University of California
Santa Barbara

Gender, Competition, & Confidence with Methodological Insights: Experimental Evidence

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Economics

by

Brianna Noelle Halladay

Committee in charge:

Professor Gary Charness, Chair
Professor Peter Kuhn
Professor Emanuel Vespa

March 2017

The Dissertation of Brianna Noelle Halladay is approved.

Professor Peter Kuhn

Professor Emanuel Vespa

Professor Gary Charness, Committee Chair

March 2017

Gender, Competition, & Confidence with Methodological Insights: Experimental Evidence

To my husband.

Acknowledgements

I've long thought about what I would write for the Acknowledgments section of my dissertation. A simple thank you to all those people on this journey with me will never fully express my gratitude.

To my patient, loving, and incredibly supportive husband - Jonathon Halladay. I could not have done this without you! You truly make me better and we are a great team.

To my sister - Nicole Venegas. We've been through everything together and I would not be finishing with such success if it were not for your support, guidance, and love. We can get through anything together.

To Monica and Dan. We've been on this journey together and I am so glad we could finish it together. Thank you for always being there to listen and relate to the craziness.

To the professors from the UCSB Economics Department, I am immensely grateful for all the knowledge and opportunities you brought into my life. Your willingness to always lend a listening ear and helpful advice did not go unappreciated. As I enter my new life as a professor, I promise to care for my students just as you've cared for me both intellectually and personally.

To the staff from the Economics Department, especially Mark Patterson. Thank you for all your help.

To Professor Peter Kuhn. I am still amazed at the vastness of your knowledge and am so grateful that you were willing to share it with me. Thank you for your generosity with your time, even when you were out of the country. Your input through this process was and remains incredibly invaluable.

To Professor Emanuel Vespa. Thank you for putting your foot down and demand-

ing we change my job market paper. Thank you for your honesty and your incredible support. You knew when to be tough on me which made the success that much better!

To my advisor, Professor Gary Charness. I literally hit the jackpot having your as my advisor. Thank you for believing in me, supporting me, being honest with me, and always being open with me. I am the economist I am today because of you. I have a career I am excited about because of you.

Curriculum Vitæ

Brianna Noelle Halladay

Education

2017	Ph.D. in Economics (Expected), University of California, Santa Barbara.
2012	M.A. in Economics, University of California, Santa Barbara.
2011	M.A. in Economics, San Diego State University.
2009	B.A. in Economics, Boston University

Publications

"Experimental Methods: Pay One or Pay All," Joint with Gary Charness and Uri Gneezy, *Journal of Economic Behavior & Organization* (November 2016)

Research Interests

Experimental and Behavioral Economics, Gender, Personnel Economics, Applied Microeconomics

Teaching Experience

Economics 1: Principles of Microeconomics, Fall 2012, Winter 2013, Fall 2013, Winter 2014, Fall 2016
Economics 100B: Intermediate Microeconomic Theory, Winter 2014, Fall 2014, Winter 2016, Winter 2017
Economics 140A: Introductory Econometrics, Spring 2013, Winter 2015, Spring 2015, Spring 2016
Economics 240C: Econometrics with Emphasis in Time Series and Forecasting (Masters), Spring 2014
PSTAT 109/5E: Statistics for Economics: Spring 2014, Winter 2015

Conference Presentations and Seminars

2016: UCSB Experimental Reading Group, UCSB Labor Lunch, Economic Science Association North America Meetings, Southern Economic Association Meetings
2015: UCSB Experimental Reading Group
2014: UCSB Experimental Reading Group, The Choice Lab and Rady School of Management Spring School in Behavioral Economics, IFREE Graduate Student Workshop in Experimental Economics

Honors and Awards

University of California, Santa Barbara
Graduate Student Association Excellence in Teaching Award (Social Sciences), 2015
Graduate Division Dissertation Fellowship, Fall 2015
Graduate Student Research and Travel Grant, Winter 2014, Spring 2015, Winter 2016
Nominated for Academic Senate Outstanding Teaching Assistant Award, 2015
Economics Department Teaching Assistant of the Year, 2014
Myerson Fellowship, 2012
Andron Fellowship, 2011-2012
Block Grant, 2011-2012

San Diego State University
McCuen Scholarship, 2010-2011
Daniel Weintraub Research Scholarship, 2010
Terhune Scholarship, 2009-2010

Boston University
University Scholarship, 2006-2009

Computer Skills

Proficient: STATA, zTree, LaTeX, Eviews
Familiar: Matlab

Abstract

Gender, Competition, & Confidence with Methodological Insights: Experimental Evidence

by

Brianna Noelle Halladay

This dissertation is a collection of two laboratory experiments addressing leading empirical questions in behavioral responses to competition including observed gender differences and one survey paper addressing methodological concerns facing experimental economists.

