over one billion dollars (USD) in annual freelancer billings[16]. Thus, Upwork represents a skilled labour market with high wages and high stakes jobs, populated by freelancers who can earn a living on the platform.

Upwork is one of the most commonly used freelance websites, with over 14 million active users[17]. Although online labour markets are not "field" or organizational settings in the traditional sense, with the rise of remote work and the gig economy, they are becoming more and more relevant field settings. Indeed, there is an increasing number of freelancer and gig workers in U.S. and Canada[18,19]. Moreover, this number is projected to increase due to the pandemic accelerating online work [20–22]. The relevance of this type of field setting is evident in multiple published papers citing Upwork as their main context [23–26].

Importantly, for our research question, decisions captured in the lab can be tightly

controlled to isolate mechanisms, but many decisions made in real life (i.e., in the field) are more complex. Thus, moving from evidence in a tightly controlled laboratory setting to testing the intervention in a field setting allows us to examine how behavioral change may or may not replicate in the field, and whether an opt-out intervention can be practically applied by organizations seeking to reduce gender disparities in competition and promotion.

Given Upwork's unique setting, we design a field experiment that maintains certain experimental aspects of the lab experiments [1], mimics aspects of real promotions in organizations, yet conforms to norms on Upwork. Below, we give an overview of the experimental design before delving into the specific details.

Overview of experimental design

We act as a real client on Upwork to hire freelancers to complete a real job. Freelancers were invited to complete a data scraping task, which comprised of two phases: an assessment phase (test project phase) and a task phase (see Burbano 2020 for another example of a multiple-phase job design on Upwork). For the assessment phase, freelancers were to complete a paid "test project", which is a common feature of many Upwork jobs[2]. The paid test project typically acts as an assessment, or skill test, before employers choose whom to hire. For the task phase, freelancers would be sorted into two possible tasks (standard or advanced) depending on their performance and choice during the assessment phase. Taking advantage of this set-up, we mimic a more realistic long-term employer-employee relationship that lasts over multiple milestones/tasks. This multiple-phase set up also allows us to retain the experimental design features of our laboratory experiments [1]. Fig. S7 shows the process of the entire experiment.

We administered the choice treatment during the test project and elicit freelancers' choice

[2] See https://www.upwork.com/hiring/startup/how-to-freelancer-test-project/

(our main outcome variable of interest) during the test project to determine sorting of freelancers into two jobs in the task phase.

Logistically, we made use of the milestone system on Upwork to implement this multiple-stage set-up on Upwork. We hired each freelancer and created individual contracts with each freelancer. The contract upon hiring started with only one milestone (the paid test project), and freelancers were informed that if they were to continue to the task phase we would add a second milestone, which would depend on their sorting. Following completion of the second milestone, we ended the contract with each freelancer and left a review. The content of the review and the overall rating were determined a priori and we detail this in the section "Ending Contracts and voluntary exit survey".

The entire experiment took place over a span of 5 weeks, from November, 2020 to December, 2020. As pre-registered, we posted 5 job postings given the limit of 99 freelancers per job posting. We posted each job posting at the same time (8:30AM EST) on the Monday of each week.

The central task that we chose was a data entry task. Our reason for doing so is threefold. First, a data entry task is less male-typed than math, which allows us to test whether the gender gap in competition replicates for a relatively less culturally male-typed task. Second, a data entry task allows us to more objectively discern performance (by counting the number of correct data entries) and allows us to retain the design of Niederle & Vesterlund[1] to assess performance within a time limit and calculate a commission. Third, data entry on Upwork is a highly populated field with many freelancers, which provides us with a large enough population from which to sample while avoiding repeat freelancers.

Procedure

Fig. S7 displays the experimental design flow of the field experiment.

Test Project. The job posting of the test project was posted on each Monday at 8:30AM EST. We advertised the job as a short data entry task that was a one-time task, but made clear in the job posting that there was a paid test project and a following "official" task to follow. Because we posted multiple job postings, we made sure to include in the job posting that we were only hiring new freelancers who had not applied to a previous version of the job. To additionally verify that only new and unique freelancers were hired each time, research assistants coded whether freelancers were indeed unique participants before adding them to the hiring list.

