Selection into Self-Improvement and Competition Pay: Gender, Stereotypes, and Earnings Volatility

David Klinowski[†]

Santiago Centre for Experimental Social Sciences

University of Oxford (Nuffield College), and

Universidad de Santiago de Chile

October 2018

Abstract

We examine whether men and women differ in their willingness to select into a contract that pays upon improving one's past performance. Experiment participants choose to perform a task under either a regular piece rate, or a larger piece rate provided they improve relative to a previous round. Women are less willing than men to select into self-improvement pay, and this gender gap is largely explained by higher risk aversion and (to a smaller extent) lower self-confidence. High earnings volatility widens the gender gap, and makes self-improvement pay less attractive than competition pay. We find no effect of gender stereotypes in the willingness to sort into self-improvement. The results provide insight into the feasibility and potential of using self-improvement contracts as gender-neutral incentive mechanisms.

Keywords: Gender, self-improvement, competitiveness

JEL codes: C91, J16, J31, D02

[†] Email: dklinowski@gmail.com. Address: Concha y Toro 32C, Santiago, Chile. I am grateful for valuable comments from participants of the 2017 IMEBESS Conference and the 2017 Antigua Experimental Economics Conference, and from four anonymous referees.

1. Introduction

Across many countries, women continue to earn less than men for comparable jobs, and to be underrepresented in high-paying occupations and leadership positions (Blau and Kahn, 2017; Bertrand, 2018). One explanation for this disparity is the possibility that women have a lower taste for competition, which may make them more reluctant to enter competitive careers and to seek promotions and pay raises (Niederle and Vesterlund, 2011). Indeed, a literature starting with Niederle and Vesterlund (2007) finds that, when performing tasks for pay, women are less willing than men to enter tournaments against other individuals, preferring instead a piece rate that depends on individual performance alone.[1] More recent work finds that a distaste for competition may affect actual labor market outcomes, as tournament entry predicts career choices and earnings differentials (Buser et al., 2014; Reuben et al., 2015; Buser et al., 2017; Kamas and Preston, forthcoming).

As much as the labor market rewards those who seek to, and do, outperform the rest, it often also rewards individuals who better themselves, independently of others' performance. The drive to challenge oneself to do better than before is regarded as a common trait among business top performers, and a trait that managers ought to nurture in their employees in order to promote professional growth (Harvard Business Review, 2016). If men and women differ in their taste for self-improvement, this difference may constitute an additional reason for the gender gap in the labor market. But if instead we find that women are as eager as men to select into challenges that pay upon self-improvement, we may be able to use this knowledge to design policies and incentives that help to close the gender gap in the labor market while potentially also promoting productivity growth.

This paper investigates whether men and women differ in their willingness to select into a self-improvement contract, and if so, what accounts for the difference. We conduct a laboratory experiment where participants must choose, prior to performing a task, between a contract that pays a regular piece rate regardless of performance, and a contract that pays twice the piece rate

[1] See for instance Cason et al. (2010), Healy and Pate (2011), Kamas and Preston (2012); Niederle et al. (2013), Brandts et al. (2014), Sutter and Rützler (2014), Wozniak et al. (2014), Petrie and Segal (2015), Saccardo et al. (forthcoming). A comprehensive review is given by Niederle (2016) and Dariel et al. (2017).

but only if performance improves relative to a previous round. We compare the rate at which men and women select into self-improvement pay. We measure the participants' abilities at the task, self-confidence, risk and ambiguity preferences, and perceptions about which gender dominates at the task, to explore whether these influence their compensation choices. A feature of our design is that we create an individually-calibrated risk-preference elicitation that approximates the risk involved in the decision to sort into self-improvement, but removes the need to self-improve. By comparing the participants' willingness to accept the contract to their willingness to accept the similar, purely monetary, bet, we can classify participants as self-improvement-averse, self-improvement-seeking, or consistent, and can examine whether men and women differ in the distribution of types.

We find that women are less willing than men to select into self-improvement. Most of this gap is explained by differences in risk aversion and self-confidence, and the residual gap after controlling for these and other factors in a regular regression framework is insignificant. An alternative analysis comparing men's and women's distribution of types also suggests no gender differences in preferences over self-improvement *per se* holding constant other participant characteristics.

We also examine the sensitivity of the results to the task performed by the participants. We use two tasks that vary in their perceived femaleness according to the participants, and in the earnings volatility that can be expected under self-improvement. Finally, we conduct sessions in which we replace the self-improvement contract with one that pays upon outperforming another participant, as is typical in the literature on gender and competitiveness. This allows us to contrast our results to this literature.

Closest to our paper are Apicella et al. (2017) and Carpenter et al. (2017), who also examine gender differences in sorting into self-improvement. As we mention in more detail in the discussion section, when we employ a similar task to theirs, our results are consistent: we find a small gender gap in willingness to self-improve that is explained largely by risk preferences, as in Apicella et al. (2017), and find females to be significantly more willing to sort into self-improvement than into competition, as in Carpenter et al. (2017). But we find that these results are likely to depend on the earnings volatility of the task. By introducing a task with higher volatility in our experiment,

we find that the gender gap in willingness to self-improve can grow substantially, and that this volatility may cause females to shy away from self-improvement more than they do from competition. This is in line with work by Flory et al. (2015), who find that females shy away from jobs with uncertain pay, and in line with the recent assessment of the literature by Bertrand (2018), who concludes that "higher levels of risk aversion among women may get them to shy away from jobs with greater earnings volatility". It is worth noting that Apicella et al. (2017) and Carpenter et al. (2017) also examine the effects of self-improvement on performance, which our design is unable to investigate.

We also contribute to the literature by studying the role of gender stereotypes. In the context of contributing ideas to a group, Coffman (2014) shows that individuals (both males and females) become less confident in their ability to answer correctly in areas that are stereotypically outside of their gender domain. Thus, it seems reasonable and important to ask whether individuals become less confident in their ability to improve and less willing to select into self-improvement pay if they must perform in what they perceive to be a task incongruent with their gender. This is also relevant for policy, for understanding the potential effect of self-improvement pay in different domains and occupations. We examine this question, and find that stereotypes do not affect the willingness to self-improve in our experiment, even in a sample where stereotypes are present.

In what comes next in the paper, Section 2 describes the experimental design, Section 3 presents the results, and Section 4 concludes with a discussion. An online appendix contains supplementary analysis and the experiment instructions.

2. Experimental Design

Our laboratory experiment consists of four incentivized parts plus an unincentivized questionnaire. Participants know at the start that there are these many parts to the experiment, but receive instructions for each part only at the beginning of the corresponding part. Four treatments are randomly assigned in a between-subjects design, with all participants in a given session receiving the same treatment. Treatments vary in the real-effort task participants are asked to perform (a math task or a verbal task), and in the compensation offered (a self-improvement contract or a competition contract, with the alternative always being regular piece rate). We describe a session

with the math task and the self-improvement contract, and note how other treatments vary when relevant.

In Part 1, participants have 5 minutes to add sets of 5 randomly generated 2-digit numbers (as in Niederle and Vesterlund, 2007), and earn 40¢ per correct sum. At the end of the 5 minutes, each participant receives feedback on the number of correct sums she solved and her corresponding earnings. Before they begin Part 1, participants practice with the task (with no incentives) for 90 seconds.

For sessions with the verbal task, participants see a set of 6 letters and have 5 minutes to form words using these letters only. Words must not repeat letters, and must have at least 3 letters. Participants earn 40¢ per valid word they form.

In Part 2, participants again have 5 minutes to perform the task. Before they start, they must choose one of the following two options for generating earnings in this part.

- Option A: Receive 40¢ per correct sum, regardless of the number of sums solved.
- Option B: Receive 80¢ per correct sum, provided the participant solves more sums than the number solved in Part 1. Otherwise the participant earns nothing in Part 2.

For sessions with the competition contract rather than the self-improvement contract, Option B offers a piece rate of 80¢ provided the participant solves correctly a greater number of sums than the number solved in Part 1 by another anonymous, randomly-selected participant in the session.

After selecting a contract, but before performing the task, participants report how many problems they expect to solve correctly in Part 2, and with what probability (a percentage integer from 0 to 100) they expect to solve at least one more sum in Part 2 than they solved in Part 1. To avoid strategic interactions with performance in Part 2, we do not incentivize answers to these questions. These questions elicit measures of the participants' confidence in their ability to improve. After responding, participants perform the task for 5 minutes, and receive feedback on their number of correct sums and their corresponding earnings in Part 2.

For sessions with the competition contract, participants report (i) how many problems they expect to solve in Part 2, (ii) the probability with which they expect to beat a random opponent, and (iii) their guess of their rank among all participants in the session based on Part-1 performance.

Part 3 elicits risk preferences in a choice as close as possible to the one faced in Part 2, presenting participants with similar risks and stakes as Options A and B, but removing the need to perform the task. Thus, in Part 3 participants are faced with the following two options.

- Receive a sure payment of X.
- Make a bet that pays Y with probability p and nothing with probability $1 - p$.

X and Y are individually calibrated for each participant, so that X is her earnings from Part 1, and Y is the earnings she would receive in Part 2 under the self-improvement contract were she to solve exactly one more sum in Part 2 than she solved in Part 1. Under this calibration, the sure payment of X approximates Option A in Part 2, while the bet that pays either Y or 0 approximates the potential earnings under self-improvement pay.[2]

In Part 3, participants must indicate the minimum probability p for which they prefer to make the bet over receiving the sure payment X. The probability reported must be a percentage integer between 0 and 100. To incentivize this decision, after the participant reports her p, the experimenter randomly draws an integer probability value q between 0 and 100, which becomes the actual probability with which the bet pays Y. If q is smaller than the reported p, the participant receives X; otherwise, the participant enters the bet, and earns either Y or 0, depending on the outcome of the bet. To determine the outcome of the bet, the experimenter randomly draws a second number, between 1 and 100. If the number is in $[1, q]$ the bet pays Y; if the number is in $[q + 1,100]$ the bet pays 0. Just as in Part 2, where the participant always learns whether she improved her score relative to Part 1 regardless of her choice of contract, in Part 3 the participant always receives feedback on the outcome of the bet regardless of whether she entered the bet.

[2] The correspondence between Option A in Part 2 and the sure payment of X in Part 3 rests on the assumption that in Part 2 the participant is certain that she can replicate her Part-1 performance. In the Appendix we examine the validity of this assumption.

Part 4 elicits ambiguity preferences by presenting participants with a multiple price list between a risky bet and an ambiguous bet. The risky bet is a 50-percent chance of receiving \$4 and a 50-percent chance of receiving nothing. The ambiguous bet is an undisclosed probability (known to be between 0 and 100 percent) of receiving Z and the complement probability of receiving nothing. The price list consists of 21 items that keep the risky option fixed and vary Z from \$2.35 to \$7.00 in increments of 23¢. The items are listed downward in increasing order of Z.