In the first paper, I use a laboratory experiment to investigate the role of the gender perception of the task in tournament selection decisions. I find that while women enter the tournament at significantly lower rates than men under the male stereotype task, more women than men enter the tournament under the female stereotype task. It appears that greater female confidence and lower male confidence in the female stereotype task drives this observed difference in entry rates across tasks.

In another laboratory experiment, I investigate whether past history with an individual, specifically negative history affects behavioral responses in a tournament with the individual. Using a laboratory experiment, I find that while women respond to the negative history with increased performance in the tournament, male performance in unaffected regardless of emotional stimuli.

In the third paper, I further the discussion of payment schemes for laboratory experiments including the necessity of providing subjects with choices that are incentive compatible.

Contents

This page intentionally left blank

Chapter 1

Perception Matters: The Role of Task Gender Stereotype on Confidence and Tournament Selection

1.1 Introduction

A gender gap in competitiveness is well documented in the literature. There are two possible channels that may be responsible for the observed gap. It is possible that this gender gap is driven mainly by women's distaste for competition. However, it is also possible that this gender gap is driven mainly by a difference in beliefs about future performance across genders. In their seminal paper, Niederle and Vesterlund (2007)[1] note, "However 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." Currently, there is little research analyzing which channel is responsible for the gender gap in competitiveness. This study will serve as a first step in distinguishing between these two possible channels.

I use a controlled laboratory setting to analyze tournament selection using a task that carries a female gender stereotype. While much of the previous research has employed tasks carrying either a male gender stereotype or a "neutral" gender stereotype, my study adds to the body of research by using a task that previous research suggests carries a female stereotype.

This experiment exploits the gender stereotype of two tasks: an arithmetic task gender stereotyped in favor of men and a facial emotion recognition task gender stereotyped in favor of women. While both men and women have the capacity to excel at math, society traditionally views math as a male domain. Lundeberg et al. (1994)[2] find evidence that in certain settings, specifically math, men were more confident than women. Nosek et al. (2002)[3] echo this finding in their series of studies. In particular, the authors consistently find that women have more negative views of math than men both implicitly and explicitly.

Neurological research substantiates the belief that women are more intuitive and emotional beings than men. In tasks where men and women assess the approachability of an individuals face, interestingly, men take significantly longer to make judgments despite no significant difference in accuracy found (Hall et al., 2012)[4]. Also, using fMRI data, Hall et al. (2012)[4] find that men show significantly more brain activity than women when making the facial judgments. As a control, brain activity was consistent across both genders when making determinations of sex based on a photograph, suggesting that social decisions are more taxing and complex for

men. In an online study for the Edinburgh International Science Festival involving more than 15,000 people, 77% of women judged themselves are highly intuitive while only 58% of men judged themselves are highly intuitive[5]. Additionally, women are better at expressing their emotions and are therefore viewed as more empathic than men. A researcher from the Greater Good Science Center, Emiliana Simon-Thomas (2007)[6], notes, "We all know the stereotype: Women are better than men at taking other people's perspectives, feeling their pain, and experiencing compassion for them." In a study conducted in Germany, 63% of Germans believed women had better intuition about selecting a romantic partner. A majority of subjects also believed that women's intuition was overall better in regards to decisions about one's personal life. Gigerenzer et al. (2013)[7] conclude that this confirms the validity of the common gender stereotype that women are between at inferring people's intentions.

While there are many factors that may influence an individual's decision to enter a competition in the field, the control of the laboratory allows me to isolate differences in behavior across task gender stereotype, absent concerns about discrimination for example. I test whether men enter a tournament more often than women when the task carries a female stereotype as opposed to a male stereotype.

For the purpose of this study, I replicated the study by Niederle and Vesterlund (2007)[1] and added an additional treatment. Niederle and Vesterlund (2007)[1] use a math task in their study, which can be thought of as carrying a male stereotype. For the additional treatment, I follow the experimental design of Niederle and Vesterlund (2007)[1], but use a task that carries a female stereotype. My experiment uses a between-subjects design such that no subject participates in more than one treatment. Subjects participate in groups of four, two men and two women. In both treatments, subjects complete the task under a piece-rate compensation scheme and then under

a tournament compensation scheme. Next, subjects can select which compensation scheme they prefer to determine compensation in the third round. In the final round, subjects choose whether they want their piece-rate performance paid according to a piece-rate compensation scheme or a tournament compensation scheme. While subjects are aware of their own absolute performance, at no point until the end of the experiment do subjects learn about their relative performance. Subjects are randomly paid for one of the four rounds. Additionally, subjects report beliefs about relative performance, participate in a Gneezy and Potters (1997)[8] risk assessment, and complete the Cognitive Reflection Test (Frederick, 2005)[9].