The test project advertised a 10-minute task that would pay freelancers $5, with an added bonus commission of $0.25 per correct data entry scraped. Although this rate seems low, this falls into the normal range of data entry tasks on Upwork. On Upwork, data entry tasks can range from short, entry-level one-time jobs that pay $5 to more long-term, advanced level jobs that pay $30/hour. Thus, our short, entry-level, fixed-amount job was typical of a data entry posting on Upwork. This is further evidenced by our ability to recruit over 400 freelancers.

As pre-registered, all unique freelancers who applied were hired, in order of their application time. Once freelancers were hired, we sent them a message that described the details of the task, where we reminded them of the multiple-phase aspect of the job. We then assigned them an anonymous worker ID, which they were instructed to use throughout all of the jobs and our correspondence. We then gave them a link to the task, which was hosted on Qualtrics.

The test project began with instructions detailing an overview of the task: it described the set-up of the multiple phases in the job, and informed freelancers that they would make a choice during the test project to determine their sorting in the second task phase. We also informed them of their compensation scheme: $5 show-up fee for the test project, and a commission of $0.25 for

the two parts that they were to complete in the test project.

The test project task consisted of two parts.

Part 1. In part 1, participants completed the standard level task, which asked them to scrape basic information (company name, revenues, profit, number of employees) about Fortune 500 companies. Participants had 5 minutes to scrape data about as many companies as they could, and were asked to input their answers into a Qualtrics form for each company. For the standard level task, participants were paid a standard rate, which is $0.25 per correct data entry.

Choice. Following Part 1, but before starting Part 2, freelancers were notified that they had to make a choice about the type of task they want to complete in the task phase (second phase). There were two possible options: the standard task or advanced task.

Freelancers who selected the *standard* task again would be invited back to the task phase for a standard task, and paid the same rate as the test project ($5 participation fee, $0.25 commission). We notified them that the difficulty would be similar.

Freelancers who selected the *advanced* task entered a competition, where their scores in Part 2 of the test project would be used to compete against other freelancers for the advanced task; only the top 25% of performers who applied would be selected to complete the advanced task. Those who applied to the promotion to the advanced task and were selected (scores were among the top 25% of performers) would be invited back to the advanced job in the task phase, and they would be paid $7.50 participation fee and an increased commission of $1 per correct data entry[3]. Those who applied to the promotion to the advanced task and were *not* selected were

[3] We determined this pay-off scheme by keeping pay-offs relatively equivalent for choosing the standard task versus competing for the advanced task, while keeping in mind some external validity concerns. For the standard task, freelancers had a 100% chance of earning $0.25 per data entry (plus a participation fee of $5). For the advanced task, freelancers had a 25% chance of earning $10 per data entry (plus a participation fee of $7.50), and 75% chance of earning nothing. We added an increase of $2.50 to the participation fee of the advanced task to mimic a promotion bonus.

simply not invited back to the task phase at all. Thus, this constituted an up-or-out promotion scheme.

In describing the choice between these two tasks, we administered our treatment and varied the default choice. As in our previous set-ups, participants were randomly assigned to an opt-in or opt-out condition. In the opt-in condition, freelancers by default enrolled in the standard task for the task phase, but had the option to apply instead to compete for the advanced task in the task phase by checking a box. In the opt-out condition, freelancers were by default enrolled in the competition for the advanced task, but had the option to opt-out of the competition and proceed instead to the standard task by checking a box[4].

Part 2. Once freelancers made the choice during the choice phase, they proceeded to Part 2 of the test project. Similar to Part 1, participants were paid a standard rate ($0.25 commission per correct data entry) and they had 5 minutes to enter data about companies.

Once freelancers completed the test project, they were paid the participation fee and a bonus commission fee across part 1 and part 2. They were then notified of the evaluation phase, which took place over 2 days.