We implement this elicitation using two physical jars, each containing 100 marbles. The risky jar contains 50 red marbles and 50 black marbles, with the contents visible and disclosed to the participants. The ambiguous jar contains an undisclosed number of red and black marbles; the number of marbles of each color can be anywhere between 0 and 100, and they add to 100. The contents of the ambiguous jar are veiled. Participants are asked to select the color on which they want to bet, and the item in the list (1 to 21) at which they prefer to switch from drawing a marble from the risky jar to drawing a marble from the ambiguous jar. The experimenter then randomly selects an item from the list, and draws a marble from the jar chosen by the participant for that item based on her reported switch point. If the color drawn matches the color selected by the participant, the participant receives either \$4 or Z, depending on whether the marble was drawn from the risky jar or the ambiguous jar; if the color does not match the one selected by the participant, the participant receives nothing.

The session concludes with an unincentivized questionnaire. We ask participants whether they would hypothetically prefer to receive \$4 with certainty or to make a bet that pays \$8.80 with 50-percent probability and \$0 with 50-percent probability. This question serves as an alternative risk-aversion measure that involves the same stakes for all participants. We also ask participants whether they think male or female participants on average perform better at the task, with a 7-point Likert scale of the form "women(men) give a substantially/somewhat/slightly larger number of correct answers than men(women)" and a middle option for "no gender difference". This question captures beliefs about the femaleness of the task. Finally, participants provide their student status, major, occupation, year of birth, and gender.

At the end of the session, either Part 1, 2, or 3 is randomly selected for payment (participants knew this feature of the design from the beginning of the session). Earnings from this part are added to those from Part 4, and to a show-up fee of $3.15. Average total earnings were $9.75. Sessions lasted approximately 45 minutes. The experiment was programmed in oTree (Chen et al., 2016) with recruitment through ORSEE (Greiner, 2015). The experiment was conducted at the Santiago Centre for Experimental Social Sciences in Chile.[3] The self-improvement sessions were conducted between November 2016 and January 2017, while the competition sessions were conducted in November and December 2017. A total of 341 undergraduate students (164 males) participated.

3. Results

3.1. Summary Statistics

We begin by presenting descriptive statistics of the data. Table 1 shows means and proportions by gender and contract type. For both tasks, males and females perform similarly under piece rate: over all sessions, performance in Part 1 is, for math, 8.89 for males and 8.70 for females (p=0.732), and, for verbal, 15.53 for males and 14.84 for females (p=0.393). See Figure A1 in the Online Appendix for the cumulative distributions. For both tasks in self-improvement sessions, males and females report similar levels of optimism about their ability to improve in Part 2 relative to Part 1, measured both by the number of answers by which they expect to improve, and the reported likelihood of improving. In competition sessions, males on average report a higher likelihood than females of beating an opponent in the math task, while there are no gender differences in the verbal task, or in the guessed rank in either task. In both tasks, males and females self-improve at equal rates (in self-improvement sessions) and outperform their opponent at equal rates (in competition sessions). Both males and females tend to believe that males do better than females on the math task, and that females do better than males on the verbal task. Females are less likely than males to prefer the risky option in the questionnaire. Males and females switch from the risky option to the ambiguous option on average at a similar row in Part 4, suggesting no differences in ambiguity aversion. Finally, males are approximately twice as likely as females to be a STEM major, which mirrors the gender proportion in STEM fields in Chilean universities (SIES, 2017). In the analysis, we control for a STEM major indicator (results are largely unchanged without this control).

[3] Throughout the paper we express earnings in US dollars (at the approximate exchange rate at time of the experiment), but in the experiment we use Chilean pesos.

Table 1: Descriptive statistics

	Self-Improvement			Competition		
	Male	Female	p-value	Male	Female	p-value
Part 1 (piece-rate) performance						
Math	9.58	8.28	0.066	7.79	9.11	0.095
Verbal	15.81	14.98	0.447	15.18	14.65	0.674
Expected improvement in Part 2 relative to Part 1						
Math	1.21	0.88	0.323	1.65	1.07	0.321
Verbal	0.47	-0.39	0.131	0.12	-1.8	0.030
Reported chance of improving/ beating an opponent						
Math	0.68	0.66	0.771	0.60	0.48	0.009
Verbal	0.58	0.59	0.903	0.57	0.51	0.168
Guessed rank in Part 1						
Math	-	-	-	0.42	0.48	0.287
Verbal	-	-	-	0.36	0.41	0.267
Self-improve / Beat opponent*						
Math	0.55	0.42	0.226	0.50	0.53	0.823
Verbal	0.19	0.24	0.614	0.32	0.30	1.000
Perceived task femaleness (-3 to 3)						
Math	-0.23	-0.14	0.530	-0.12	-0.33	0.211
Verbal	0.86	0.80	0.746	0.74	0.53	0.344
Agrees with task gender stereotype*						
Math	0.28	0.21	0.481	0.24	0.27	0.799
Verbal	0.63	0.47	0.146	0.50	0.53	1.000
Prefers risky choice*	0.58	0.40	0.014	0.54	0.29	0.003
Ambiguity tolerance score	10.59	11.07	0.390	11.03	10.81	0.743
STEM major*	0.47	0.25	0.002	0.41	0.18	0.002
Sample size						
Math	53	43	-	34	45	-
Verbal	43	49	-	34	40	-

Notes: Values indicate means, or proportion of participants for variables with "*". For the calculation of the expected improvement in Part 1 relative to Part 2 in the *Competition* sessions, we remove one observation from a female participant who reports an expected performance in Part 2 of 1250 sums. Guessed rank is normalized by the session size, so that it is a value from 0 to 1, where 0 is the top performer and 1 the bottom performer in the session. P-values from t-tests for means and Fisher's exact tests for proportions.

3.2. Differences between the Math and Verbal Tasks

Before going in depth into the results, we discuss some important differences between the math and the verbal tasks that matter when drawing conclusions form the results. In choosing these two tasks, our design goal was to select tasks that varied in terms of the participants' perceptions of the femaleness of the task, in order to explore the influence of gender stereotypes on compensation choices. We succeeded in this respect: as reported previously, both male and female participants consider the math task to be male-dominated and the verbal task to be female-dominated. The average perceived femaleness is -0.21 for the math task, and 0.73 for the verbal task, both

significantly different from 0 ($p<0.001$ from t-tests). But an additional, incidental feature of our choice of tasks is that difficulty varies from Part 1 to Part 2 more so in the verbal task than the math task. Adding five 2-digit numbers is essentially equally challenging regardless of the numbers presented. But one's ability to form words from sets of 6 letters depends crucially on the specific letters shown. The sets of letters shown to participants in Parts 1 and 2 are the two sets of letters with the largest number of valid answers we could find. We selected these sets in order to maintain difficulty as similar as possible across parts within the verbal task, and to avoid ceiling effects whereby participants gave all the valid answers before the five minutes elapsed.[4] Yet, difficulty did vary from Part 1 to Part 2 in the verbal task much more than in the math task, as reflected in performance. In the math task, Part-1 mean performance is 8.79, with standard deviation of 3.46, while Part-2 mean performance is 9.15, with standard deviation of 3.31. The standard deviation is statistically similar in Part 1 and Part 2 ($p=0.540$), and the change in means from Part 1 to Part 2 corresponds to an improvement of merely 4 percent. In the verbal task, on the other hand, Part-1 mean performance is 15.16, with standard deviation of 5.25, while Part-2 mean performance is 12.80, with standard deviation of 4.45. The standard deviation in Part 1 is statistically different from that in Part 2 ($p=0.034$), and the change in means from Part 1 to Part 2 corresponds to a decrease of 16 percent. This indicates that variability in difficulty is larger in the verbal task that in the math task. Participants appear to have recognized this difference: they report a lower likelihood of self-improving in the verbal task than in the math task (58.3 percent vs. 67.0 percent, $p=0.004$), while their reported likelihood of beating a random opponent—which should not be influenced by round-to-round changes in difficulty, as all participants experience the same changes—does not vary by task (53.6 percent for verbal, 53.2 percent for math, $p=0.888$).

It is important to bear in mind that tasks differ in the uncertainty participants experience over how difficult Part 2 will be relative to Part 1, because this uncertainty may affect compensation choices. Uncertainty over whether one can improve past performance may make the self-improvement contract less attractive relative to the regular piece rate, but may not affect preferences over competition (since one does not need to self-improve in order to outperform an opponent). Gender stereotypes of the task may or may not affect self-confidence in one's ability to improve or beat

[4] The letters shown are "ETROCA" for Part 1 (118 valid answers) and "OASNEM" for Part 2 (65 valid answers). The answers were obtained from <http://www.sensagent.com/en/anagrams-dictionary>.

an opponent, or one's willingness to select into a specific contract. These are empirical questions. As the math and verbal tasks differ both in terms of the level of uncertainty over difficulty and the stereotypes associated with the task, in the analysis we are careful not to attribute differences in compensation choices across tasks precisely to either single factor. Yet, the fact that tasks vary along these two dimensions, and that we can measure these dimensions (with performance variability and beliefs about the ability to improve on the one hand, and beliefs about the femaleness of the task on the other hand), gives us an opportunity to recognize and explore the influence these factors may have on preferences over self-improvement and competition pay, especially within task. Recently, Bertrand (2018) highlights the potential relevance and interplay of both of these factors in explaining gender gaps in the labor market, when, in reviewing the literature, she notes that "higher levels of risk aversion among women may get them to shy away from jobs with greater earnings volatility", and that "these gender differences in psychological attributes might not be fixed traits, but rather are more likely to manifest themselves in tasks or activities that are perceived as more 'male'."

3.3. Baseline Gender Differences in Compensation Choices

Table 2 reports the fraction of participants who select into self-improvement and competition for each task, overall and by gender. In the math task, the fraction of participants who select into self-improvement is larger than the fraction of participants who select into competition (56 percent vs. 30 percent, $p<0.001$ Fisher's exact test). The reverse is true for the verbal task, as fewer participants select into self-improvement than into competition (though not statistically so, 27 percent vs. 35 percent, $p=0.311$). The difference-in-difference in entry is highly significant ($p<0.001$, from an OLS regression of compensation choice on indicators of contract type, task, and their interaction). Separately, both males and females are less likely to select into self-improvement than into competition, although not significantly (males by 23 percentage points and females by 5 percentage points, $p=0.354$ and $p=0.792$ from Fisher's exact tests). Thus, a first finding is that, in our experiment, self-improvement is not necessarily more attractive than competition (relative to regular piece rate). We explore possible explanations for this in the following sections. This finding is potentially relevant for understanding the take-up we might expect to see from mechanisms that seek to move individuals away form a piece-rate regime.