By comparing tournament entry rates across gender and across tasks, I determine if the task affects the tournament entry decision. Specifically, I analyze how female tournament selection responds to a task carrying a female stereotype compared to a task carrying a male stereotype. Higher tournament entry rates for men than women in both tasks is evidence for the distaste for competition channel, while observing more tournament entry for women in the female stereotype task than men provides evidence for the gender difference in beliefs about future performance channel.

The main finding is, controlling for performance, men enter the tournament significantly more than women in the male stereotype task, yet more women than men enter the tournament when the task carries a female stereotype, though this difference is not statistically significant. Additionally, female (male) confidence in relative performance appears higher when the task carries a female (male) stereotype and male (female) confidence in relative performance appears lower when the task carries a female (male) stereotype. Further, my findings in the math task replicate the findings of Niederle and Vesterlund (2007)[1].

The paper proceeds as follows. Section 2 provides an overview of previous research

related to the my research question. Section 3 identifies my experimental design while section 4 explores my hypotheses and predictions. I present my results in section 5 and discuss the implications of these results in section 6. Section 7 concludes.

1.2 Literature Review

In studies where a competitive environment is introduced and subjects do not have the ability to opt out of the competition, differences in changes in performance across genders depend on task and opponent gender. Gneezy et al. (2003)[10] use a maze task to assess responses to competition. They find that while male performance in the task rises significantly in the presence of competition, female performance remains static when competing against men. However, in single sex tournaments, women did respond to competition. Gneezy and Rustichini (2004b)[11] assess responses to competition among children. The authors observe no difference in running speed across boys and girls in the noncompetitive environment, but do observe a significant increase in the running speed of boys after introducing a competition. The running speed of girls was unaffected by the introduction of the competition. Replicating and extending the work by Gneezy et al. (2003)[10], Günther et al. (2010)[12] find that men respond to competition while women do not in the maze task (which they denote the male stereotype task), men and women both respond to competition in the word task (which they denote the neutral stereotype task), and women respond to competition while men do not in the memory/distraction task (which they denote the female stereotype task). It is not clear why this is considered a female stereotype task, nor do the authors have any evidence about gender differences in response to competition when subjects are given the opportunity to opt out of the tournament. Replicating and extending Gneezy and Rustichini (2004b)[11], Dreber et al. (2011)[13]

find that there is no difference in responses to competition across genders for any of the tasks selected (running, jumping rope, and dancing). I find that in both tasks, men and women increase performance when the competitive environment is introduced. Additionally, there is no gender difference in this increase in performance in either task. Another important question to ask is how subjects respond when given the opportunity to opt out of a competition.

Gneezy and Rustichini (2004a)[14] use two tasks varying in task gender stereotype. The authors selected a sports task as the male task and a verbal task as the female task. Men select into the tournament more often than women in both tasks but the gender gap is smaller with the female task. Given that subjects observe the gender of their selected opponent prior to making their compensation scheme decision, Gupta et al. (2005)[15] find that men choose the competition more often than women. They suggest that risk aversion primarily drives the choice for women. Despite similar performances in the task for men and women, Niederle and Vesterlund (2007)[1] find that men enter the tournament twice as often as women. The authors documented a preference for competition, but did not distinguish whether it is driven by distaste for competition or a belief about future performance in the competition. My study will help to distinguish whether the observed gender gap is arising out of a gender difference in preference for competition or a gender difference in beliefs about future performance. Again, despite similar performance, in the arithmetic task, I find that more men enter the tournament than women while in the facial emotion task, more women than men enter the tournament.

Recent research has provided evidence about environments that aid in closing the gender gap in response to competition. [16] Using a maze task, Niederle and Yestrumskas (2008)[17] assess gender differences in preferences for performing harder tasks in