Evaluation Phase. The evaluation phase took place after all freelancers completed the paid test project, typically around day 3 of the week. During the evaluation phase, we determined the sorting of freelancers into tasks for the task phase. All of those who chose to stay with the standard task were assigned to be invited back to the standard task. For those who applied to the advanced task, we determined the cut-off score for the top 25 percentile of performance of those who applied. Those who fell above or on that cut-off score were selected for the promotion to the advanced task. For cases where there were more than 25% of freelancers who scored on or above

[4] Given the complexity of this choice, we worked with a copy-editor to ensure that our instructions were clear.

the cut-off point, we took the ties at the cut-off score and randomly selected the number of winners that would bring us to exactly 25% of applicants selected.

All of the selection was done by code in R. Following this evaluation phase, we notified all freelancers of their task phase.

Task Phase. During the task phase, freelancers were notified of their sorting. We invited those in the standard task to complete the standard task, again hosted on Qualtrics. As an extended version of the paid test project, participants had 10 minutes to scrape data about as many companies as they could and were asked to input their answers into a Qualtrics form for each company. Rather than simply scraping basic company data about Fortune 500 companies, this time the standard task asked freelancers to scrape data about Fortune 500 companies' websites but was similar in length and difficulty to the test project. Again, participants were paid a standard rate of $5 participation fee and $0.25 per valid data entry.

We invited those who were selected for the promotion to the advanced task to a more difficulty task. Again, they had 10 minutes to scrape data about as many companies as they could and were asked to input their answers into a Qualtrics form for each company. For the advanced task, freelancers were asked to scrape data about the company's mission statements and diversity statements and was more difficult and involved more web research and navigation on the company's site. Freelancers were paid an advanced rate of $7.50 participation fee, and an increased commission of $10 per valid data entry.

Finally, for those who applied but were not selected for the advanced task, we notified them of this result and ended their contract. Please see section on "ending contracts" for details of this process.

Ending Contracts. On Upwork, ending contracts with freelancers requires a review of the

freelancers, which involves ratings out of 5 on a multiple item scale. We determined the performance ratings a priori based on freelancers' performance on the test project[5]. Those who scored in the bottom quartile of performance on the test project (the cut-off was almost always 3 or less correct data entries) received a rating of 4.80 out of 5 on the task (they were given 4/5 on the "skills" component of the rating matrix). We left a review that stated that the freelancer did a "good job" on the task, and that they completed their task in a timely manner. Those who scored above the cut-off received a rating of 5/5 on the task, and we left a review that stated that the freelancer did an "excellent job" on the task and added that we would definitely hire them back for another job.

Participants

Our final sample was 482 freelancers over 5 job postings. We excluded 5 participants whose timer malfunctioned in the test project, and our final sample was 477 freelancers. The final sample was comprised of 173 women (36%) and 304 men (64%)[6]. Table S14 displays a breakdown of descriptive statistics of the overall sample.

[5] We chose to base the rating on the test project given that some freelancers would not advance at all to the task phase and thus have no performance upon which to base the rating.
[6] These numbers are based on self-identified gender, which we consider to be most accurate. We also have coded gender as an additional variable, which was coded by research assistants based on profile photos of the freelancers.

Detailed Results from Field Experiment

First, we examine performance on the data entry task, operationalized by the number of correct data entry that freelancers completed within the time limit. The evidence is consistent with no overall gender differences in performance on the task: For Part 1 of the Test project, men's average performance (M = 2.72, SD = 2.52) was no different from women's average performance (M = 2.84, SD =2.84) (p = 0.63). There was no significant difference in the overall distribution (p from Kolmogorov-smirnov test = 0.83). In Part 2 of the Test project, again there were no differences between men's average performance (M = 2.94, SD = 2.54) and women's average performance (M = 3.21, SD=2.96) (p = 0.30). Again, there was no significant difference in the overall distribution (p from Kolmogorov-smirnov test = 0.67). This evidence is consistent with men and women performing similarly on the task.

