Overview

Compensation packages based on performance pay, such as bonuses, commissions, and piece-rate payments, have risen in popularity relative to hourly/salaried pay, especially among workers in the highest tiers of occupations (Cuñat & Guadalupe, 2005; Hall & Liebman, 1998; Lemieux, MacLeod, & Parent, 2009; Murphy, 1999). There is evidence that the increasing use of performance pay lends itself to wage inequality. Lemieux et al. (2009) showed that an increased dependence on performance pay during the late 1970's and early 1990's accounted for 21% of the observed growth in variance of male wages. Bonuses and commissions, arguably the most competitive compensation schemes, may be especially important in driving the large disparity between the highest and lowest percentile earners within organizations (Bell & Van Reenen, 2010, 2014; Bénabou & Tirole, 2016). Importantly, performance pay may contribute to the gender wage gap too. Using data from the National Longitudinal Surveys of Youth, McGee, McGee, & Pan (2015) show that women are less likely to be employed in occupations that receive bonuses, and simultaneously are more likely to receive piece-rate pay – the least competitive of all forms of performance pay, where workers are paid based on their absolute output.

Since competition is relevant to labor market outcomes, researchers began to focus on how a person's gender affects their response to competition as a means of understanding persistent gender gaps in labor market outcomes (for review, see Niederle & Vesterlund, 2011). Seminal work on gender differences in competitiveness operationalized competitiveness as the choice of a tournament payment scheme that reaps potentially higher earnings but requires outperforming an opponent over a piece-rate scheme (Niederle & Vesterlund, 2007). This work found that women are less competitive than men, on average, even if they would have earned more by competing (Niederle & Vesterlund, 2007). Additionally, this laboratory measure of competitiveness predicts labor market outcomes, such as education choices (Buser, Niederle, & Oosterbeek, 2014; Zhang, 2012), entrepreneurial decisions (e.g., investment, employment; Berge, Bjorvatn, Garcia Pires, & Tungodden, 2015), and earnings (Reuben, Sapienza, & Zingales, 2015). Thus, competitive preferences may contribute to gender differences in labor market outcomes (Blau & Kahn, 2017).

Follow-up research with nearly identical procedures has replicated the effect of gender on the choice to opt into tournaments (see Niederle & Vesterlund, 2011 for review). Notably, this effect has been replicated in diverse populations (e.g., across age groups and cultures) (Andersen, Ertac, Gneezy, List, & Maximiano, 2013; Apicella & Dreber, 2015; Buser et al., 2014; Buser, Peter, & Wolter, 2017; Dreber, Essen, & Ranehill, 2014; Mayr, Wozniak, Davidson, Kuhns, & Harbaugh, 2012; Sutter, Glätzle-Rützler, Balafoutas, & Czermak, 2016; Sutter & Rutzler, 2010) and with a diverse set of tasks (Apicella & Dreber, 2015; Bjorvatn, Falch, & Hernæs, 2016; Frick, 2011; Saccardo, Pietrasz, & Gneezy, 2018; Samek, 2019; Sutter & Glätzle-Rützler, 2015). However, there is evidence that the task used during competition affects the size of the gender gap. For instance, some research suggests that when the task is female-typed or gender-neutral, the gender gap in willingness to compete may be reduced or eliminated (Apicella & Dreber, 2015; Boschini, Dreber, Essen, Muren, & Ranehill, 2019, 2014; Dreber, Essen, & Ranehill, 2011; Dreber et al., 2014; Grosse & Riener, 2010; Günther, Ekinci, Schwieren, & Strobel, 2010; Iriberri & Rey-Biel, 2017; Shurchkov, 2012). Drawing from the psychology literature on stereotype threat (Spencer, Logel, & Davies, 2016; Spencer, Steele, & Quinn, 1999; Steele, 1997), negative stereotypes about women's ability to perform male-typed tasks (e.g., math, mental rotation) may produce anxiety and undermine performance. As a result, women may decide not to engage in a competition because they either believe the stereotype or because the stereotype provokes enough anxiety to reduce performance (Grosse & Riener, 2010; Günther et al., 2010; Iriberri & Rey-Biel, 2017; Shurchkov, 2012).

While competitions are generally motivating and designed to improve performance through increased effort (Connelly, Tihanyi, Crook, & Gangloff, 2014; Miller, Petrie, & Segal, 2019; Murayama & Elliot, 2012), some research suggests that men perform better under competitive payment schemes relative to non-competitive payment schemes, while women's performance does not respond to competitions (Gneezy, Niederle, & Rustichini, 2003; Gneezy & Rustichini, 2004; Günther et al., 2010; Samak, 2013). Gneezy et al. (2003) show that there is no gender difference in performance when participants are solving mazes following a piece-rate payment scheme, but a significant gender difference in performance arises under a tournament payment scheme, with males performing better. Günther et al. (2010) replicate the effect of competition on gender differences in performance for a male-typed task, but find no gender differences in performance during competition for female-typed or gender-neutral tasks. Relatedly, during repeated competition, women tend to perform worse in subsequent performance rounds after losing, even if the monetary prize they lost was relatively meager, while men only perform worse in subsequent rounds if they lost the chance to win a large monetary prize (Gill & Prowse, 2014). Other research suggests women stop competing altogether after losing if given the choice. Buser & Yuan (2019), who examine the effects of losing while competing in the Dutch Math Olympiad on the choice to compete in subsequent years, show that men are just as likely to compete even if they lost the previous year, while women are less likely to compete again if they lost before. This body of literature suggests that competitions may differentially impact women and men.