Table 2: Baseline Compensation Choice by Gender

	Self-Improvement				Competition			
	Overall	Male	Female	M-F p-value	Overall	Male	Female	M-F p-value
Math	0.56	0.62	0.49	0.218	0.30	0.47	0.18	0.007
Verbal	0.27	0.37	0.18	0.060	0.35	0.50	0.23	0.016

Notes: Fraction of participants who select into self-improvement or competition. Overall refers to both genders combined. P-values from Fisher's exact tests of the difference in choices between males and females.

In terms of gender differences, Table 2 shows that females select into both self-improvement and competition at a lower rate than males—directionally so for self-improvement in the math task, and statistically so for the three other cells. The size of the gaps, and the resulting underrepresentation of females among participants who enter into a given contract, seem economically relevant in all four cases: relative to females, 27 percent more males enter into self-improvement in the math tasks, 106 percent more males enter into self-improvement in the verbal task, 161 percent more males enter into competition in the math task, and 117 percent more males enter into competition in the verbal task. To examine the significance and the determinants of these gaps more precisely, we conduct regression analysis. We predict the participant's compensation choice for each contract type and task separately, with probit regressions and clustering standard errors at the session level. All regressions control for a STEM major indicator. Results appear in Table 3 for the math task and Table 4 for the verbal task.

The *Baseline* specification includes as regressors only a female indicator, and thus shows the gender gap in entry without accounting for potential explanatory factors. In the math task, females are 12 percentage points less likely than males to select into self-improvement ($p=0.084$), and 28 percentage points less likely to select into competition ($p<0.001$). The difference-in-difference is significant ($p=0.005$). In the verbal task, the gender gap in entry is 19 percentage points for self-improvement ($p=0.007$), and 32 percentage points for competition ($p<0.001$). The difference-in-difference is not significant ($p=0.523$). Thus, for the math and verbal tasks in our experiment, females shy away from both self-improvement and competition pay relative to males. In the math task, the gender gap in entry shrinks significantly under self-improvement relative to competition, while in the verbal task the gap does not significantly differ between self-improvement and competition.

Table 3. Probability of Selecting into Self-Improvement and Competition Pay, Math Task

	Self-Improvement							Competition						
	Baseline	Ability	Confid.	Risk	Ambig.	Stereot.	Full	Baseline	Ability	Confid.	Risk	Ambig.	Stereot.	Full
Female	-0.122*	-0.123*	-0.143	-0.066	-0.120*	-0.124*	-0.067	-0.277***	-0.299***	-0.202***	-0.244***	-0.276***	-0.283***	-0.179**
	(0.039)	(0.063)	(0.094)	(0.096)	(0.071)	(0.072)	(0.083)	(0.057)	(0.069)	(0.058)	(0.070)	(0.063)	(0.059)	(0.082)
Performance in Part 1		-0.001					-0.013		0.014					-0.001
		(0.011)					(0.014)		(0.021)					(0.020)
Expected improvement			0.309***				0.276***			0.022*				0.028***
			(0.050)				(0.047)			(0.013)				(0.009)
Belief chance improving			-0.001				-0.002							
			(0.001)				(0.002)							
Belief chance beat opponent										0.002				0.001
										(0.003)				(0.003)
Guessed rank in Part 1										-0.096*				-0.109**
										(0.055)				(0.046)
Chose risky option				0.321***			0.204***				0.129			0.144
				(0.092)			(0.050)				(0.134)			(0.116)
Ambiguity aversion					-0.003		0.006					-0.016		-0.016
					(0.011)		(0.004)					(0.011)		(0.014)
Agrees with stereotype						0.010	0.099**						0.003	-0.077*
						(0.149)	(0.050)						(0.090)	(0.059)
Female*agrees stereotype						-0.111	-0.030						-0.310**	-0.157*
						(0.178)	(0.123)						(0.121)	(0.086)
Observations	96	96	96	96	96	96	96	79	79	78	79	79	79	78
Pseudo R^2	0.014	0.014	0.417	0.091	0.015	0.016	0.487	0.087	0.096	0.154	0.103	0.106	0.106	0.210

Notes: Marginal effects from probit regressions. Performance in Part 1 measured as the number of correct answers. Expected improvement is the number of correct answers the participant expects in Part 2 relative to Part 1. Guessed rank in Part 1 is a measure from 1 to 5, where 1 corresponds to the top 20 percent, …, and 5 corresponds to the bottom 20 percent. Ambiguity aversion is the switch point (1-21) in Part 4, where a larger value indicates higher ambiguity aversion. Agrees with the stereotype is an indicator of believing that males perform better than females in the math task in the experiment. All regressions control for a STEM major indicator. Standard errors clustered at the session level in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Table 4. Probability of Selecting into Self-Improvement and Competition Pay, Verbal Task

	Self-Improvement							Competition						
	Baseline	Ability	Confid.	Risk	Ambig.	Stereot.	Full	Baseline	Ability	Confid.	Risk	Ambig.	Stereot.	Full
Female	-0.191***	-0.197***	-0.144**	-0.160*	-0.190***	-0.184**	-0.116	-0.318***	-0.322***	-0.272*	-0.291***	-0.312***	-0.309***	-0.205**
	(0.070)	(0.072)	(0.072)	(0.088)	(0.069)	(0.092)	(0.101)	(0.079)	(0.079)	(0.142)	(0.080)	(0.065)	(0.059)	(0.099)
Performance in Part 1		-0.006					-0.006		-0.007					-0.015
		(0.009)					(0.008)		(0.019)					(0.017)
Expected improvement			0.049**				0.049**			0.010				0.011
			(0.019)				(0.020)			(0.021)				(0.022)
Belief chance improving			-0.002				-0.002							
			(0.002)				(0.001)							
Belief chance beat opponent										0.002				0.004
										(0.003)				(0.003)
Guessed rank in Part 1										-0.047				-0.083*
										(0.063)				(0.044)
Chose risky option				0.159			0.170				0.107			0.118
				(0.113)			(0.113)				(0.104)			(0.097)
Ambiguity aversion					-0.001		-0.003					0.012*		0.015**
					(0.010)		(0.010)					(0.007)		(0.006)
Agrees with stereotype						0.032	-0.026						0.035	0.031
						(0.131)	(0.133)						(0.125)	(0.095)
Female*agrees stereotype						0.068	0.234						-0.432**	-0.564***
						(0.184)	(0.197)						(0.219)	(0.199)
Observations	92	92	92	92	92	92	92	74	74	74	74	74	74	74
Pseudo R^2	0.039	0.043	0.116	0.066	0.039	0.042	0.158	0.079	0.082	0.093	0.088	0.087	0.121	0.197

Notes: Marginal effects from probit regressions. Performance in Part 1 measured as the number of correct answers. Expected improvement is the number of correct answers the participant expects in Part 2 relative to Part 1. Guessed rank in Part 1 is a measure from 1 to 5, where 1 corresponds to the top 20 percent, …, and 5 corresponds to the bottom 20 percent. Ambiguity aversion is the switch point (1-21) in Part 4, where a larger value indicates higher ambiguity aversion. Agrees with the stereotype is an indicator of believing that females perform better than males in the verbal task in the experiment. All regressions control for a STEM major indicator. Standard errors clustered at the session level in parentheses. *p<0.1, **p<0.05, ***p<0.01.

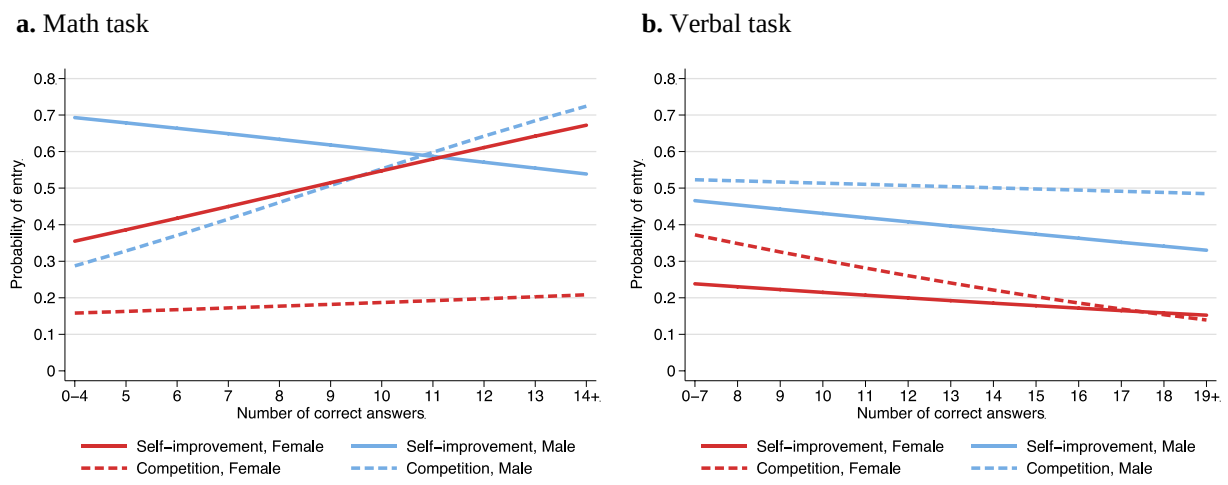

Figure 1: Probability of Selecting into Self-Improvement and Competition by Ability

3.4. Mechanisms: Ability

Previously we saw that there is no difference in ability on average, as measured by Part-1 performance, between males and females on either task. This gives us some confidence that the gender gap in entry cannot be attributed to differences in ability.[1] We confirm this with a specification that includes Part-1 performance as additional control (*Ability* column, Table 3 and Table 4). For both tasks and compensation structures, the gender gap in entry is practically unchanged relative to the baseline specification, and the coefficient on ability is not significantly different from 0.

Despite the fact that ability does not explain the gender gap in compensation choices at the mean, looking at the pattern of selection across different levels of ability gives insight into some heterogeneity with respect to ability and task. Figure 1 shows the likelihood of a participant entering into a given pay structure, by task and gender separately, for different levels of Part-1 performance. In math, self-improvement attracts more females than competition does, especially high-ability females. In fact, female entry into self-improvement resembles quite closely male

[1] One way in which we depart from Niederle and Vesterlund (2007) is that, in our experiment, participants do not perform a round of forced competition (or self-improvement). This arguably simplifies the experiment, but implies that we cannot be sure that differences in ability under competition (or under self-improvement) drive the differences in entry, since we do not measure performance under those incentive structures.

entry into competition. For males, on the other hand, self-improvement is more attractive than competition only for low-to-middle-ability participants, while for high-ability males, competition is more attractive than self-improvement. Thus, in the math task self-improvement can close, and even reverse (for high-ability individuals), the gender gap in entry we see for competition. But in the verbal task, self-improvement is equally if not less appealing than competition for all levels of ability and for both males and females. Thus, in the verbal task, self-improvement does little to change the gap in entry we see for competition. As mentioned previously, the differences across tasks may be due to differences in stereotypes associated to the tasks, or differences in earnings volatility associated with the self-improvement contract (and/or possibly other factors). As we examine several mechanisms in the following sections we try to evaluate these explanations.