Upon this baseline of equal performance, we next examine the choice to apply for the advanced task versus staying with the standard task in the task phase. Figure 2 Panel A displays the percentage of men and women who chose to apply to compete for the advanced task in the task phase, by experimental condition. We find that on the baseline of equal performance on the task, in the opt-in condition women (57.3%) are significantly less likely than men (72.5%) to apply for the advanced task (p = .015). In the opt-out promotion, the gender gap shrinks and is no longer significant between men (71.6%) and women (66.7%) (p = .43). Opt-out framing attenuated the gender gap by around 10% -- but although this difference in participation rates is substantial, it is not statistically significant (p = .25).

We next explored the role of performance on part 1 of the test project on likelihood of applying to the advanced job. Although the results of part 1 of the test project were not immediately communicated to the freelancers, the data entry task simply involved finding a company and

entering that data into the survey – task that was fairly straightforward and involved little to no computation. Indeed, the total number of attempted data entry was highly correlated with the correct number of data entry ($r = .90$, $p < .001$). Thus, it is reasonable to assume a strong relation between effort and performance. As a result, although freelancers were not given feedback about their performance directly, they likely had a good grasp of their performance as judged by the amount of data entries they attempted.

The binned scatterplots in Figure 2 Panel B report the share of workers who applied for the advanced task, by their score in the first part. Note that the only case where the participation decision is significantly "responsive" to the part 1 performance is for women in the opt-in condition. When participating in a tournament requires an active decision (i.e., to opt-in), women's participation is much more sensitive to ability. Women who have a "bad" signal are more likely to not compete. Note that men do not show this tendency; this means that there are women likely to be more able than men, who do not compete.

Table 15 confirms the descriptive evidence of Figure 2 Panels B1 and B2. This table shows the regression estimates from the analysis examining the relationship between performance on part 1 of the test project and participation in the competition for the advanced task, by gender and experimental condition. We display the regression estimates broken down by each group, but also the full sample. The outcome variable is a 0-1 indicator for the choice to apply to the advanced task. Columns 5 and 6 show that for men in the opt-in and opt-out condition, scores in part 1 did not affect their likelihood of applying to the advanced task. However, Columns 6 and 7 show that scores in part 1 was a strong positive predictor of women's likelihood of applying in the opt-in condition; this estimate becomes negative for women in the opt-out condition. We interpret this as opt-out framing dulling women's "sensitivity" to bad signals by creating a strong norm to apply.

These analyses around the "threshold" for when women and men apply under different choice conditions reveal an interesting mechanism: in the opt-in condition, only "excellent" women apply for the advanced job, suggesting that women see a higher "bar" for the competition compared to men. However, in the opt-out condition where the default is participation, women are less "sensitive" to bad signals as the decision to apply for the competition requires less deliberation. The analyses on performance on part 2 of the test project (Table 2, columns 3 and 4) suggest that once women choose to apply, they perform well – and so when we nudge women in the opt-out condition to compete more often even though they perform slightly worse in part 1 of the test project, they step up and improve, performing just as well as women who entered with a higher bar under the opt-in scheme when given the opportunity.

Fig. S7. Experimental design and procedure of field experiment.