order to potentially increase one's payoff. Men select the harder task significantly more often than women despite no differences in performance and beliefs about relative performance. Given a more flexible choice such that subjects can adjust their choice between rounds, there is no longer a gender gap in selection of the harder task. In a study assessing differences in responses to competition across a matrilineal society and a patrilineal society, Gneezy et al. (2009)[18] find that women choose the tournament more often than men in the matrilineal society while men choose the tournament more often than women in the patrilineal society. Team competition is another environment in which the gender gap in response to competition appears to eliminate the gender gap in response to competition (Dargnies, 2012[19]; Kuhn and Villeval, 2014[20]). Evidence provided by Booth and Nolen (2012)[21] suggests that "nurture" may play a significant role in the commonly observed gender gap in tournament entry decisions. In a study assessing the tournament entry decisions of girls from single-sex schools and girls from coed schools, the authors find that girls from single-sex schools choose the competition at the same rate as boys. Using the math task from Niederle and Vesterlund (2007)[1], Niederle et al. (2013)[22] implement an affirmative action quota in the laboratory and find that women significantly increase their tournament entry while men significantly decrease their tournament entry. Cassar et al. (2016)[23] find that creating tournament prizes that benefit one's child such as workplace daycare as opposed to monetary prizes erased the gender gap in tournament selection. Halladay (2016)[24] finds that women will perform in a competition just as well as men in the presence of negative emotions towards the opponent.

A few studies have attempted to analyze the effect of a non-male stereotype task on tournament selection. Grosse et al. (2014)[25] use a neutral stereotype task

and two male stereotype tasks. The male stereotype tasks include a sports task where subjects throw tennis balls in a bucket and the math task as in Niederle and Vesterlund (2007)[1]. The neutral stereotype task involved a list of five words that a subject needed to place in order to form a grammatically correct sentence. While the authors observe difference in tournament selection rates across genders in the male stereotype tasks, there is no observed difference in the neutral task. However, these results are a bit concerning given the complex nature of the experimental design and the fact that performance was not similar across genders for the selected tasks. To avoid this concern, I have chosen two tasks that ex-ante have similar performance across genders.

In a study perhaps most closely related to my research, Dreber et al. (2014)[26] utilize both a male stereotype task and a female stereotype task in order to assess gender differences in self-selection into tournaments. The authors choose a math task similar to the one used in Neiderle and Vesterlund (2007)[1] for the male stereotype task and a crossword-like puzzle for the female stereotype task. While math tasks undoubtedly carry a male dominant perception, gender perception is less clear for verbal tasks. IT is not clear that verbal tasks carry a female stereotype. A meta-analysis by Hyde and Linn (1988)[27] of gender differences in verbal tasks finds an insignificant effect size of +0.11 in favor of females while other studies mentioned above have used verbal tasks as "neutral". Additionally, previous studies assessing gender differences in self-concept in verbal tasks find no gender difference (Hyde et al., 1990[28]; Meece et al., 1982[29]; Skaalvik and Skaalvik, 2004[30]). Also, by selecting a math and a verbal task in conjunction with a within-subjects design, there are certain problematic interactions regarding how individuals perceive their math versus verbal ability. Because subjects are likely to view math and verbal abilities as dichotomous

and are completing both tasks, the selection of these two specific task categories can influence subjects choices and behavior. The authors find that men select into the tournament more than women in both the math and verbal tasks but the difference in tournament selection is only significant in the math task. While this result is interesting, given the concerns above, more research is warranted. My study will help to clarify the validity of these results absent the concerns mentioned above.

1.3 Experimental Design

The experimental design follows almost exactly that of Niederle and Vesterlund (2007)[1]. Instructions were adapted from the instructions available on the Lise Vesterlund's website and available in the appendix.

The experiment takes place in four main stages, lasting five minutes each followed by a few simple tasks. The stages differ by the compensation scheme while the task remains the same across all stages. Participants only learn of the compensation scheme differences immediately before each stage.

There are two treatments differing by the task. One task is identical to the math task of Niederle and Vesterlund (2007)[1] while the other task requires subjects to identify facial emotions in a series of images.

In the math task subjects must add up a series of five randomly selected two-digit numbers. Subjects have five minutes to correctly answer as many problems as they can. Subjects cannot use a calculator, but can use the provided scratch paper.

The facial emotion task utilizes professionally classified images obtained from The Great Good Science Center at the University of California, Berkeley. In each stage subjects view a series of 15 images, each appearing on the screen for two seconds, and subjects attempt to correctly select the depicted emotion out of four options. Subjects

have 20 seconds to submit their answer after each image is displayed. Subjects know that the same emotions can repeat but the same image will never appear more than once.

In both tasks, subjects immediately observe whether the answer submitted was correct or not while also generating a new series of numbers or image depending on the treatment. Consistent with Niederle and Vesterlund (2007)[1], in each stage, subjects continually observe their absolute performance, but subjects never observe their relative performance until the end of the experiment.

In the first stage, payment follows a piece-rate compensation scheme. For each correct answer, subjects earn 50 cents.