To date, most of the research on gender differences in competitions has focused on either i) explaining the sources of the gender difference (e.g., Veldhuizen, 2017) or ii) designing interventions to encourage women to compete more (Alan & Ertac, 2018; Balafoutas & Sutter, 2012; Brandts, Groenert, & Rott, 2015; Brandts et al., 2015; Cassar, Wordofa, & Zhang, 2016; Healy & Pate, 2011; Niederle, Segal, & Vesterlund, 2013; Sutter et al., 2016). Less consideration has been paid to how competitions differentially, and negatively, impact women. However, as the research on gender differences in performance during competition suggests (Buser & Yuan, 2019; Gill & Prowse, 2014; Gneezy et al., 2003; Gneezy & Rustichini, 2004; Günther et al., 2010; Miller et al., 2019; Samak, 2013), it is also important to consider potential downstream consequences of women's entry into competitions.

The present proposal builds on prior research by examining how competitions affect gender differences in the amount of time spent preparing for competitions. We hypothesize that women will spend more time preparing than men, especially before competitions, in part because they are, on average, less risk-seeking (Bertrand, 2010; Croson & Gneezy, 2009; Dohmen et al., 2011; Eckel & Grossman, 2008) and confident (Barber & Odean, 2001; Bertrand et al., 2010; Croson & Gneezy, 2009; Lundeberg, Fox, & Puncochaf, 1994; Mobius, Niederle, Niehaus, & Rosenblat, 2011) than men. Indeed, both confidence and risk attitude have been implicated in driving the gender gap in willingness to compete (Niederle & Vesterlund, 2011; Veldhuizen, 2017).

The extent to which confidence and risk attitude account for the gender gap in willingness to compete is debated. The seminal research in this literature suggests that confidence and risk attitude do not completely explain the gender gap in the choice to compete, since there remains a residual gap in the choice to compete after controlling for these factors (Niederle & Vesterlund, 2007). The unexplained component of the original gender effect was taken as evidence of a distinct "competitiveness" trait, separate from risk attitude and confidence (Niederle & Vesterlund, 2007, 2011). However, recent work correcting for measurement error (Gillen, Snowberg, & Yariv, 2019) and using experimental techniques to isolate the effects of the competitiveness trait (Veldhuizen, 2017) suggests that risk attitude and confidence may fully explain the gender gap in the choice to compete. Regardless of whether competitiveness is a stand-alone trait, it is clear that confidence and risk attitude can generate differences in how men and women react to competitions, possibly including the decision to prepare before a competition.

Confidence is conceptualized as the accuracy of one's perceived performance or ability on a task (Beyer & Bowden, 1997). Since competitions, by definition, compare the performance among two or more individuals, they naturally lead to self-evaluation and comparative judgments of self with others - processes that are intimately linked to confidence. To the extent that confidence influences how much individuals think they need to prepare in order to win, we may expect to see women preparing more than men, particularly in competitive contexts, which naturally invoke self-other assessments. Thus, less confident individuals may prepare more. Moreover, they may prepare more in order to reduce the negative feelings caused by low confidence independent of any ambitions to win. Indeed, mastery is an important driver of confidence (for review, see Gist & Mitchell, 1992; Usher & Pajares, 2008) and there is no theoretical or empirical reason to suspect that women would be less concerned with mastery than men. In fact, research suggests that women are just as likely as men to compete when competing against their own past performance, suggesting an equal desire for self-improvement (Apicella et al., 2017b).

There is ample research to suggest that women are less (over)confident on average than men across a number of domains (Bertrand, 2010; Beyer, 1990; Beyer & Bowden, 1997; Croson & Gneezy, 2009; Lundeberg et al., 1994; Mobius et al., 2011; Niederle & Vesterlund, 2007, 2011). Within the literature on the gender gap in competitiveness, confidence is operationalized as the belief about one's relative performance during a competition, where individuals who have inaccurately high ratings of their performance are deemed overconfident. If an individual does not feel as though their performance is higher than individuals they are competing against, they are unlikely to make the decision to compete for fear of missing the opportunity to earn money, even if they would otherwise outperform their opponent. Although both men and women tend to be overconfident, men are far more likely to fall into the trap of overconfidence, which leads them to compete more often than they should, given their actual ability (Niederle & Vesterlund, 2007).

Another variable that has been identified as a possible explanation for gender differences in competitiveness is risk attitude, typically construed as the preference for a certain gain over a gamble, even if the gamble has an equal or greater monetary expectation (Kahneman & Tversky, 1982). For instance, a risk averse person would prefer a sure gain of $80 over a gamble where they have an 85 percent chance of winning $100 and a 15 percent chance of winning $0, even though the monetary expectation in the latter case is higher (e.g., average of $85 in earnings relative to an average of $80) (Kahneman & Tversky, 1982). Payment based on the outcomes of a competition are inherently riskier than non-competitive payment schemes (e.g., guaranteed payment per unit of output) because in most cases, there is uncertainty surrounding one's relative performance (Niederle & Vesterlund, 2011). Several studies across diverse settings have documented a gender difference in risk attitudes, where women tend to be more risk-averse than men on average (Apicella et al., 2017a; Bertrand, 2010; Croson & Gneezy, 2009). Because competitions are riskier and women tend to be more risk-averse, women may prepare more than men before competing to reduce some of the uncertainty of performing during competition.

The current proposal examines gender differences in preparation in competitive contexts. Because women tend to be more risk-averse (Bertrand, 2010; Croson & Gneezy, 2009; Dohmen et al., 2011; Eckel & Grossman, 2008), less confident (Barber & Odean, 2001; Bertrand et al., 2010; Croson & Gneezy, 2009; Lundeberg et al., 1994; Mobius et al., 2011) and prefer to opt out of competitions (Niederle & Vesterlund, 2011), they may engage in more coping strategies, such as preparation, before entering competitions. While there is no literature examining how competitions might impact gender differences in preparation, there is a small literature suggesting that women are more likely than men to value dedication and mastery (Kenney-Benson, Pomerantz, Ryan, & Patrick, 2006; Leslie, Cimpian, Meyer, & Freeland, 2015), emphasize the importance of hard work (Hirt & Mccrea, 2009; Mccrea et al., 2008a, 2008b), and spend more time preparing than men in general (Kimble & Hirt, 2005; Lucas & Lovaglia, 2005). For instance, in a study examining school-aged children's approach to learning math, researchers found that girls, compared to boys, reported being more motivated to "master" their schoolwork and engage in more effortful learning strategies (Kenney-Benson et al., 2006).