3.5. Mechanisms: Self-confidence

The literature on competitiveness often finds that females are less self-confident than males, which explains at least partly why females shy away from competition (Niederle, 2016). To examine self-confidence in the competition sessions in our experiment, Figure 2 plots the fraction of participants who rank themselves—their Part-1 performance—in a given quintile within the session they participated in. For each quintile we divide participants depending on whether they are overconfident (ranked themselves better than their actual quintile), are underconfident (ranked themselves worse than their actual quintile), or guessed their quintile correctly. We see that, in the math task, males are underconfident 6 percent of the time, correct 56 percent of the time, and overconfident 38 percent of the time. If we assign them a "confidence score" of -1 for being underconfident, 0 for being correct, and 1 for being overconfident, their average confidence score is 0.32. Females, on the other hand, are underconfident 40 percent of the time, correct 24 percent of the time, and overconfident 36 percent of the time. Their average confidence score is -0.04. The distribution of confidence scores is significantly different across gender ($p<0.001$, Fisher's exact test). Thus, in the math task, males appear overconfident and females slightly underconfident.[2] In the verbal task, both males and females appear overconfident, with average confidence scores of 0.382 and 0.275 respectively. Males are underconfident 24 percent of the time, correct 15 percent

[2] In contrast to this result, work on competitiveness tends to find that both males and females are overconfident about their relative abilities, with males being more overconfident. An exception is Dreber et al. (2014), who find that male and female adolescents in Sweden tend to be underconfident.

of the time, and overconfident 62 percent of the time. Females are underconfident 23 percent of the time, correct 28 percent of the time, and overconfident 50 percent of the time. The distributions are not significantly different across gender (p=0.421, Fisher's exact test). The larger rate of overconfidence by both males and females in the verbal task results in the right-skewed distribution in Figure 2b; i.e., a "better-than-average" effect, in which more participants rank themselves in the top percentiles than is statistically possible.

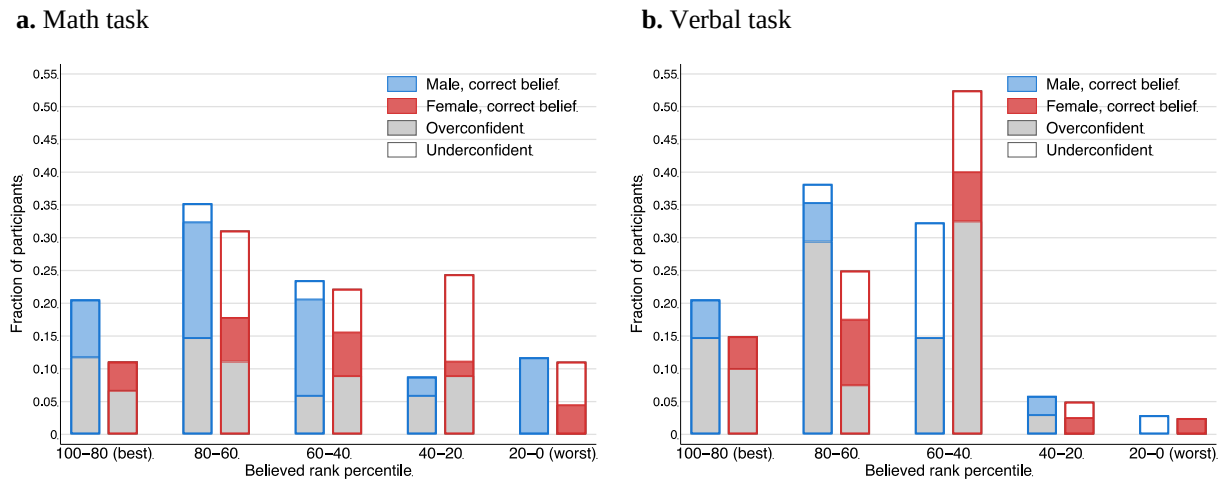

Figure 2: Self-Confidence Based on Guessed Rank

Notes: The height of the bars shows the fraction of participants who rank their Part-1 performance in a given quintile of ability in the session. The gray shade indicates the proportion of those participants who ranked themselves a better than their actual rank. The blue/red shade indicates the proportion who ranked themselves correctly. The white shade indicates the proportion who ranked themselves worse than their actual rank.

We can also explore the participants' rank beliefs about their Part-1 performance with regression analysis. In Table A1 (Panel a) in the Online Appendix, we regress the participant's rank guess on a female indicator, controlling for actual rank and a STEM major indicator. In Column 1 we see that, on average, females rank themselves significantly worse than males in the math task—by 1.41 rank places (p=0.038). In the verbal task (Column 2), there are no significant differences in how males and females rank themselves (p=0.871). A difference-in-difference regression shows that, while males' self-confidence does not change across tasks, females' self-confidence drops significantly (and differently than males') in the math task relative to the verbal task (p=0.049 in the interaction term, Column 3). Since beliefs about performance relative to other participants in

Part 1 are arguably unaffected by the variance in difficulty of the task across parts, it is reasonable to think that this drop in female self-confidence in their relative ability in the math task relative to the verbal task may be influenced by stereotypes about the femaleness of the task, rather than by the variance in difficulty of the task. Later analysis suggests that the drop in self-confidence is driven largely by participants who hold stereotypical views about the femaleness of the tasks.

In addition to looking at gender differences in overconfidence about relative ability—i.e. overplacement in Moore and Healy's (2008) terminology—we examine whether females are more or less optimistic about their ability to beat an opponent in the competition sessions. In Table A1 Panel b we replicate the previous analysis, but change the outcome of interest to the participant's reported likelihood of beating an opponent, rather than the rank guess. In the math task, males report on average a likelihood of 61 percent and females report a likelihood of 47 percent; the difference is significant (p=0.019). In the verbal task, the gap is smaller but still significant: males report a likelihood of 58 percent and females report a likelihood of 50 percent (difference p=0.050). The difference-in-difference in the gap across task is not significant at standard levels (p=0.175). Thus, we find that in the math task, females are less optimistic than males about their ability to beat an opponent, while the gap reduces somewhat but still persist for the verbal task. Later in the analysis we show that, as in the case of beliefs about relative ability, the difference in the gender gap in optimism across tasks appears to be driven to a large extent by participants who hold stereotypical views about the femaleness of the tasks.

While females in the experiment are less confident about their relative performance and less optimistic about their ability to beat an opponent, especially in the math task, we see a different pattern with respect to beliefs about the ability to self-improve. Figure 3 plots the distribution of participants over their expectations about improving their Part-2 performance relative to Part 1. We classify beliefs according to whether the participant expects to improve, expects to stay exactly the same, or expects to worsen. Looking at the height of the bars (ignoring colors within the bars), we see that, in the math task, 62 percent of males expect to improve, 22 percent of males expect to stay exactly the same, and 16 percent of males expect to worsen, while 63 percent of females expect to improve, 26 percent of females expect to stay the same, and 6 percent of females expect to worsen. The distributions are statistically different (p=0.063, Fisher's exact test), suggesting

that in the math task females are *more* optimistic than males about their ability to self-improve. In the verbal task, 42 percent of males expect to improve, 14 percent of males expect to stay the same, and 44 percent of males expect to worsen, while 28 percent of females expect to improve, 21 percent of females expect to stay the same, and 51 percent of females expect to worsen. Thus, in the verbal task, females are somewhat *less* optimistic than males, although not statistically so at standard levels (p=0.160, Fisher's exact test).

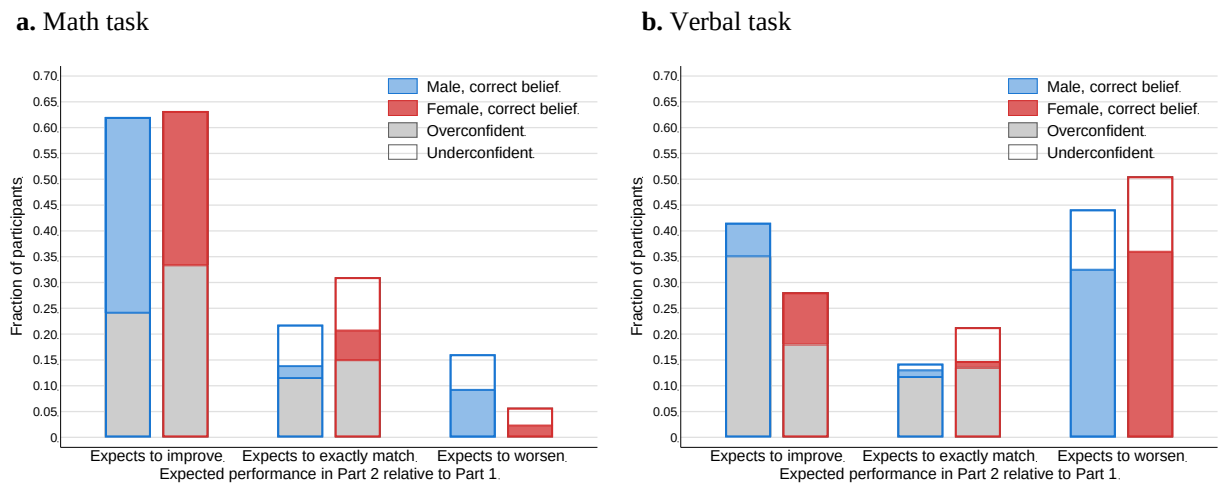

Figure 3: Self-Confidence Based on Reported Probability of Beating an Opponent

Notes: The height of the bars shows the fraction of participants who expect to improve, exactly match, or worsen their performance in Part 2 relative to Part 1. The gray shade indicates the proportion of those participants who expected a better outcome than actually occurred (namely, expected to improve but ended up exactly matching or worsening their performance, or expected to exactly match but ended up worsening their performance). The blue/red shade indicates the proportion whose expectations were correct. The white shade indicates the proportion who expected a worse outcome than actually occurred (namely, expected to worsen but ended up exactly matching or improving their performance, or expected to exactly match but ended up improving their performance).