In the second stage, payment follows a tournament compensation scheme. Payment depends on the performance of the other three subjects in a participant's group. If a participant submits more correct answers than her other three group members, she receives $2 for each correct answer, and nothing otherwise.

The third stage gives subjects the opportunity to select which compensation scheme they want applied to their performance in the third stage. If a subject selects the piece-rate compensation scheme, she receives 50 cents per correct answer in the third stage. If a subject selects the tournament compensation scheme, her earning depend on her performance in the third stage relative to the performance of her other three group members in the second stage tournament. If the subject submits more correct answers in the third stage than her other three group members did in the second stage tournament, she receives $2 per correct answer, and nothing otherwise. Niederle and Vesterlund (2007)[1] note that they designed this stage specifically as an individual choice, independent of the choices of the other three group members. They also note that by comparing a subjects Task 3 performance to others' Task

2 performance, subjects need only base their choice on their beliefs about relative performance in a tournament and not beliefs about who will or will not enter the tournament. Lastly, Niederle and Vesterlund (2007)[1] explain that an observed gender gap in the choices of subjects in Task 3 indicated a gender difference in preference for performing in competitive environments.

The fourth, and final main stage gives subjects another opportunity to select a compensation scheme. However, in the fourth stage, subjects do not need to complete any additional tasks. Task 4 was designed to distinguish between a difference in preference for competition across genders or perhaps differences in overconfidence, risk, or feedback aversion across genders. Subjects decide whether they want their performance from the first stage piece-rate round to be submitted to a tournament or to remain piece-rate. If a subject selects the piece-rate compensation scheme, she receives 50 cents per correct answer. If a subject selects the tournament compensation scheme, her earnings depend on her performance in the first stage relative to the performance of her other three group members in the first stage. If the subject submits more correct answers in the first stage than her other three group members did, she receives $2 per correct answer, and nothing otherwise. Before submitting their choices, subjects are reminded of their first stage performance.

At the conclusion of the fourth stage, subjects report their beliefs about their relative performance. First subjects guess their rank in the first stage piece-rate task and then subjects guess their rank in the second stage tournament. Guesses are incentivized. Subjects earn $1 for each correct guess.

Additionally, subjects complete a risk elicitation. The risk elicitation follows the adaptation of the Gneezy and Potters (1997)[8] method utilized by Charness and Gneezy (2010)[31]. Subjects have 200 tokens ($2) to keep or invest in a risky project.

Subjects can choose to invest any integer value of their tokens between zero and 200, inclusive. The risky project has a 50% chance of success. If the project is successful, a subject receives 2.5 to 1 on the invested amount. If the project fails, whatever is not invested is kept. We randomly draw a number between one and four. If we draw a one or a two, the project is successful. If we draw a three or a four, the project is a failure.

After observing the risk elicitation results, subjects answer three logical questions based on the Cognitive Reflection Test (Frederick, 2005)[9]. Subjects have five minutes to answer all three questions. This experiment uses three questions in the same spirit of the Cognitive Reflection Test, but not the exact three proposed by Frederick (2005)[9] because of a concern that a significant portion of the subject pool had been previously exposed to the Cognitive Reflection Test. The three questions used follow those of Gillen, Snowberg, and Yariv (2015)[32]. Subjects can earn 50 cents per correct answer. After the five minutes elapse, subjects are asked how many of the three questions they believe they answered correctly and also what percent of subjects they believe answer more questions correctly than they did.

One of the four main stages is randomly selected for payment. In addition to payment from the randomly selected stage, subjects earn a $5 show-up fee, payment for correct guesses about relative performance, payment from the results of the risk elicitation, and payment from the Cognitive Reflection Test.

The experiment was programmed in zTree (Fischbacher, 2007)[33]. All sessions took place in the Experimental and Behavioral Economics Laboratory at the University of California, Santa Barbara. The University's ORSEE system was used to recruit subjects. Three or four groups of four participants participated in each session. Subjects were seated four per row, each row consisting of two males and two

females although gender was never explicitly discussed. Groups consisted of the other participants sitting in the same row as a subject. A total of 154 students participated, 76 subjects (38 males and 38 females) in the math task, and 72 subjects (36 males and 36 females) in the facial emotion task. No subject participated in more than one session or more than one treatment. Average earnings were $14.48 and each session lasted approximately an hour.

1.4 Hypotheses

This experiment addresses three main hypotheses about how women and men respond to competitive environments.