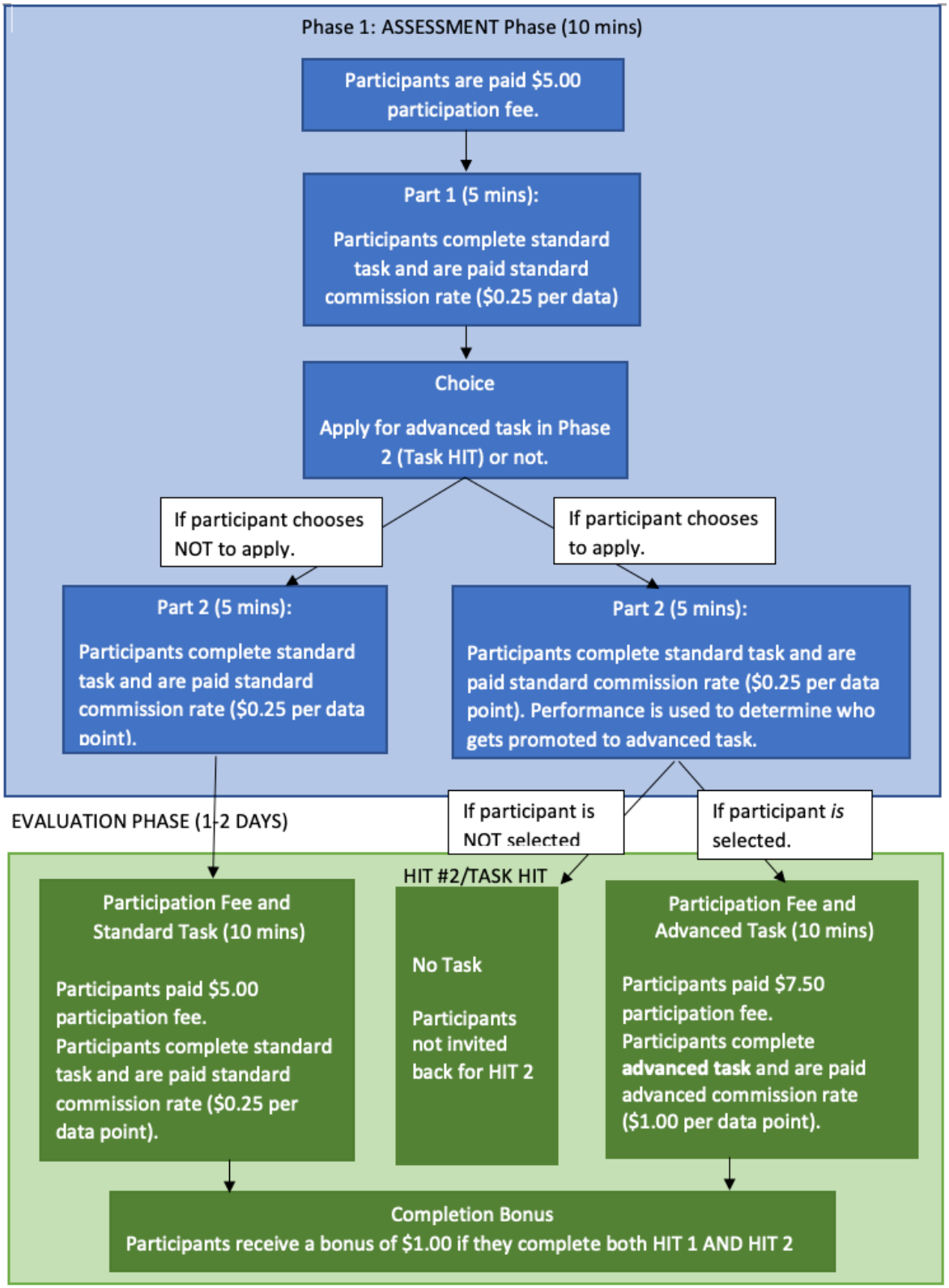


Table S14. Table of Descriptive Data of Upwork Freelancers

Variable	n	mean	sd	median	trimmed	mad	min	max	range
Bachelor's degree (1 = yes)	475	0.705	0.456	1	0.756	0	0	1	1
Master's degree (1 = yes)	475	0.225	0.418	0	0.157	0	0	1	1
New to Upwork (1 = yes	474	0.549	0.498	1	0.561	0	0	1	1
Hourly Rate on Upwork	474	7.932	7.151	5	6.648	1.483	1	100	99
Total Earned on Upwork	464	3732.651	12726.197	30	704.167	44.478	0	100,000	100,000
Number of Jobs Completed on Upwork	472	16.763	50.347	2	5.86	2.965	0	667	667
Job Success (%)	180	92.072	10.09	97	93.931	4.448	56	100	44
Average Rating from Last 5 Jobs	274	4.85	0.568	5	4.968	0	0	5	5
From US/Canada (1 = yes)	475	0.038	0.191	0	0	0	0	1	1
From Europe (1 = yes)	475	0.063	0.244	0	0	0	0	1	1
From Asia (1 = yes)	475	0.823	0.382	1	0.903	0	0	1	1
From Latin America (1 = yes)	475	0.008	0.091	0	0	0	0	1	1
Fluent in English (1 = yes)	474	0.949	0.219	1	1	0	0	1	1