Additionally, in my own research, I discovered a sizable gender difference in effort, where women were more likely than men to choose to prepare before completing a multiplication task (Richards et al., in prep).[1] The goal of this work was to explore how preparation might influence women's willingness to compete. The first preliminary study in this line of work manipulated participants' (N = 1056) knowledge of whether they would have unlimited time to prepare before they made their decision to compete. We expected that participants who knew they had time to prepare would be more inclined to compete compared to participants who were not aware of the opportunity to prepare. While we did not find that knowledge of preparation affected participants' decision to compete, there was a sizable gender difference in the choice to prepare. Controlling for the decision to compete, women were 75% more likely to choose to prepare compared to men when offered the opportunity, $b = 0.56$, 95% CI [0.31, 0.82], $z = 4.37$, $p < .001$ (see Figure 1). Additionally, at the end of the experiment, participants were incentivized to correctly predict which gender they believed would be more likely to prepare in the study. Both men (74%) and women (84%) correctly believed that women would spend more time preparing for the task, $\chi^2(1, n = 1056) = 447.11$, $p < .001$. We found similar results when asking participants which gender prepares more in general - 82% of men and 87% of women believed that women prepare more in general, $\chi^2(1, n = 1056) = 447.11, p < .001$.

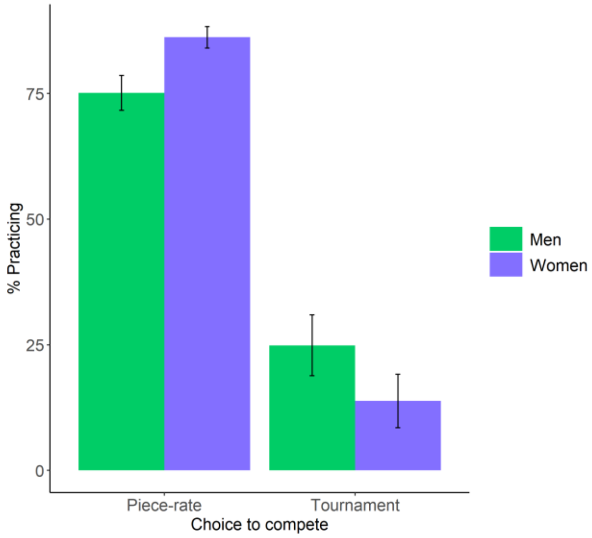

Figure 1. Proportion of participants who chose to prepare based on participant gender and choice to compete from first study. Error bars represent standard error.

In a follow-up study designed to examine the role of forced preparation on the decision to compete, we recruited 1076 participants from Amazon Mechanical Turk (MTurk), who were assigned to either a condition where they were required to complete several rounds of practice for the upcoming paid multiplication task or several rounds of a filler task before choosing their payment scheme. Again, our manipulation had no effect on the choice to compete in men or women. However, after completing their respective rounds of practice and choosing their payment scheme, participants in both conditions had the option to spend (extra) time preparing for the multiplication task. Here, we replicated the effect of gender on preparation, where 40% of women and 35% of men chose to complete the optional preparation while controlling for the choice to compete, $b = 0.32$, 95% CI [0.06, 0.57], $z = 2.44$, $p = .014$ (see Figure 2). In fact, even in the condition where participants were forced to prepare for 12 rounds, lasting on average two minutes, there was a significant effect of gender on the choice to prepare, $b = 0.42$, 95% CI [0.05, 0.79], $z = 2.20$, $p = .028$. Here too, we find that participants correctly predicted that women prepared more for the task, $\chi^2(1, n = 1076) = 394.16$, $p < .001$. They also reported that women are more likely to prepare in general, $\chi^2(1, n = 1076) = 511.06$, $p < .001$. Overall, our previous work provides compelling evidence that women are more likely to prepare than men, even after being forced to do so.

[1] All prior studies were pre-registered on Open Science Framework

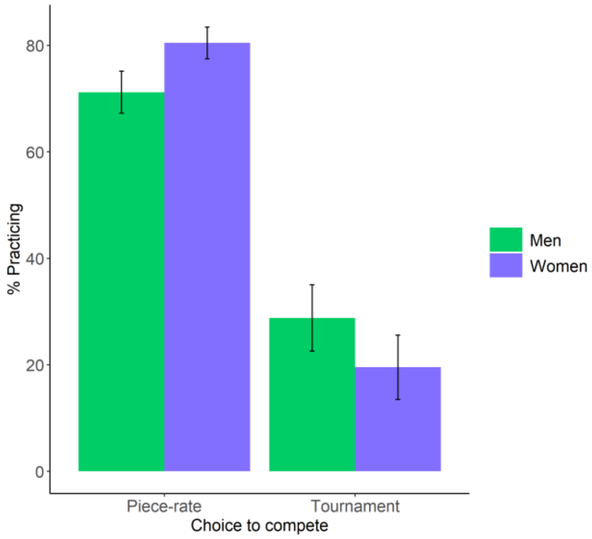

Figure 2. Proportion of participants who chose to prepare based on participant gender and choice to compete from second study. Error bars represent standard error.