Both experimental tasks, the arithmetic task and the facial emotion task, involve very little prior knowledge. There is no general consensus that one gender is better than another at either task but each task carries a gender stereotype. In their selection of this specific arithmetic task, Niederle and Vesterlund (2007)[1] note that in a meta-analysis of over 100 gender studies, there was no observed difference in male and female performance on computational math tasks. In a study assessing gender differences in facial emotion recognition, Hoffmann et al. (2010)[34] find that when the emotional stimuli is highly expressive, men and women perform equally well on the task. Additionally, Hoffmann et al. (2010)[34] point out evidence from previous research is mixed, some studies report no gender difference while others suggest a female advantage. The studies that report a female advantage provide very small mean effect sizes. Therefore, it seems very reasonable to hypothesize that there will be no gender difference in performance for either task. Equal performance across genders is a very important task characteristic for the purpose of this study, allowing me to rule out gender differences in tournament entry decisions arising from differing

performance levels. To summarize:

Hypothesis 1: There is no gender difference in performance for either task.

There appears to be a significant confidence gap between men and women. The Harvard Business Review mentions a Hewlett-Packard study that found women would only apply for a promotion when they felt they met 100% of the qualifications while men would apply when they met at least 60% of the qualifications [35]. However, because evidence discussed above alludes to the fact that women are more confident in their ability to harness their "female intuition" as opposed to their math ability, I hypothesize that women will be more confident in the facial emotion task. Succinctly,

Hypothesis 2: Women are more confident than men in the facial emotion task, while men are more confident than women in the arithmetic task.

If women are more confident in their relative ability in the facial emotion task, it is logical to hypothesize that women will therefore be more willing to enter a tournament in the facial emotion task as opposed to the math task. Additionally, in their experiment, Charness et al. (2016)[36] find that differences in confidence across genders can explain the gender difference in tournament entry rates. Claude Steele (1997)[37] details how an individual who identifies with a specific domain, in particular a domain that carries a relevant stereotype, places themselves in a potentially self-threatening position. As a result, in many instances, individuals will either avoid the identification all together or disidentify with the relevant domain. In the arithmetic task, a female who selects the tournament (therefore identifying with math) is thus in a position to possibly validate the existing stereotype that math is for men if she fails

to win the tournament. Consequently, females avoid the tournament in the arithmetic task. Nosek et al. (2002)[3] point out that because math is so heavily aligned with men, the stronger the association, the more positively men view math, hence one can expect to view higher tournament entry rates for men in the arithmetic task. With the facial emotion task, the stereotype threat is absent for women and therefore I expect to see increased tournament selection. Explicitly:

> **Hypothesis 3: Women (men) will select into the tournament more often than men (women) in the facial emotion task (math task), not controlling for confidence.**

1.5 Results

All statistical tests performed in the following analysis are two-tailed unless otherwise specified.

> **Finding 1: Regardless of task, there is no difference in male and female performance under either the piece-rate and tournament compensation schemes.**

Consistent with Niederle and Vesterlund (2007)[1], men and women perform equally well in the arithmetic task under both the piece-rate and tournament compensation schemes. Men solve an average of 9.55 tasks under the piece-rate compensation scheme while women solve 8.58 tasks under the piece-rate compensation scheme. The null hypothesis that these averages are equal cannot be rejected by a two-sided test (p=0.233). The results are similar under the tournament compensation scheme. Men solve an average of 11.42 tasks while women solve 10.13 tasks. Again, the null hypothesis that these averages are equal cannot be rejected by two-sided test (p=0.183).

Additionally, using a KolmogorovSmirnov test there is no difference in the distribution of male and female performance on the arithmetic task under either the piece-rate and tournament compensation schemes (p=0.449 and p=0.287).

Furthermore, there is no difference in male and female performance in the facial emotion task in either the piece-rate or tournament compensation schemes. Men correctly identify an average of 8.19 and 9.72 emotions under the piece-rate and tournament compensation schemes, respectively. Similarly, women correctly identify 8.08 and 9.77 emotions under the piece-rate and tournament compensation schemes, respectively. The differences between these figures are not significant using a two-sided test of averages (p=0.806 and p=0.894) and there is also no statistically significant difference in the distribution of male and female performance under either compensation scheme using a Kolmogorov-Smirnov test (p=0.413 and p=0.965). Figure 1 and Figure 2 depict the cumulative distribution functions for men and women in both tasks under both the piece-rate and tournament compensation schemes.