Table S15. Regression analyses of the choice of applying to the advanced task in the task phase in the field experiment, by gender, treatment condition, and number of attempted data entries in part 1 of the test project.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Outcome variable:	Applied for advanced task									
Sample:	Men	Women	Full sample	Full sample	Men-Opt in	Men-Opt out	Women-Opt in	Women-Opt out	Full sample	Full sample
Woman			-0.150**	-0.090					-0.405***	-0.308**
			(0.063)	(0.067)					(0.104)	(0.120)
Opt-out	-0.009	0.094	-0.009	0.005					0.042	0.013
	(0.052)	(0.074)	(0.054)	(0.057)					(0.095)	(0.101)
Opt-out: Woman			0.092	0.036					0.373***	0.321***
			(0.079)	(0.089)					(0.066)	(0.082)
Attempted tasks in stage 1					0.018	0.003	0.108***	-0.046**	0.019	0.024
					(0.016)	(0.015)	(0.028)	(0.022)	(0.017)	(0.018)
Attempted tasks in stage 1: Woman									0.080**	0.065**
									(0.032)	(0.033)
Attempted tasks in stage 1: Opt-out									-0.016	-0.004
									(0.023)	(0.024)
Attempted tasks in stage 1: Woman: Opt-out									-0.127***	-0.113***
									(0.041)	(0.044)
Overconfident				0.074						0.120**
				(0.046)						(0.049)
Underconfident				0.027						0.009
				(0.090)						(0.093)
Observations	304	173	477	422	149	155	89	84	477	422
(Pseudo) R2	0.00008	0.00700	0.0115	0.0110	0.00827	0.000160	0.159	0.0419	0.0539	0.0451

Notes: The table reports estimated marginal effects from probit regressions, with robust standard errors. The variable "Overconfident" takes a value of one if a freelancer predicted that their performance was in the top quartile but their actual performance was not, and zero otherwise. The variable "Underconfident" takes a value of one if a freelancer predicted that their performance was not in the top quartile but their actual performance was, and zero otherwise. * = p < .05, ** = p < .01, *** = p < .001.