While this research did not find that preparation affects decisions to compete, we did uncover a sizable gender difference in decisions to prepare. In light of this discovery, we have flipped our research question and now ask whether and how competitions affect men and women's decisions to prepare. To be clear, there was no interaction between gender and choice to compete on the choice to prepare in either of the previous studies. That is, women prepared more than men regardless of which payment scheme they had chosen. However, it is not possible to draw conclusions from this because i) we did not manipulate the payment scheme, so there could have been selection effects on one's choice to prepare across payment schemes, such that those who were more likely to choose to compete may have been less likely to prepare, and ii) there was little power to detect any possible interaction effects. For instance, in the first study, only 11% of women chose to compete, and in the second, 14% of women made this decision. Through the proposed experiments, we intend to address these limitations by directly manipulating participants' payment scheme and recruiting a large sample to provide power to detect small effects.

Here, we propose two studies. In Study 1, we will test whether competition exacerbates previously established gender differences in preparation by manipulating participants' assigned payment scheme (i.e., competitive or non-competitive). We will also explore whether the effect of confidence and risk attitude on preparation differ based on participant gender and condition. Based on research suggesting that task type affects women's decision to compete (Apicella & Dreber, 2015; Boschini et al., 2019, 2014; Dreber et al., 2011, 2014; Grosse & Riener, 2010; Günther et al., 2010; Iriberri & Rey-Biel, 2017; Shurchkov, 2012), Study 2 will explore whether beliefs about how each gender will perform on a matching task affects gender differences in preparation before competitions. Here, we expect women will spend significantly more time preparing for the matching task when participants believe, based on how previous research is described, men will outperform women, but the gender difference in preparation will be reduced when participants believe women will outperform men.

Study 1: Does competition elicit gender differences in effort?

Participants

Participants will be recruited to complete a study on "decision-making and performance" through MTurk, with a guaranteed payment and the opportunity to earn bonuses depending on their performance and the performance of others. Recruiting participants on this platform allows for efficient data collection while meeting acceptable psychometric standards, such as high test-retest and alpha reliability (Buhrmester, Kwang, & Gosling, 2011; Rand, 2012). Since we anticipate completing the required parts of the study will take no more than 10 minutes on average, we will pay participants $2.50 (i.e., double the federal and Pennsylvania minimum wage), with the opportunity for bonuses, outlined below. Participants will only be included if they indicate that they are 18 years or older, are American citizens, and identify as female or male while answering initial demographic questions.

Given the difficulty of powering interaction effects (see Simonsohn, 2014; Giner-Sorolla, 2018), we conducted a power analysis to determine an adequate sample size for the main hypothesized interaction effect in the primary analysis (simulations modeled after code from Hughes, 2017). We ran 5000 simulations while varying the sample size (*N* = 3000, 3250, 3500) and the effect size for the interaction effect (*b* = .2, .3, .4). Based on these simulated estimates, we will recruit 3250 participants to achieve at least 80% power for a relatively small effect (*b* = .2) (see Figure 3).

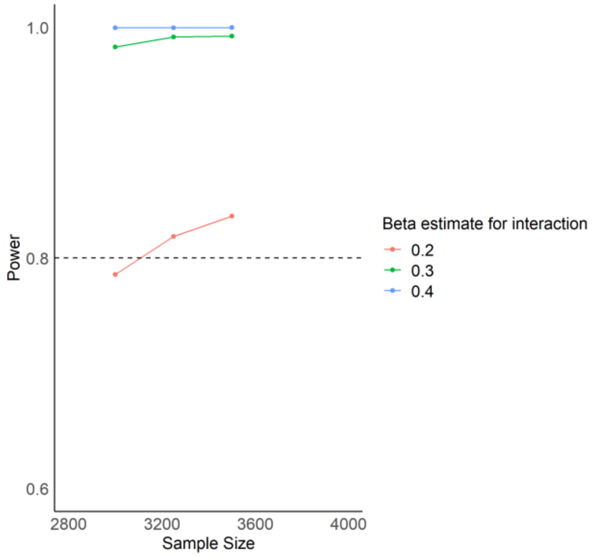

Figure 3. Plot of simulation output used to determine necessary sample size for at least 80% power in both proposed studies.

Manipulation

Participants will be randomly assigned to follow either a competitive or noncompetitive payment scheme for one round (2 minutes) of multiplication problems. The payment scheme will be manipulated between subjects, where participants in the competition (tournament) condition will be paid 4 cents per problem on the task, but only if they beat another randomly assigned MTurker, while participants assigned to the noncompetitive (piece-rate) payment scheme will be paid 2 cents per problem. Although a within-subjects design would provide more power in detecting the hypothesized interaction effect, we opted to use a between-subjects design to avoid carryover effects. If we followed a within-subjects design, we would only be able to confidently interpret the results for whichever condition were presented first because there would be several carryover effects that could affect the decision to prepare (e.g., fatigue and/or learning effects reducing participants' desire to prepare, demand effects for preparation if participants believe they are expected to prepare more in one condition compared to the other).

Dependent variable

After participants in each condition are told which payment scheme they will be following, they will have the option to prepare for the task by completing unlimited practice problems, which they will be told could improve their performance on the subsequent task. To measure their desire to prepare for the task, we will first ask participants whether they would like to spend any time practicing multiplication problems. We chose a multiplication task because we expect participants will improve with practice. Indeed, research suggests that rehearsing and recalling associative memories can speed up retrieval of those memories (Rundus, 1971). Moreover, we have already established a robust gender difference in both the choice to prepare and compete using this task (Richards et al., in prep). For participants who agree to practice, they will be be able to practice for as long as they want, with the option to pause in case of any unexpected interruptions, such as children coming into the room. Also, participants will have the option to exit the preparation and move onto the task at any point via an "Exit" button in the bottom right corner of the survey screen. The dependent variable will be quantified as the total number of seconds of preparation, excluding the amount of time participants paused during the practice.