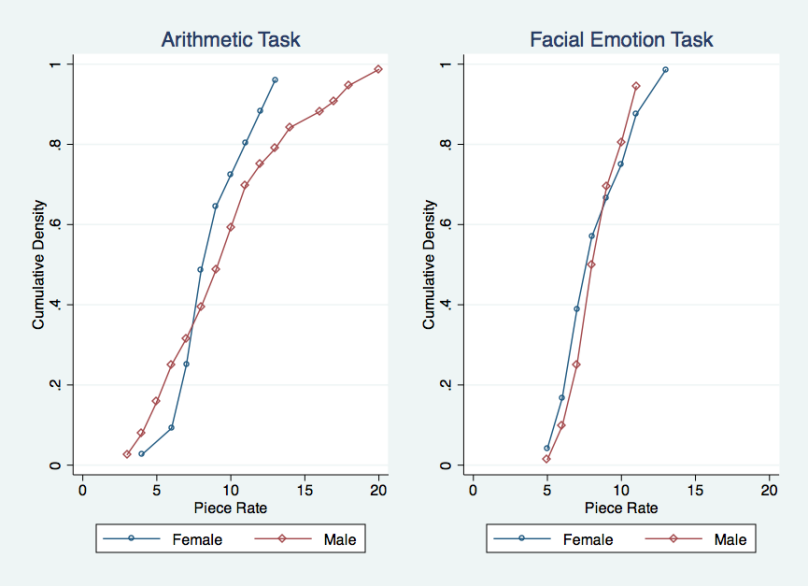

Figure 1.1: There is no difference in male and female performance in either treatment under the piece-rate compensation scheme.

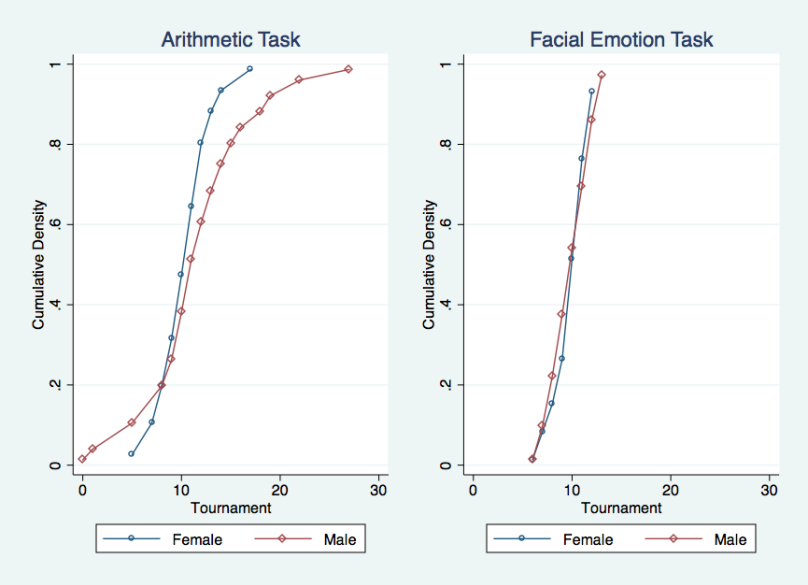

Figure 1.2: There is no difference in male and female performance in either treatment under the tournament compensation scheme.

Finding 2: In both tasks, both men and women perform significantly better in the tournament than in the piece-rate task. Additionally, for both tasks, there is no gender difference in the increase in performance between the piece-rate and tournament compensation schemes.

In both tasks, male and female performances across the piece-rate and tournament compensation schemes are highly correlated. Additionally, in both tasks, both genders perform significantly better in the Task 2 tournament ($p<0.000$ for all four one-sided hypothesis tests). Furthermore, the improvement in performance from the piece-rate to the tournament does not differ across genders for either task ($p=0.546$ for the arithmetic task and $p=0.727$ for the facial emotion task). The distributions of performance improvements across gender are also not statistically different for either task ($p=0.651$ and $p=0.615$).

Because we are unable to reject the null hypothesis that performance is equal across genders for both tasks, we can investigate how tournament selection differs across genders and compare these results to those of Niederle and Vesterlund (2007)[1] without too much concern that any observed differences are due to gender differences in performance.