We examine the differences in optimism about self-improvement in more detail with regressions. In Table A2, Panel a, we predict the number of correct answers a participant expects to reach in Part 2 on a female indicator, controlling for the number of correct answers in Part 1 and a STEM major indicator. Males and females have similar expectations in the math task (0.395 fewer expected correct answers for females, p=0.216), while in the verbal task females have significantly lower expectations than males (1.547 fewer expected correct answers, p=0.016). The difference-in-difference in the gender gap across tasks is significant (p=0.095). This indicates that something

about the verbal task makes females become relatively pessimistic about their ability to self-improve. Analysis presented later suggests that this cannot be explained by agreement with stereotypes of the tasks. A plausible explanation is that the increased earnings volatility in the verbal task depresses female confidence in their ability to improve.

We return to Figure 3 to examine how correct males and females are about their expectations to self-improve. The shades of the bars in Figure 3 indicate, for a given belief group, the fraction of participants who turned out to be overconfident, correct, or underconfident in their beliefs about self-improving.[3] In the math task, males are underconfident 15 percent of the time, correct 49 percent of the time, and overconfident 36 percent of the time, while females are underconfident 14 percent of the time, correct 38 percent of the time, and overconfident 48 percent of the time. The difference in distributions is not significant (p=0.228, Fisher's exact test). If we assign a confidence score as before, the average score is 0.21 for males and 0.34 for females. Thus, both genders appear overconfident, with females being slightly but not significantly more overconfident. In the verbal task, males are underconfident 13 percent of the time, correct 40 percent of the time, and overconfident 47 percent of the time, while females are underconfident 21 percent of the time, correct 47 percent of the time, and overconfident 32 percent of the time. The differences in distributions are only marginally significant (p=0.107, Fisher's exact test). The average confidence score is 0.34 for males and 0 for females. Thus, in the verbal task, now females appear slightly less confident than males. Again, a plausible explanation is that higher earnings volatility depresses females' but not males' confidence.

We can also examine beliefs about the likelihood of improving. In Table A2, Panel b, we predict the reported likelihood of improving in Part 2 relative to Part 1, on a female indicator, controlling for actual rank in Part 1 and a STEM major indicator. We see that in both tasks, males and females are equally optimistic about their likelihood of improving. In the math task, the average belief is 66 percent for males and 68 for females (p=0.497); in the verbal task the average belief is 58

[3] We classify participants as overconfident if they (i) expected to improve but ended up either exactly matching or falling short of their previous scores, or (ii) expected to exactly match their previous scores but ended up falling short. Similarly, participants are underconfident if they (i) expected to decrease but ended up either matching or improving their previous scores, or (ii) expected to exactly match their previous scores but ended up improving. In all other cases, participants' beliefs are correct.

percent for males and 59 percent for females (p=0.747). The difference-in-difference in the gap across tasks is not significant (p=0.940, column 3 in Table A2 Panel b). Analysis on stereotypes in a later section suggests that agreement with the stereotypes about the task do not explain any change in the gender gap in beliefs about the likelihood of improving across tasks.

In summary so far, with respect to beliefs that relate one to others—beliefs about relative performance and the likelihood of beating an opponent—females tend to be less confident and less optimistic than males. This is especially so in the math task, and as we explore later, the change in the gender gap in confidence and optimism across tasks seems to be influenced by agreement with stereotypes. With respect to beliefs that relate one only to oneself—beliefs about the level of improvement and the likelihood of improving—females tend to be equally if not more confident and optimistic than males in the math task, while, only for the level of improvement, their confidence and optimism drop below males' in the verbal task. In analysis reported later we find no evidence that the change across tasks is explained by agreement with stereotypes. It is plausible then that the drop in female confidence and optimism in the verbal task is related to the other key distinction in this task, namely higher earnings variability.

To conclude this section, we examine how beliefs affect compensation choices. In Tables 3 and 4, column *Confidence*, we regress the compensation choice on a female indicator as before, but now change the controls to the number of answers by which the participant expects to improve and the reported likelihood of improving. For competition sessions, we include instead the number of answers by which the participant expects to improve, the reported likelihood of beating an opponent, and the guessed rank in Part 1 from 1 to 5 (1: top 20 percent, …, 5: bottom 20 percent). For the math task, we see that one additional answer in the expected improvement is associated with an increase by 31 percentage points in the likelihood of selecting into self-improving (p<0.001). The reported likelihood of improving does not significantly affect entry (p=0.489). Relative to the baseline specification, the introduction of these controls actually increases the gender gap by 2 percentage points, likely due to the fact that, as we saw, females are more optimistic than males in their expected improvement. The gender gap in entry is only marginally significant (p=0.129). In the verbal task, one additional answer in the expected improvement is associated with an increase in the likelihood of selecting into self-improvement by 5 percentage

points (p=0.011). The reported likelihood of improving does not significantly affect entry (p=0.201). Controlling for these factors, the gender gap in entry shrinks by 25 percent and remains significant (p=0.046). For competition sessions, in the math task, the likelihood of selecting into competition increases by 2 percentage points for each additional answer by which the participant expects to improve (p=0.088), is not significantly affected by the reported likelihood of beating an opponent (p=0.475), and increases by 10 percentage points for each superior quintile the participant ranks themselves in (p=0.081). These controls explain 27 percent of the baseline gender gap in competitiveness, but a substantial gap of 20 percentage points remains (p=0.001).[4] In the verbal task, neither confidence control is statistically significant on its own. Their joint introduction shrinks the gender gap in competitiveness by 14 percent, but a significant gap of 27 percentage points remains (p=0.055).

3.6. Mechanisms: Risk Aversion

Risk-averse individuals may dislike the self-improvement and competition contracts, as both involve a risk of failing to earn money. If females are more risk averse than males, as the literature often finds (Croson and Gneezy, 2009; Charness and Gneezy, 2012), then gender differences in compensation choices may be due to differences in risk aversion. In the experiment, we have two measures risk preferences: a hypothetical binary choice between a risky bet and a safe payoff, where payoffs are the same for all participants, and an elicitation of the probability that makes the participant indifferent between a risky bet and a sure payment, where the payoffs are individually calibrated to approximate the payoffs faced by the participant in the compensation choice.

Females appear more risk averse in the hypothetical binary choice, as 57 percent of males and 35 percent of females prefer the risky bet (p<0.001, Fisher's exact test). To explore whether this difference in risk aversion explains the gender gap in selection into self-improvement and competition, we include an indicator of the risky choice as a control in the regression that predicts

[4] Thus, we are able to investigate a question Niederle and Vesterlund (2007) raise in discussing their results but are unable to answer, namely that "to the extent that there are gender differences in the participants' beliefs about their future performance and that these influence tournament entry, our study incorrectly attributes such an effect to men and women having different preferences for performing in a competition." In our study, males are more optimistic than females about their ability to beat an opponent, while females are more optimistic about their ability to self-improve. The introduction of these controls is not able to explain away the gender gap in competitiveness.

compensation choice. Results appear in Tables 3 and 4, column *Risk.* In the math task, choosing the risky bet is associated with an increase in the likelihood of selecting into self-improvement of 32 percentage points ($p<0.001$). Controlling for this factor alone reduces the gender gap in entry into self-improvement by 46 percent, and the remaining gap of 6.6 percentage points is insignificant ($p=0.493$). Risk aversion is by far the largest single explanatory factor for the gender gap in selection into self-improvement in the math task. In the verbal task, choosing the risky bet is associated with an increase in the likelihood of selecting into self-improvement by 16 percentage points, but this coefficient is not significant at standard levels ($p=0.161$). Controlling for this factor alone reduces the gender gap in entry by 16 percent, but the remaining gap of 16 percentage points remains significant ($p=0.067$). Thus, we have some evidence that risk preferences play a small to moderate role in the decision to accept a self-improvement contract. If females are more risk averse, they may be less willing than males to accept this contract.

Risk aversion seems to play a smaller role in the decision to compete.[5] The coefficients are not significant in the math task ($p=0.334$) or the verbal task ($p=0.303$), and, relative to baseline, introducing this control reduces the gender gap in entry by 12 percent in the math task and 8 percent in the verbal task, with both gaps remaining highly significant ($p\leq0.002$).

While the response to the hypothetical risk-elicitation question correlates with behavior, one may worry that this measure captures risk preferences that may affect compensation choices only too crudely, because the response is binary and not incentivized (Charness et al., 2013). As a second approach to examining whether there is a gender gap in preferences over selecting into self-improvement and competition conditional on risk preferences, we use the responses to the individually-calibrated risk elicitation. For self-improvement sessions, we compare the participant's compensation choice given her reported probability of improving, to her indifference point between a risky and a safe payoff that approximate the compensation choice, but that do not require the participant to perform the task. Comparing the two choices allows us to classify participants into three types: self-improvement averse, self-improvement seeking, and consistent. The categorization is illustrated in Table 5, and is derived as follows.

[5] As in Niederle and Vesterlund (2007).

Table 5: Participants types based on comparing their willingness to select into self-improvement, and their indifference point between a risky and a safe payoff that approximate the contract choice but do not require the participant to self-improve

	$\boldsymbol{\beta \geq \delta}$	$\boldsymbol{\beta \leq \delta}$
Selects into self-improvement	Consistent choices	Self-improvement seeking
Does not select into self-improvement	Self-improvement averse	Consistent choices

Notes: β denotes the probability with which the participant believes she will improve in Part 2 relative to Part 1, and δ denotes the minimum probability of the bet in Part 3 paying off for which the participant accepts to make the bet instead of receiving the sure payoff.

Suppose the participant completed Part 1 and is now deciding whether to select into self-improvement in Part 2. Let X be the participant's earnings in Part 1, Y be what her earnings would be in Part 2 if she selects into self-improvement and improves her Part-1 performance by 1 additional correct answer, and β be her reported probability of improving in Part 2 relative to Part 1. Purely in terms of monetary risk, the choice of contract is approximately a choice between a bet that pays Y with probability β and 0 with probability $1-\beta$, and a sure payment of X.[6] The risk elicitation in Part 3 presents the participant with a choice between these same risky and sure payments, and asks the participant to state the minimum probability of the bet paying off for which she would accept the bet over receiving the sure payoff. Denote the participant's reported probability as δ. In this risk elicitation, the participant reveals that she accepts the bet if it pays with probability of at least δ. We can contrast her preference over the bet given at least δ with her preference over contracts given β. If she accepts the bet when it pays with probability of at least δ but does not select into the self-improvement contract when her reported probability of improving is $\beta \geq \delta$, her choices suggest a distaste for the self-improvement contract net of the risk element. Similarly, if she accepts the bet when it pays with probability of at least δ but enters into self-improvement when her believed probability of improving is $\beta < \delta$, her choices suggest a taste for the self-improvement contract net of the risk element. In the remaining cases (selecting into self-improvement when $\beta \geq \delta$, and not selecting into it when $\beta < \delta$) her choices are consistent with each other.