Task performance

After practicing, participants in each condition will complete the paid multiplication task. Participants' scores on the task will be quantified as the number of questions correct within the two-minute time frame allotted, without any penalties for incorrect responses. Afterwards, participants will be informed of the number of questions they answered correctly. We do not include any information about their relative performance since we ask them to guess their relative performance in the confidence measure. Thus, participants following the tournament payment scheme will not be told whether they won, since this serves as an indicator of relative performance.

Post-manipulation measures

After completing the task, participants will complete a series of measures to be used for exploratory analyses. All questions will be counterbalanced. A confidence measure will incentivize participants to guess their relative performance compared to all other participants that completed the task by indicating the decile of their score relative to other participants. If correct, participants will earn $.25. We use a measure of relative performance, rather than a measure of absolute performance (e.g., asking participants to guess their score on the task) because perceptions of relative performance will likely be predictive of the choice to practice, especially when an individual is required to compete. The confidence measure draws from previous research (Niederle & Vesterlund, 2007), but instead of asking participants to indicate whether they won against a randomly selected opponent, we ask them to guess their relative decile to provide us with more information about their relative confidence. Given the difficulty of guessing one's exact percentile without any information about other participants, deciles are used rather than percentiles to make earning the bonus seem more achievable. Also, the item will be phrased so participants do not need to understand the word "decile," but will be asked "If my performance is compared to that of all participants that completed the task, I think my score was…" with the options for responses ranging from "Better than all other participants" to "Better than none of the other participants" with 10% increments in between (e.g., "Better than 50% of participants"). Since task-specific confidence measures tend to be better predictors of behavior than general measures of confidence (see Oney & Oksuzoglu-Guven, 2015 for review), the confidence measure assesses participants' beliefs within the context of the task used. We will also measure risk attitude by asking participants to indicate on a 0-10 scale "How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" (Dohmen et al., 2011). There is evidence that risky behavior (i.e., lottery choices) is strongly associated with the risk measure included in the current proposal (Dohmen et al., 2011). Additionally, risk attitude tends to be explained by one underlying trait, with a relatively smaller amount of variation in risk attitude explained by context (e.g., risk attitude during career, health, or financial decisions). Thus, across contexts, risk attitude is likely to be stable and predictive of behavior (Dohmen et al., 2011). These measures are included after completing the task largely because the confidence measure requires participants to state their perceived relative performance on the task.

Concerns: calculator use and attrition

There may be concern that participants will use a calculator to answer the multiplication questions, which could affect the interpretation of the results if there is a gender difference in calculator use and/or calculator use is related to the choice to practice. Our previous work suggests participants are unlikely to use calculators to complete the task and more importantly, there are no gender differences in the choice to use a calculator. In our first study we ran using a multiplication task, participants who completed the task were asked i) whether they thought using a calculator would help them answer the multiplication problems more quickly and ii) whether they used a calculator to complete the multiplication task (they were told their response would not affect their payment). Based on their responses, it is unlikely that participants will use calculators in the first place, since 86% of participants indicated that they thought using a calculator to answer the multiplication questions would slow them down and 93% of participants said they did not use a calculator. Importantly, there were no gender differences in perceptions of how calculators would affect performance, $\chi^2(1, n = 1056) = 0.42$, $p = .519$. Additionally, we did not find evidence of gender differences in actual calculator use, $\chi^2(1, n = 1056) = 1.70$, $p = .193$. Since we are recruiting participants through the same platform using the same task in the current study, we expect these findings will generalize to the current study, and thus, do not have evidence that gender differences in calculator use will be a confound when interpreting our results.

Attrition can threaten an experiment's internal validity (Zhou & Fishbach, 2016). Fortunately, our previous work suggests that condition-dependent attrition is unlikely (Richards et al., in prep). In our prior study where participants completed one round of a similar multiplication task under each type of payment scheme, only a small proportion of participants (6%) dropped out during the study. And, of the participants who dropped out, all did so at the end of the study, after completing both of the main tasks. Nevertheless, we will still take several steps to counteract the possibility of condition-dependent attrition, which has the potential to lead to misleading conclusions (Zhou & Fishbach, 2016), especially if women and men drop out of the study at different rates based on condition. First, we will employ three costless strategies (i.e., personalization, forewarning of study content, and an appeal to participants' conscience) suggested by Reips (2000) and shown in Zhou & Fishbach (2016) to be effective in reducing dropout rates by at least half. When participants enter the study, they will read a message that serves as both a forewarning and an appeal to their conscious (modified from Zhou & Fishbach, 2016): "This is an anonymous survey consisting of multiple questions. If a sizable number of people quit a survey partway, the data quality of that survey would be compromised. However, our research depends on good quality data, so we ask that you are willing to participate in the survey for its entirety." Then, participants will enter their MTurk ID as a means of establishing personalization. Notably, Zhou & Fishbach (2016) acknowledge that this is not a foolproof solution, since screening participants in advance in this way may reduce external validity. In this case, we want to have the capacity to establish the anticipated effect in the first place, so we are prioritizing internal validity.

In addition to these preventive measures, we will collect information about the rates of attrition during each study. Turkprime provides a metric for the overall rate of attrition, while Qualtrics offers the option to view partial responses from dropouts. For participants who drop out during or after learning about the manipulation, we will create an indicator variable for survey completion based on partial responses from Qualtrics, which will be coded as 1 if participants finish the study and 0 otherwise. This indicator will then be submitted as the dependent variable to a logistic regression with $\beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Condition} + \beta_3 \text{Gender} \times \text{Condition}$ as predictors. If we find a significant interaction effect between gender and condition, this would suggest that we should interpret our results with caution because internal validity may be threatened, which will be explicitly stated in any reports on the studies, along with overall attrition rates and condition-dependent attrition rates (Zhou & Fishbach, 2016).