REFERENCES

1. Niederle, M. & Vesterlund, L. Do Women Shy Away From Competition? Do Men Compete Too Much? *Q. J. Econ.* **122**, 1067–1101 (2007).
2. Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* **10**, 171–178 (2007).
3. Marteau, T. M. & Bekker, H. The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *Br. J. Clin. Psychol.* **31**, 301–306 (1992).
4. Spielberger, C. & Gorsuch, R. *State-trait anxiety inventory for adults: Manual and sample: Manual, instrument and scoring guide*. (Consulting Psychologists Press, 1983).
5. Tluczek, A., Henriques, J. B. & Brown, R. L. Support for the reliability and validity of a six-item state anxiety scale derived from the State-Trait Anxiety Inventory. *J. Nurs. Meas.* **17**, 19–28 (2009).
6. Johnson, S. K., Murphy, S. E., Zewdie, S. & Reichard, R. J. The strong, sensitive type: Effects of gender stereotypes and leadership prototypes on the evaluation of male and female leaders. *Organ. Behav. Hum. Decis. Process.* **106**, 39–60 (2008).
7. Koenig, A. M. Comparing Prescriptive and Descriptive Gender Stereotypes About Children, Adults, and the Elderly. *Front. Psychol.* **9**, (2018).
8. Prentice, D. A. & Carranza, E. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychol. Women Q.* **26**, 269–281 (2002).
9. Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E. & Nauts, S. Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *J. Exp. Soc. Psychol.* **48**, 165–179 (2012).
10. Brescoll, V. L. Leading with their hearts? How gender stereotypes of emotion lead to biased evaluations of female leaders. *Leadersh. Q.* **27**, 415–428 (2016).
11. Moss-Racusin, C. A. & Rudman, L. A. Disruptions in Women's Self-Promotion: The Backlash Avoidance Model. *Psychol. Women Q.* **34**, 186–202 (2010).
12. Ramarajan, L., Rothbard, N. P. & Wilk, S. L. Discordant vs. Harmonious selves: The effects of identity conflict and enhancement on sales performance in employee-customer interactions. *Acad. Manag. J.* **60**, 2208–2238 (2017).
13. Schmader, T. Gender Identification Moderates Stereotype Threat Effects on Women's Math Performance. *J. Exp. Soc. Psychol.* **38**, 194–201 (2002).
14. Popiel, P. "Boundaryless" in the creative economy: assessing freelancing on Upwork. *Crit. Stud. Media Commun.* **34**, 220–233 (2017).
15. Eha, B. P. The Freelance Economy Is Booming. But Is It Good Business? *Entrepreneur* (2013). Available at: https://www.entrepreneur.com/article/229277. (Accessed: 19th April 2021)
16. Brier, E. & Pearson, R. Upwork's SVP of Marketing Explains What It Takes To Perfect An Offering That Relies On People. *TechDay* (2017). Available at: https://techdayhq.com/community/articles/upwork-s-svp-of-marketing-explains-what-it-takes-to-perfect-an-offering-that-relies-on-people. (Accessed: 19th April 2021)
17. Woodward, A. Snag Announces Appointment Of Fabio Rosati As Chairman And CEO | Snagajob. *Snagajob* (2018). Available at: https://www.snagajob.com/blog/post/snag-announces-appointment-of-fabio-rosati-as-chairman-and-ceo. (Accessed: 19th April 2021)

18. McCaw. The Outlook of the Gig Economy. *Wonder* (2020). Available at: https://askwonder.com/research/outlook-gig-economy-wxvivqigu. (Accessed: 19th April 2021)
19. Lim, J. Gig economy work in Canada is growing, Stats Can says. *iPolitics* (2019). Available at: https://ipolitics.ca/2019/12/16/gig-economy-work-in-canada-is-growing-stats-can-says/. (Accessed: 19th April 2021)
20. Berliner, U. Millions Turn To Freelancing During The Pandemic, Trend May Last. *npr* (2020). Available at: https://www.npr.org/2020/09/16/912744566/jobs-in-the-pandemic-more-are-freelance-and-may-stay-that-way-forever. (Accessed: 19th April 2021)
21. Pofeldt, E. Pandemic Fuels A Freelancing Boom. *Forbes* (2020). Available at: https://www.forbes.com/sites/elainepofeldt/2020/09/29/pandemic-fuels-a-freelancing-boom/?sh=3507d1246625. (Accessed: 19th April 2021)
22. Maurer, R. How the Coronavirus Has Changed Freelancing. *SHRM* (2020). Available at: https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/how-the-coronavirus-has-changed-freelancing.aspx. (Accessed: 19th April 2021)
23. Burbano, V. C. Social responsibility messages and worker wage requirements: Field experimental evidence from online labor marketplaces. *Organ. Sci.* **27**, 1010–1028 (2016).
24. Burbano, V. C. The demotivating effects of communicating a social-political stance: Field experimental evidence from an online labor market platform. *Manage. Sci.* **67**, 1004–1025 (2021).
25. Lyons, B. J., Pek, S. & Wessel, J. L. Toward a 'sunlit path': Stigma identity management as a source of localized social change through interaction. *Acad. Manag. Rev.* **42**, 618–636 (2017).
26. Leung, M. D. & Koppman, S. Taking a Pass: How Proportional Prejudice and Decisions Not to Hire Reproduce Gender Segregation. *Am. J. Sociol.* **124**, 762–813 (2018).