[6] This assumes that the participant is certain that she can replicate her Part-1 performance. In the Appendix we examine the validity of this assumption.

Table 6: Gender Gap in the Probability of Being Self-Improvement Averse, Consistent, and Self-Improvement Seeking.

	Math		Verbal		Pooled	
	(1)	(2)	(1)	(2)	(1)	(2)
Averse	0.152**	0.108	-0.046	-0.119	0.044	-0.013
	(0.068)	(0.081)	(0.067)	(0.081)	(0.054)	(0.056)
Consistent	-0.105	-0.113	0.040	0.058	-0.019	-0.012
	(0.075)	(0.089)	(0.094)	(0.107)	(0.058)	(0.062)
Seeking	-0.048	0.005	0.006	0.061	-0.024	0.025
	(0.074)	(0.082)	(0.077)	(0.068)	(0.051)	(0.049)
Addit. controls	No	Yes	No	Yes	No	Yes
N	96	96	92	92	188	188

Notes: Marginal increase in the probability of females (relative to males) of being classified as self-improvement averse, consistent, or self-improvement seeking. Estimates from multinomial probit regressions that control for a STEM major indicator. Additional controls are Part-1 performance, the level of expected improvement, the reported probability of improving, the choice in the hypothetical risk elicitation, the switch point in the ambiguity elicitation, and an indicator of agreement with the stereotype (interacted with the female indicator in the *Pooled* specification). *Pooled* specifications pool data from both tasks in self-improvement sessions. Standard errors clustered at the session level in parentheses. *p<0.1, **p<0.05, ***p<0.01.

To examine how males and females distribute over the three types, we run multinomial probit regressions that predict the participant's type on a female indicator, controlling for a STEM major indicator. Results appear in Table 6 Column 1, for each task separately and also with tasks pooled together. Without controlling for additional factors, females are 15 percentage points more likely than males to be self-improvement averse (p=0.025) in the math task, while in the verbal task and with the pooled data there are no significant gender differences in the distribution of types. In later analysis we include additional controls in the regression and find no gender differences in either task separately or in the pooled data. This suggests, in an alternative analysis to the regressions in Tables 3 and 4, that after controlling for the risk involved in the compensation choice and conditional and other participant characteristics, males and females have equal preferences over self-improvement per se.

We perform a similar exercise for the competition sessions. For these sessions, we define β as the participant's reported probability of beating an opponent. The elicitation in Part 3 involved a choice between a sure payment of X and a risky bet paying either Y or 0, as defined previously. Thus, the choice in Part 3 approximates the choice in Part 2 of a participant who is certain that she

can replicate her Part 1 performance, and considers that if she beats an opponent she will do so with a score one point higher than her Part-1 performance. We classify a participant as competition averse if she accepts the bet with probability of at least δ but does not select into competition when her reported probability of beating an opponent is $\beta \geq \delta$. Similarly, the participant is competition seeking if she selects into competition when $\beta < \delta$. In all other cases we classify the participant as consistent in her choices.

Table 7: Gender Gap in the Probability of Being Competition Averse, Consistent, and Competition Seeking.

	Math		Verbal		Pooled	
	(1)	(2)	(1)	(2)	(1)	(2)
Averse	0.026	0.122*	0.050	0.108	0.041	0.125**
	(0.077)	(0.074)	(0.141)	(0.125)	(0.074)	(0.056)
Consistent	0.018	-0.010	0.183	0.145	0.090	0.036
	(0.124)	(0.089)	(0.134)	(0.104)	(0.078)	(0.071)
Seeking	-0.044	-0.112***	-0.233***	-0.253***	-0.132***	-0.161***
	(0.053)	(0.026)	(0.051)	(0.068)	(0.051)	(0.039)
Addit. controls	No	Yes	No	Yes	No	Yes
N	78	78	74	74	152	152

Notes: Marginal increase in the probability of females (relative to males) of being classified as competition averse, consistent, or competition seeking. Estimates from multinomial probit regressions that control for a STEM major indicator. Additional controls are Part-1 performance, the level of expected improvement, the reported probability of beating an opponent, guessed rank quintile for Part-1 performance, the choice in the hypothetical risk elicitation, the switch point in the ambiguity elicitation, and an indicator of agreement with the stereotype (interacted with the female indicator in the *Pooled* specification). *Pooled* specifications pool data from both tasks in competition sessions. Standard errors clustered at the session level in parentheses. *p<0.1, **p<0.05, ***p<0.01.

In Table 7 Column 1 we replicate the analysis in Table 6, but for competition sessions. Without controlling for additional factors, we find no gender differences in the likelihood that the participant is classified as competition averse or consistent, while we find that females are less likely to be competition seeking only in the verbal task (by 22 percentage points, p<0.001) and with the tasks pooled (by 13 percentage points, p=0.004). Thus, in this alternative exercise, we have some evidence that, controlling for risk preferences, males and females have similar preferences over self-improvement, while females shy away from competition, particularly in the verbal task. Later in the analysis we include additional controls and find stronger evidence for both

tasks that females are more likely to be competition averse and less likely to be competition seeking than males.

3.7. Mechanisms: Ambiguity Aversion

Beliefs about how one will do under self-improvement and competition are likely to be imprecise (rather than point) estimates. Thus, the decision to select into these contracts could be influenced by the participant's attitude toward ambiguity.[7] We proxy for ambiguity aversion in our analysis with the row at which the participant switches from preferring the risky bet to preferring the ambiguous bet in the elicitation in Part 4. This switch point is an integer from 1 to 21; a larger number indicates greater ambiguity aversion. In a regression controlling for a STEM major indicator, the average switch point is 10.8 for males and 10.9 for females; the difference is insignificant (p=0.797). To examine whether ambiguity aversion explains compensation choices and the gender gaps in them, we include the switch point as a control in Tables 3 and 4, Column *Ambiguity*. We find that ambiguity aversion plays a significant role only in the decision to compete in the verbal task: an increase in ambiguity aversion by one point is associated with an increase in the likelihood of competing by 1 percentage point (p=0.073), which is counter to the direction of the effect we would expect ambiguity preferences to have, if at all (i.e., that higher ambiguity aversion would be associated with lower likelihood of entry). We are unable to explain the direction of the effect we find. In neither task or contract type does the inclusion of ambiguity aversion explain the gender gap in entry.

3.8. Mechanisms: Stereotypes about the Femaleness of the Task

As part of the final questionnaire, participants answer whether they think males or females perform better in the task (math or verbal, depending on the session). They provide an integer from -3 to 3, such that -3/-2/-1 correspond to "men give a *substantially/somewhat/slightly* larger number of correct answers than women," 0 corresponds to "no gender difference," and 1/2/3 correspond to "women give a *slightly/somewhat/substantially* larger number of correct answers than men." Both male and female participants in the experiment tend to believe that, on average, males do better than females in the math task and females do better than males in the verbal task. In the math task,

[7] Ambiguity preferences are largely unexplored in the literature on competitiveness. An exception is Saccardo et al. (forthcoming).

the mean belief is -0.184 for males and -0.239 for females ($p=0.032$ and $p<0.001$ respectively, from t-tests). The median belief is 0 for both males and females, but a signed rank test indicates the distribution of beliefs is biased toward negative values for males and females ($p<0.001$ for both genders, from Wilcoxon signed rank tests). In the verbal task, the mean belief is 0.805 for males ($p<0.001$, t-test) and 0.674 for females ($p<0.001$, t-test). The median belief is 1 for males and 0 for females, and both distributions are biased toward positive values ($p<0.001$, Wilcoxon signed rank test). As we saw, for both tasks, these stereotypes are incorrect within the experiment, as there are no gender differences in Part-1 performance on either task.

To examine whether these stereotypes may play a role in producing gender differences in behavior in the experiment, we begin by exploring an association between agreement with the stereotype and self-confidence and optimism. We classify participants as agreeing with the gender stereotype of the task if their answer to the question is less than 0 for the math task, and greater than 0 for the verbal task. In Table A1, Panel a, Column 4, we predict the participant's guessed rank in competition sessions on a female indicator, a math task indicator, an indicator of agreement with the task, and the triple interaction (and all pairwise interactions). The positive and significant coefficient on the triple interaction suggests that the widening in the gender gap in self-confidence in the math task relative to the verbal task is explained in part by agreement with the stereotypes. Figure A2 illustrates the relationship. In the math task (Panel a), males who agree that math is male-dominated rank themselves better (i.e., down the scale in the top percent) than males who disagree, while females who agree rank themselves slightly worse than females who disagree. That is, males who agree with the stereotype increase their self-confidence while females slightly decrease it. The opposite seems to be the case for the verbal task (Panel b): males who agree that verbal is female-dominated decrease their self-confidence while female self-confidence is hardly affected. A similar pattern appears if we examine the reported likelihood of beating an opponent. Table A1, Panel b, Column 4, shows a significant coefficient on the triple interaction, which is illustrated in Figure A2 Panels c and d. Male optimism goes up for those who agree with the stereotype in the math task, while it goes down for those who agree with the stereotype in the verbal task. The reverse is the case for females. Thus, it seems that stereotypes may play role in the experiment in increasing self-confidence and optimism about relative ability when participants perceive the task as congruent with their own gender.

In Table A2, Panels a and b, Column 4, we look for a similar interaction between agreement with the stereotype and the gender gap in beliefs about the level of self-improvement and the likelihood of self-improving. In neither case do we find a significant effect, which suggests that gender stereotypes about the task play less, if any, of a role in shaping the participants' self-confidence and optimism about their ability to self-improve.

Next, we explore whether agreement with stereotypes about the task influence selection into contracts. In Tables 3 and 4, column *Stereotype*, we include the indicator of agreement with stereotype and its interaction with the female indicator in the regression that predicts entry into self-improvement and competition. Stereotypes seem to play no role in the decision to select into self-improvement in either task. On the other hand, agreement with the stereotype is related to an increase in the gender gap in selection into competition for the math task and *also* for the verbal task. This is illustrated in Figure A3. In the math task, among participants who disagree with the stereotype, 42 percent of males and 22 percent of females select into competition; the gap of 20 percentage points is significant ($p=0.008$). Among participants who agree with the stereotype, 60 percent of males and 9 percent of females select into competition; the size of the gap is now 51 percentage points ($p<0.001$). On the other hand, for the verbal task, among participants who disagree with the stereotype, 38 percent of males and 30 percent of females select into competition; the gap of 8 percentage points is not significant ($p=0.380$). Among participants who agree with the stereotype, 65 percent of males and 14 percent of females select into competition; the gap of 52 percentage points is significant ($p<0.001$).