Primary hypotheses and analyses

We will be using two-tailed tests during all hypothesis testing ($p < .05$) and all analyses will be conducted using *R*. To control the false-discovery rate during exploratory analyses, we will apply the Benjamini-Hochberg correction to all exploratory analyses. All analyses will be pre-registered on Open Science Framework.

We expect that women will choose to prepare more than men, especially before a competition. We will test the interaction between gender and condition (competitive or noncompetitive pay) using a linear regression with amount of time a participant chose to prepare (log-transformed) as the dependent variable. We will include the number of practice problems completed as a control for differences in participants' ability, along with a control for the total time participants spent pausing during the practice. Thus, the following linear regression will be run:

$$\text{Log(Time Preparing)} = \beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Condition} + \beta_3 \text{Number of problems completed} + \beta_4 \text{Total pause time} + \beta_5 \text{Gender} \times \text{Condition},$$

where the piece-rate payment scheme and men will be coded as the reference groups for Condition and Gender, respectively. A positive beta coefficient for the interaction term (β_5) would support our hypothesis, indicating that the effect of gender on time spent preparing is greater in the tournament condition. Additionally, we expect positive beta coefficients for the main effects of gender and condition, suggesting the women and participants following the competitive pay scheme spent more time preparing. Finally, we will check that participants from different demographic groups were successfully randomized equally to each condition by running four separate logistic regressions with age, race/ethnicity, education, and income predicting condition (e.g., $\text{Condition} = \beta_0 + \beta_1\text{Age}$). Since both proposed experiments include large sample sizes (N = 3250), it is unlikely demographic variables will become exceptionally imbalanced across conditions to the extent that they will explain our observed effects (Bowers, 2011). However, we will control for any variables that were not successfully randomized if we find significant differences in demographics across groups.

Exploratory analyses

We will test whether the role of confidence or risk attitude on time spent preparing differs based on participant gender and condition. To this end, we will run separate multiple regression analyses to test whether confidence or risk attitude interact with gender and condition to affect time spent preparing (after log transformation) as the dependent variable. Therefore, the model will be structured as follows, where X_1 is either participants' confidence or risk attitude:

$$\text{Log(Time Preparing)} = \beta_0 + \beta_1\text{Gender} + \beta_2\text{Condition} + \beta_3 X_1 + \beta_4\text{Gender} \times \text{Condition} + \beta_5\text{Condition} \times X_1 + \beta_6\text{Gender} \times X_1 + \beta_7\text{Gender} \times X_1 \times \text{Condition}.$$

The reference groups will be the piece-rate payment scheme and men for condition and gender, respectively. Given the previous literature on gender gaps in confidence/risk attitude and competitiveness, we would expect a three-way interaction between gender, condition, and confidence/risk attitude on preparation, where women's confidence/risk plays a larger role in time spent preparing for women following the competitive payment scheme, relative to men following either the competitive or piece-rate payment scheme and women following the piece-rate payment scheme. A three-way interaction may be underpowered with the current sample size, so the analysis of the three-way interaction will serve as the foundation for future work using the effect sizes found.

Study 2: Do task stereotypes elicit gender differences in effort during competition?

Participants

Study 2 will explore the boundary conditions of gender differences in the choice to prepare during competition by manipulating stereotypes about gender differences in performance on the task. Like Study 1, participants will be recruited on Amazon's Mechanical Turk to complete a study on "decision-making and performance." To attain power for the hypothesized interaction between condition and gender on time spent preparing, we will recruit 3,250 participants (see Figure 3). Participants will only be included if they indicate that they are 18 years or older, are American citizens, and identify as female or male while answering initial demographic questions. We estimate the study will take an average 10 minutes to complete, so we will again pay participants $2.50 to complete the study. There will also be the opportunity to earn a bonus payment.

Manipulation

All participants will be entered into a competition, where they will earn 4 cents per problem if they outperform a randomly assigned partner. To reserve power for the main interaction effect of interest between gender and task stereotype on time spent preparing, we do not manipulate payment scheme. Participants are told that they will be paid to complete a one-minute "matching task," where they are first presented a legend containing numbers and their corresponding letters. Using this legend, participants must, as quickly as possible, enter the letters on their keyboard that correspond to the sequences of two-digit numbers presented to them. Upon learning about the task, participants will be told about the results of a study examining gender differences in manual skill (i.e., Majeres, 1983). However, in order to avoid deception, we will selectively share the study's results with the participants. Half of the participants will be told about one set of findings, where the researcher found that males excel, and the other half will be told about a separate set of findings where females excel. Participants in both groups will be told that this could result in males (or females) being better at the matching task. Specifically, participants in each condition will be told the following:

Condition 1: "A 1983 study published in the scientific journal, *Intelligence*, found that men have a motor-speed advantage compared to women (Majeres, 1983). Thus, men may be better at the matching task than women."

Condition 2: "A 1983 study published in the scientific journal, *Intelligence*, found that women have an advantage when it comes to programming sequences of manual movements (Majeres, 1983). Thus, women may be better at the matching task than men."

We validated the matching task in a pilot study with 337 MTurkers. There was no significant gender difference in performance on the task, $\Delta M = 3.92$, 95% CI $[-0.45, 8.28]$, $t(307.07) = 1.77$, $p = .078$ (see Figure 4), and we did not find a gender difference in competitiveness, $\chi^2(1, n = 337) = 0.36$, $p = .551$. These results provide tentative support for the notion that our manipulation will be believable in both conditions. Also, should participants look for the study on the internet, they will learn that we are not deceiving them. In summary, we developed this novel task to increase the likelihood that participants will believe our manipulation. If we used the multiplication task or another task that participants were familiar with, it is possible participants may not believe there were gender differences in performance, or have pre-conceived ideas about which gender would perform better, based on any previous experience with the task they may have had.