Thus, it appears the size of the gender gap in competitiveness is substantially reduced in the math task, and practically eliminated in the verbal task, for participants who do not subscribe to the gender stereotype of the task. As we saw previously, stereotypes seem to influence self-confidence and optimism in one's relative ability, so in later analysis we explore the effect of stereotypes conditional on beliefs, as these may be driving the current result. It is notable that agreement with the belief that the verbal task is female-dominated *widens* the gender gap in competitiveness in the verbal task, which seems counterintuitive. One might have expected females to become more competitive, and males less competitive, if they perceive the task as female-dominated. We can

only speculate about this result, but one possibility is that agreement with the gender stereotype of the task is capturing agreement with more general gender stereotypes held by the participant, for instance, that competition is a male activity. Unfortunately, we are unable to explore this or other potential explanations, but in future research it may be interesting to elicit beliefs about the femaleness of competition per se, in addition to the femaleness of the task.

3.9. All Mechanisms Jointly

We conclude the analysis by considering a full model that estimates the gender gap in compensation choices controlling for all mechanisms explored before jointly. Column *Full* in Tables 3 and 4 presents the results. Of the baseline gender gap in selection into self-improvement, the inclusion of controls explains 55 percent in the math task, and 61 percent in the verbal task. The size of the residual gap is 7 percentage points for the math task (p=0.419), and 12 percentage points for the verbal task (p=0.248). Neither is statistically significant. Higher self-confidence in the level of improvement significantly predicts entry, consistently across tasks (p<0.001 for math and p=0.013 for verbal). Lower risk aversion also predicts entry, significantly in the math task (p<0.001) and only marginally significantly in the verbal task (p=0.135). As before, we do not find that agreement with the stereotype is able to explain the gender gap in entry into self-improvement.

In terms of the gender gap in selection into competition, the inclusion of all controls explains 65 percent of the baseline gap in the math task, and 64 percent of the baseline gap in the verbal task. The size of the residual gap is 18 percentage points in the math task (p=0.030) and 21 percentage points in the verbal task (p=0.039). Higher self-confidence in one's relative ability, measured by the guessed rank in Part 1, is associated with a higher likelihood of entry, consistently across tasks (p=0.017 for the math task and p=0.056 for the verbal task). Agreement with stereotypes continues to play an important role in explaining the gender gap in entry for both tasks, as indicated by the significant coefficient on the *female*agreement-with-stereotype* term. We illustrate this relationship in Figure 4. In the math task, among participants who disagree with the stereotype, 39 percent of males and 25 percent of females select into competition; the gap of 14 percentage points is not significant at standard levels (p=0.184). Among participants who agree with the stereotype, 45 percent of males and 12 percent of females enter into competition; the gap of 33 percentage points is significant (p=0.002). On the other hand, in the verbal task, among participants who

disagree with the stereotype, the gender gap is actually in favor of females, although not significantly. 30 percent of males and 38 percent of females select into competition; the gap of 8 percentage points is not significant (p=0.548). Among participants who agree with the stereotype, 62 percent of males and 14 percent of females select into competition; the gap of 48 percentage points is significant (p<0.001). Thus, conditional on other controls, females appear as competitive as males among participants who do not subscribe to the stereotype, while they are less competitive than males among participants who do subscribe to it, for both tasks in the experiment.

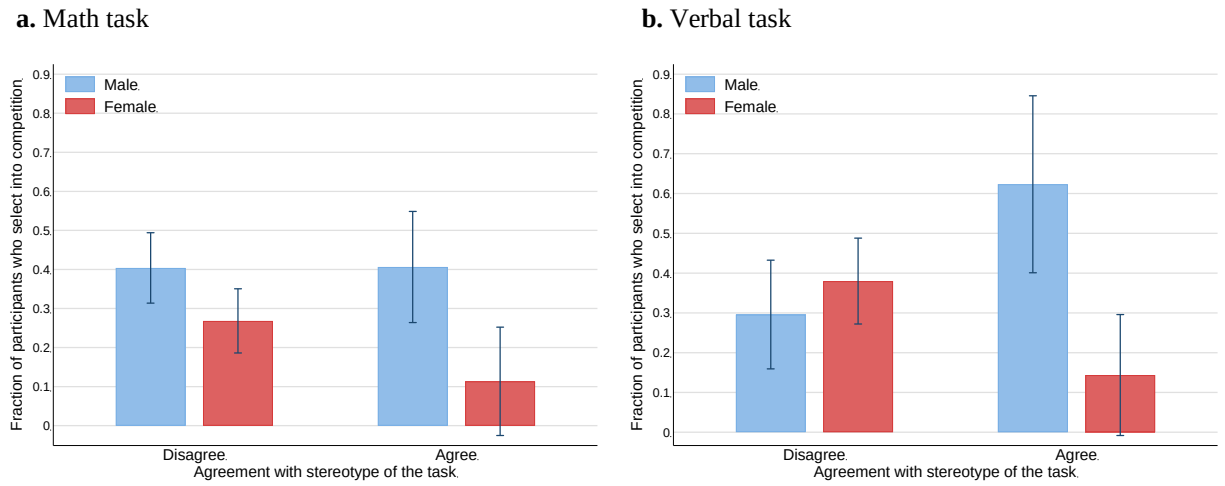

Figure 4: Gender Gap in Selection into Competition, by Agreement with Stereotype

Notes: Estimates from Tables 3 and 4, Column *Full*. Range bars are 90-percent confidence intervals.

We reach a similar conclusion if we look at the distribution of participant types, based on the comparison between compensation choices and preferences over purely monetary bets, this time controlling for other factors measured in the experiment. Table 6, Column 2, shows the gender gap in the likelihood of being classified as self-improvement averse, consistent, or self-improvement seeking (as defined previously), controlling for all factors included in the *Full* model in Tables 3 and 4. For either task separately, and with tasks pooled together, there are no gender differences in the distribution of types. This suggests that males and females have similar preferences over self-improvement, conditional on other measured preferences. The distribution is plotted in Figure 5, Panel a. Table 7, Column 2, presents a similar analysis for competition sessions. Females are significantly more likely than males to be classified as competition-averse in the verbal task, and, for both tasks, they are significantly less likely to be classified as competition-seeking. When we

pooled the tasks together, females are 13 percentage points more likely than males to be competition averse (p=0.025), and 16 percentage points less likely than males to be competition seeking (p<0.001). The distribution of types appears in Figure 5, Panel b. Thus, the differences in distributions suggest that, conditional on other preferences measured in the experiment, males and females have similar preferences over self-improvement, while females are less willing than males to compete.

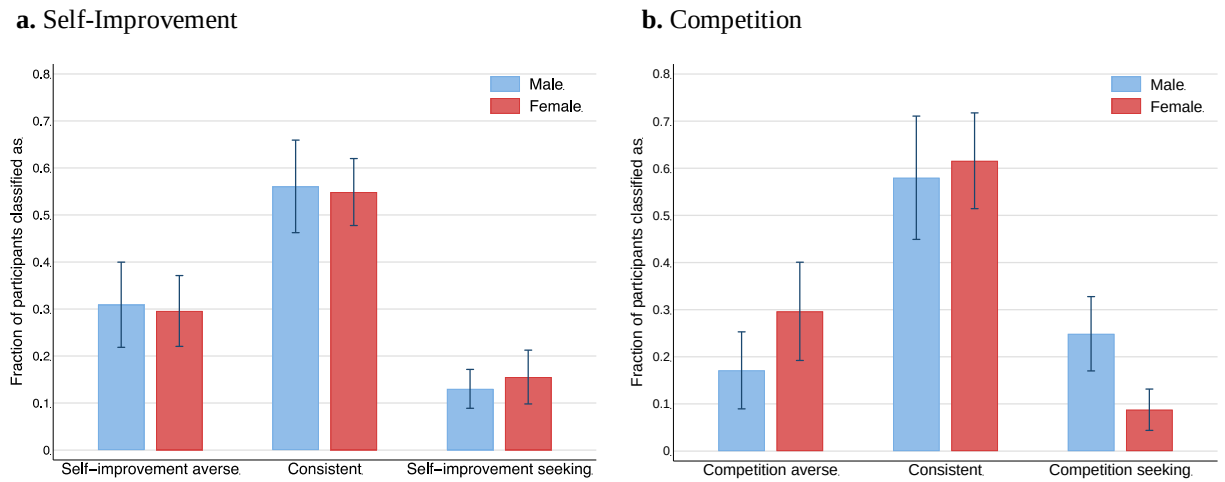

Figure 5: Distribution of Types, Conditional on Individual Characteristics

Notes: Estimates from Table 6 for Panel a, and Table 7 for Panel b, *Pooled* specification with all controls (Column 2). Range bars are 90-percent confidence intervals.

4. Discussion

We examine gender differences in preferences over compensation in a real-effort task, and find that females are less willing than males to accept contracts that pay upon self-improving or upon outperforming another individual. Higher female aversion to self-improvement is largely explained by their relatively higher risk aversion and lower self-confidence in their ability to improve. On the other hand, these and other factors are unable to account fully for why females dislike competition.

The findings provide insight into the design of effective ways to attract men and women into challenging environments. A mechanism that leverages people's willingness to self-improve may not appeal to men and women equally, given gender differences in risk tolerance and self-

confidence. Thus, these factors may need to be considered in the design phase for self-improvement pay to be a gender-neutral incentive. A possible way to address these differences seems to be to employ a task with relatively stable difficulty: our experiment suggests that self-improvement pay is unattractive to either gender, but especially to women, if it involves high earnings volatility. This is consistent with Flory et al. (2015) in the competition literature, who find that women are relatively less likely to enter a job in which payment is more uncertain.

In relation to other work on gender and self-improvement, where most comparable, our results are similar to those of Apicella et al. (2017). With a math task administered in the laboratory, they find a baseline gender gap in selection into self-improvement of 13 percentage points (p=0.176), which is accounted for almost entirely by the inclusion of controls, particularly risk aversion. With the math task in our experiment, we find a baseline gender gap of 12 percentage points (p=0.084), which shrinks substantially and becomes insignificant with the inclusion of controls, particularly risk aversion. Apicella et al. (2017) also conduct an online experiment with an M-Turk sample and a task of counting zeros from tables of zeros and ones. These features make this part of their work harder to compare to ours. Yet, to the extent that the task of counting zeros involves low variability in its difficulty, which seems possible,[8] our findings would predict a small gender gap in selection into self-improvement for this task, which is what Apicella et al. (2017) find: they find an insignificant baseline gender gap in favor of women of 5 percentage points (p=0.446).

Carpenter et al. (2017) also employ a math task in the laboratory. Their main finding is that women are 36 percentage points more likely to select into self-improvement than into competition (p=0.009). We find a similar result in our experiment: in the math task, women are 31 percentage points more likely to select into self-improvement than into competition (p=0.003, Fisher's exact test). However, whereas Carpenter et al. (2017) reasonably conclude from this that "women are more willing to compete against themselves than others" and draw implications to the labor market, our results suggest more caution. We document with the verbal task a case in which

[8] Round-to-round performance seems to be less variable in the counting task than the math task. In Apicella et al. (2017), average performance in the math task (*Self* and *Other* sessions) goes from 8.36 in Part 1 to 9.52 in Part 2; a 14 percent increase. In the counting task (*Self* and *Other* sessions), average performance goes from 2.96 in Part 1 to 3.26 in Part 2; a 10 percent increase.

females (and males) are actually less willing to self-improve than to compete, possibly due to the higher earnings variability in this task.

Thus, while our findings in the math task are largely consistent with the small literature on gender and self-improvement, we believe additional features in our experiment contribute to this literature and to potential policy discussion. The earnings volatility of the task may be an important factor in understanding what size of a gender gap in entry into self-improvement we might see, and how attractive self-improvement pay might be in general relative to competition. And our finding that stereotypes, even when they exist in the population, do not seem to affect sorting into self-improvement, is promising for policy, as it suggests that men and women alike may be attracted to self-improvement pay even in contexts and occupations generally perceived as incongruent with their gender.

Or result that females are less competitive than males in the verbal task is somewhat at odds with the literature on competitiveness. Wozniak et al. (2014) also find that females shy away from competition in a verbal task, but Shurchkov (2012), Dreber et al. (2014), Grosse et al. (2014), and Halladay (2017) all find females to be equally, if not more, competitive than males in non-stereotypically-male tasks. We can only speculate about the differences between our study and these latter set, but one possibility may be that general gender stereotypes (for instance, a notion that competitiveness is a male trait) may be more entrenched in our sample. According to the 2017 Global Gender Gap Report (Hausmann et al., 2017), Chile is ranked 63 out of 144 in the overall gender equality index, while Sweden is ranked 5, Germany 12, and the United States 49 (countries where the studies cited above have been conducted). An even within our sample, females are less competitive than males only among participants who endorse the stereotype of the task, while females are equally, if not more, competitive than males in our sample among participants who do not endorse the stereotype. It seems counterintuitive that lower female competitiveness in the verbal task emerges for the group who perceive that females have an advantage in this task. Perhaps agreement with this stereotype is capturing more general gender stereotypes held by the participants, which may driving the gap in competitiveness.[9] A potentially fruitful avenue to

[9] Recent work finds an important role for stereotypes in selection into tournaments (Hernandez-Arenaz, 2018) and performance under competition (Iriberri and Rey-Biel, 2017).

continue to investigate the interplay of stereotypes and competitiveness may be to measure (and perhaps seek to affect) beliefs about the femaleness of the action of engaging in competition, rather than about the femaleness of the task.

Ultimately, the productivity gain for men and women from implementing self-improvement pay in organizations would depend on how output responds to this contract, which is something that our design cannot examine. Doing so would require exogenous assignment (rather than self-selection) into the self-improvement contract, and the use of a task with demonstrated output elasticity.[10] Although some studies find that competition against others improves men's but not women's performance (Gneezy et al., 2003; Günther et al., 2010; Shurchkov, 2012), there is a basis to conjecture that self-improvement incentives can push productivity for both genders: a look at sixteen lab and field experiments by Bandiera et al. (2016) finds that men and women respond equally, and positively, to performance pay. If a similar response is seen for performance pay conditional on self-improvement, then, conditional on self-confidence and risk preferences, a self-improvement contract may represent a more gender-neutral way of encouraging challenge-taking and boosting productivity in organizations.

REFERENCES

Apicella, C., Demiral, E., and Mollerstrom, J., 2017. No Gender Difference in Willingness to Compete when Competing against Self. *American Economic Review: Papers and Proceedings*, 107.5: 136-140.

Araujo, F. A., Carbone, E., Conell-Price, L., Dunietz, M., Jaroszewicz, A., Landsman, R., Lamé, D., Vesterlund, L., Wand, S. W., and Wilson, A. J., 2016. The Slider Task: An Example of Restricted Inference on Incentive Effects. *Journal of the Economic Science Association*, 2.1: 1-12.

Bandiera, O., Fischer, G., Prat, A., and Ytsma, E., 2016. Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments. *CEPR Discussion Paper 11724*.

Bertrand, M., 2018. The Glass Ceiling: 2017 London School of Economics Coase Lecture. *Economica*, 85.338: 205-231.

[10] See Araujo et al. (2016) for a discussion on the importance of the choice of task when examining incentive effects on output.

Blau, F. D., and Kahn, L M., 2017. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55.3: 789-865.

Brandts, J., Groenert, V., and Rott, C., 2014. The Impact of Advice on Women's and Men's Selection into Competition. *Management Science*, 61.5: 1018-1035.

Buser, T., Niederle, M., and Oosterbeek, H., 2014. Gender, Competitiveness, and Career Choices. *Quarterly Journal of Economics*, 129.3: 1409-1447.

Buser, T., Peter, N., and Wolter, S., 2017. Gender, Competitiveness, and Study Choices in High School: Evidence from Switzerland. *American Economic Review: Papers and Proceedings*, 107.5: 125-130.

Carpenter, J., Frank, R., and Huet-Vaughn, E., 2017. Gender Differences in Interpersonal and Intrapersonal Competitive Behavior. IZA Discussion Paper No. 10626.

Cason, T. L., Masters, W. A., and Sheremeta, R. M., 2010. Entry into Winner-Take-All and Proportional-Prize Contests: An Experimental Study. *Journal of Public Economics*, 94: 604-611.

Charness, G., and Gneezy, U., 2012. Strong Evidence for Gender Differences in Risk Taking. *Journal of Economic Behavior and Organization*, 83: 50-58.

Charness, G., Gneezy, U., and Imas, A., 2013. Experimental Methods: Eliciting Risk Preferences. *Journal of Economic Behavior and Organization*, 87: 43-51.

Chen, D. L., Schonger, M., and Wickens, C., 2016. oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments. *Journal of Behavioral and Experimental Finance*, 9: 88-97.

Coffman, K. B., 2014. Evidence on Self-Stereotyping and the Contribution of Ideas. *Quarterly Journal of Economics*, 129.4: 1625-1660.

Croson, R., and Gneezy, U., 2009. Gender Differences in Preferences. *Journal of Economic Literature*, 47.2: 448-474.

Dariel, A., Kephart, C., Nikiforakis, N. and Zenker, C., 2017. Emirati Women Do Not Shy Away from Competition: Evidence from a Patriarchal Society in Transition. *Journal of the Economic Science Association*, 3.2: 121-236.

Dreber, A., von Essen, E., and Ranehill, E., 2014. Gender and Competition in Adolescence: Tasks Matter. *Experimental Economics*, 17.1: 154-172.

Flory, J. A., Leibbrandt, A., and List J. A., 2015. Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions. *Review of Economic Studies*, 82: 122-155.

Gneezy, U., Niederle, M., and Rustichini, A., 2003. Performance in Competitive Environments: Gender Differences. *Quarterly Journal of Economics*, 118.3: 1049-1074.

Greiner, B., 2015. Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE. *Journal of the Economic Science Association*, 1.1: 114-125.

Grosse, N., Riener, G., and Dertwinkel-Kalt, M., 2014. Explaining Gender Differences in Competitiveness: Testing a Theory on Gender-Task Stereotypes. *Working Paper*.

Günther, C., Ekinci, N. A., Schwieren, C., and Strobel, M., 2010. Women Can't Jump? An Experiment on Competitive Attitudes and Stereotype Threat. *Journal of Economic Behavior and Organization*, 75.3: 395-401.

Halladay, B., 2017. Perception Matters: The Role of Task Gender Stereotype on Confidence and Tournament Selection. *Working Paper*.

Harvard Business Review, 2016. *HBR Guide to Delivering Effective Feedback*, Harvard Business Press.

Hausmann, R., Tyson, Laura, and The World Economic Forum, 2017. *The Global Gender Gap Report*. http://www3.weforum.org/docs/WEF_GGGR_2017.pdf

Healy, A., and Pate, J., 2011. Can Teams Help to Close the Gender Competition Gap? *Economic Journal*, 121: 1192-1204.

Hernandez-Arenaz, I., 2018. Stereotypes and Tournament Self-Selection: A Theoretical and Experimental Approach. Working Paper.

Iriberri, N., and Rey-Biel, P., 2017. Stereotypes Are Only a Threat When Beliefs are Reinforced: On the Sensitivity of Gender Differences in Performance Under Competition to Information Provision. *Journal of Economic Behavior and Organization*, 135: 99-111.

Kamas, L., and Preston, A., 2012. The Importance of Being Confident; Gender, Career Choice, and Willingness to Compete. *Journal of Economic Behavior and Organization*, 18.1: 82-97.

Kamas, L., and Preston, A., forthcoming. Competing with Confidence: The Ticket to Labor Market Success for College-Educated Women. *Journal of Economic Behavior and Organization*.

Moore, D. A. and Healy, P. J., 2008. The Trouble with Overconfidence. *Psychological Review*, 115.2: 502-517.

Niederle, M., 2016. Gender. In: Kagel, J., and Roth, A. E., (Eds.). *Handbook of Experimental Economics*, Vol. II. Princeton University Press.

Niederle, M., Segal, C., and Vesterlund, L., 2013. How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness. *Management Science*, 59.1: 1-16.

Niederle, M., and Vesterlund, L., 2007. Do Women Shy Away from Competition? Do Men Compete Too Much? *Quarterly Journal of Economics*, 122.3: 1067-1101.

Niederle, M., and Vesterlund, L., 2011. Gender and Competition. *Annual Review of Economics*, 3: 601-630.

Petrie, R., and Segal, C., 2015. Gender Differences in Competitiveness: The Role of Prizes. *Working Paper*.

Reuben, E., Sapienza, P., and Zingales, L., 2015. Taste for Competition and the Gender Gap Among Young Business Professionals. NBER *Working Paper 21695.*

Saccardo, S., Pietraz, A., and Gneezy, U., forthcoming. On the Size of the Gender Difference in Competitiveness. *Management Science*.

Servicio de Información de Educación Superior (SIES), 2017. *Brechas de Género en Educación en Chile 2016*, Ministerio de Educación de Chile. Data available at http://www.mifuturo.cl/index.php/estudios/estudios-recientes

Shurchkov, O., 2012. Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints. *Journal of the European Economic Association*, 10.5: 1189-1213.

Sutter, M., and Rützler, D., 2014. Gender Differences in Competition Emerge Early in Life and Persist. *Management Science*, 61.10: 2339-2354.

Wozniak, D., Harbaugh, W. T., and Mayr, U., 2014. The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices. *Journal of Labor Economics*, 32.1: 161-198.