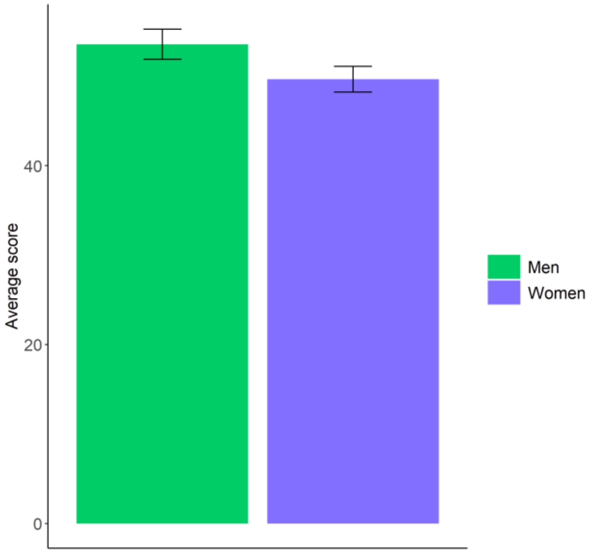

Figure 4. Average performance by gender from the pilot study. Error bars represent standard error.

The pilot data also showed that 80% of participants believed their score on the matching task would have improved with practice if they had been given the chance, $\chi^2(1, n = 337) = 112.81$, $p < .001$. Therefore, participants should be motivated to practice before the matching task compared to other tasks where one's score is unlikely to improve with practice. Finally, because pilot data suggests participants complete the problems in the matching task twice as fast as problems in the multiplication task, we limit the task time to one-minute to reduce total study costs.

Dependent variable

Before entering the paid matching task, participants across both conditions will have the option to prepare for as much time as they want, which will serve as our main dependent variable. Like Study 1, participants will be able to exit the preparation at any point or pause if they are interrupted. The dependent variable will only include the time spent practicing (i.e., excluding paused parts of the preparation).

Task performance

After practicing, participants in each condition will complete the paid matching task. Participants' scores on the task will be quantified as the number of questions correct within the one-minute time frame allotted, without any penalties for incorrect responses. Afterwards, participants will be informed of the number of questions they answered correctly. We do not include any information about their relative performance since we ask them to guess their relative performance in the confidence measure.

Post-manipulation measures

After completing the paid task, participants will complete the measures of risk attitude and confidence from Study 1, along with a manipulation check, where participants are asked to identify whether the previous study we described suggested that, on average: a) men were expected to perform better on the task, b) women were expected to perform better on the task, or c) there were no expected gender differences in performance on the task. The presentation of these options will be counterbalanced across participants. Participants will be incentivized to answer the confidence and manipulation check measures at the same rate (i.e., $.25).

Concerns: demand effects

One concern is that the manipulation will elicit demand effects. In this situation, participants may choose to prepare more when they are told their gender performs poorly on the task because they may be able to recognize our hypothesis and want to behave in ways that align with the hypothesis. We argue that demand effects are not problematic for interpreting the results for two reasons. First, it is unlikely participants will be sufficiently motivated by the unpaid preparation to succumb to demand effects, even if they know our hypothesis. Only 12% of U.S. MTurkers indicate that "MTurk money is irrelevant" and another 12% indicate that "MTurk is my primary source of income" (Mason & Suri, 2012), suggesting that many MTurkers try to maximize the amount of money they make in a given amount of time while on the platform, and are unlikely to be motivated by unpaid work. Even if participants are motivated to align their behavior with our hypothesis, there are many tasks in the real world where task stereotypes about gender differences in performance are either implicitly or explicitly stated (Grosse & Riener, 2010), with the assumption that one gender must exert more effort to "compensate" for a lack of ability. Thus, participants' behavior in the study, even if driven by demand, will likely mirror effects we see in the real world, and as a result, will have real-world implications, especially if women are preparing more than necessary based on inaccurate stereotypes.

Primary hypotheses and analyses

We predict women will choose to prepare more than men before a competition, especially when the task is male-typed. We will test the interaction between gender and condition (male-typed or female-typed task) using a linear regression with amount of time a participant chose to prepare (log-transformed) as the dependent variable. We will also control for the number of problems participants completed in the amount of time allotted and the total time they paused the practice. Thus, the following linear regression will be run:

$$\text{Log(Time Preparing)} = \beta_0 + \beta_1\text{Gender} + \beta_2\text{Condition} + \beta_3\text{Number of problems completed} + \\ \beta_4\text{Total pause time} + \beta_5\text{Gender} \times \text{Condition},$$

where the female-typed task and men will be coded as the reference groups for Condition and Gender, respectively.

A positive beta coefficient for the interaction term between the gender variable and the task condition would suggest that the manipulation elicited greater practice in women when they expected men to perform better. We do not expect the female-typed task will encourage men to practice significantly more than women, largely because men tend to be more confident on average than women (Bertrand, 2010;

Beyer, 1990; Beyer & Bowden, 1997; Croson & Gneezy, 2009; Lundeberg et al., 1994; Mobius et al., 2011; Niederle & Vesterlund, 2007, 2011).

Exploratory analyses

We will test whether the role of confidence or risk attitude on time spent preparing differs based on participant gender and condition. To this end, we will run separate multiple regression analyses to test whether confidence or risk attitude interact with gender and condition to affect time spent preparing (after log transformation) as the dependent variable. Therefore, the model will be structured as follows, where X_1 is either participants' confidence or risk attitude:

$$\text{Log(Time Preparing)} = \beta_0 + \beta_1\text{Gender} + \beta_2\text{Condition} + \beta_3 X_1 + \beta_4\text{Gender} \times \text{Condition} + \beta_5\text{Condition} \times X_1 + \beta_6\text{Gender} \times X_1 + \beta_7\text{Gender} \times X_1 \times \text{Condition},$$

where the female-typed task condition is the reference group for the condition variable and men are the reference group for the gender variable.

Additionally, we will compare participants' time spent preparing (log-transformed) based on their responses to the manipulation check. Instead of excluding participants who fail the manipulation check, we will use this as a source of information about how participants' beliefs affect their decision to prepare, even if they did not pay attention to the manipulation. To this end, we will run a linear regression with gender and participants' response to the manipulation check as predictors and their time spent preparing as the outcome. Thus, the following linear regression will be run:

$$\text{Log(Time Preparing)} = \beta_0 + \beta_1\text{Gender} + \beta_2\text{Manipulation check} + \beta_3\text{Gender} \times \text{Manipulation check},$$

where responses that gender differences in performance were not found in the previous study and men will be coded as the reference groups for Manipulation check and Gender, respectively. A positive beta coefficient for the interaction term between the gender variable and the manipulation check responses where participants believed men performed better would suggest that the manipulation elicited greater practice in women when they held that belief. For the interaction between gender and the manipulation check responses where participants believed women performed better, we expect the coefficient will either be close to zero and nonsignificant or significant and positive (albeit smaller than the interaction between gender and beliefs that men performed better based on the manipulation check), like our predictions in the primary hypothesis.

Broader impacts

Much of the research on gender differences in competitiveness has focused on designing interventions that increase women's willingness to compete. Less work has paid attention to the downstream consequences of said interventions. If we find that women spend more time preparing on average than men, and possibly overprepare, this would challenge prevailing views that gender differences in labor market outcomes could be reduced or eliminated by simple interventions. Indeed, there are opportunity costs to (over)preparing, including both economic and social costs, such as lost time with family and friends and missed advancement opportunities.

Relatedly, if women *expect* that they will prepare more in competitive environments, this may, in turn, impact whether they even enter competitive environments. Thus, while our prior work suggests that merely giving women more time to prepare does not make them more willing to compete (Richards et al., in prep), anticipated effort could still influence labor market outcomes by affecting women's decisions to enter certain fields or compete for promotions, for instance. In our studies, we use relatively unimportant tasks that are unlikely to greatly impact one's earnings. Yet, our previous work shows a striking gender difference in preparation, suggesting that our study likely *underestimates* gender differences in choices to prepare for tasks that are more important for one's career and economic prospects. In this way, our study is providing a conservative test of the gender differences in effort and preparation in the real world.

Also, the NSF DDIG will improve Ms. Richards' ability to produce high-impact work that will enable her to pursue a tenure-track faculty position as a woman of color. Women of color are currently underrepresented in both the fields of psychology and economics. As a tenure-track professor, she will be able to serve as an important role model for young women and people of color in the academy. Indeed, Ms. Richards is dedicated to promoting diversity in academic STEM disciplines both in her service and research activities. Outside of research, Ms. Richards serves as a mentor with the University of Pennsylvania College Achievement Program Graduate School Mentoring Initiative, which helps undergraduate students from disadvantaged backgrounds (e.g., first-generation, low-income) apply to graduate school. As research coordinator for the Upward Bound Math and Science Summer Scholars Academy at the University of Pennsylvania, Ms. Richards led a group of first-generation and/or low-income high school students in preparing a competitive application for the George Washington Carver Science Fair by providing guidance and feedback, improving their chance of earning an academic scholarship from Temple University. Over the course of the 6-week program, Ms. Richards also cultivated the students' passion for research by teaching them how to review background literature, generate novel hypotheses and appropriate methodology to test them, analyze results, and craft a scientific poster based upon their research.

Finally, the findings will be submitted to high-impact journals and communicated to the general public. Dr. Apicella has published in top academic journals (e.g., *American Economic Review*, *Nature*, *PNAS*) and is also active in science outreach, including authoring *New York Times* editorials, lecturing at local museums and appearing as an expert guest on podcasts and television shows (e.g., *Brain Games*).

Future directions

There are a number of avenues for future research in this area. First, we would like to test the robustness of gender differences in preparation outside of online and laboratory settings. Do these findings translate to real-world settings? Exploring the gender difference in preparation cross-culturally would also shed light on the universality of the finding and help to identify cultural, ecological and social factors that exacerbate it. Given Dr. Apicella's expertise on cross-cultural research, including work with hunter-gatherers (Apicella, 2014, 2018; Apicella et al., 2014, 2017a, 2007a, 2007b, 2018; Apicella & Barrett, 2016; Apicella & Crittenden, 2015; Apicella & Dreber, 2015; Apicella & Feinberg, 2009; Apicella, Marlowe, Fowler, & Christakis, 2012), exploring the generalizability of these findings across cultures is a real possibility.

A second important extension of the work would be to examine how anticipated preparation or workload influences women's decisions to enter competitive environments. While we did not find that giving women time to prepare makes them more likely to compete, it is still possible that women know that they will end up preparing more in competitive situations and thus, select out of them. As mentioned earlier, there are opportunity costs to preparing. A third extension of the current work would be to examine whether women are overpreparing. Does preparation negatively impact women? Does it help women? To determine whether men or women are preparing more (or less) than needed, future research should test whether gender and time chosen to prepare interact to affect a participants' probability of winning a competition (see Niederle & Vesterlund, 2007). Another follow-up study could manipulate whether there is a monetary cost for preparing to explore whether gender differences in the choice to prepare persist despite a clear cost, and whether this leads to gender differences in earnings within the study.

While we build off an extensive and laudable literature on gender differences in competitiveness, we have unearthed a gender difference in preparation. As this is a new area of research, there are many promising and exciting avenues for future exploration, all of which have the potential to inform policies that mitigate gender disparities in the labor market.