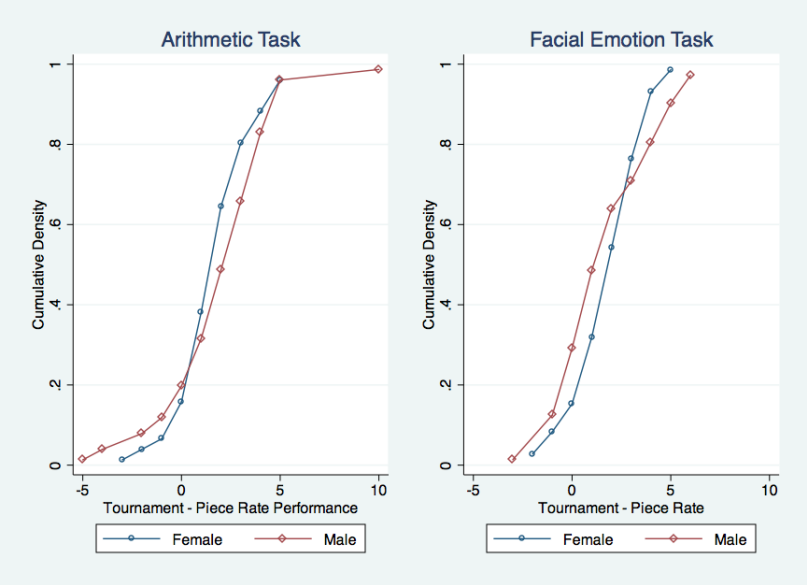

Figure 1.3: There is no difference in the increase in male and female performance between the piece-rate and tournament compensation schemes in either task.

Finding 3: In the arithmetic task, males enter the Task 3 tournament significantly more than females despite the similar Task 1 and Task 2 performances.

Given the similar performances by men and women under both the piece-rate and tournament compensation schemes in either task, one might expect, absent a concern about confidence, tournament entry decisions should also not vary by gender. In contrast to this logical expectation, Niederle and Vesterlund (2007)[1] observe males choosing to enter the tournament significantly more than females. In the arithmetic

task, I find a similar result. While 66% (25/38) of males choose to enter the tournament in Task 3, only 37% (14/38) of females choose to enter the tournament in Task 3 despite the similar Task 1 and Task 2 performances. This difference in tournament entry rates for the arithmetic task is highly significant using a one-tailed test of proportions (p=0.006).

Finding 4: In the facial emotion task, 45% more women than men choose to enter the Task 3 tournament.

In the facial emotion recognition task, the tournament selection results contrast with those of the arithmetic task. More women than men choose to enter the tournament in the facial emotion recognition task. While 31% (11/36) of men choose to enter the Task 3 tournament, 45% (16/36) of women choose to enter the Task 3 tournament, a nearly 50% increase. Strikingly, 45% more women than men choose to enter the tournament in the facial emotion task. A one-tailed test of proportions (consistent with the directional hypothesis) gives p = 0.112, which is not significant but is certainly in the right direction.

Figure 3 depicts tournament entry proportions by gender and task. Running a probit regression of Task 3 compensation choice as a function of a treatment dummy variable, a gender dummy variable, and an interaction of these two variables allows a test of a test of differences-in-differences between male and female tournament selection in Task 3. Each of the reported marginal effects is highly significant suggesting there is something systematically different about tournament selection decisions across tasks (p=0.003, p=0.012, and p=0.009, respectively). A joint hypothesis test about the effect of gender is significant as well as a joint hypothesis test about the treatment effect (p=0.020 and p=0.008, respectively).

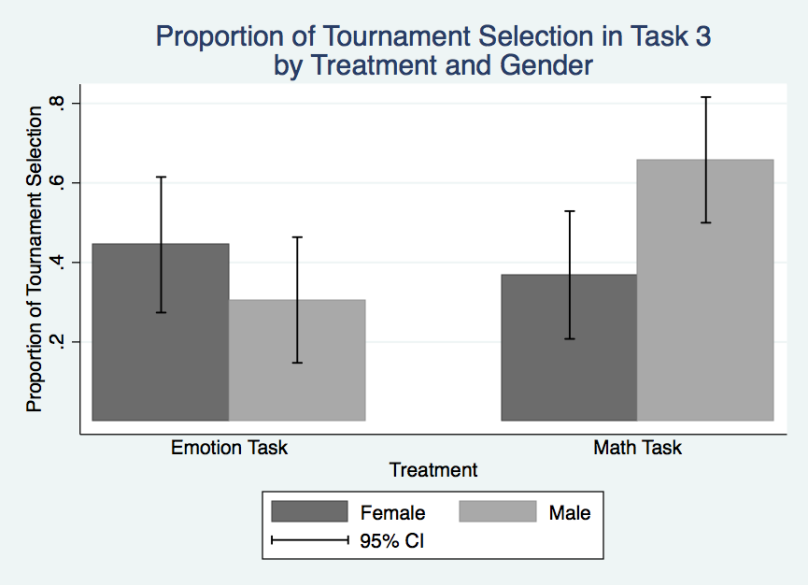

Figure 1.4: While significantly more males than females enter the tournament in the arithmetic task, there is no statistically-significant difference in tournament entry rates across genders in the facial emotion task.

Finding 5: Tournament entry decisions are not driven by performance differences across gender for either the arithmetic task or the facial emotion task.

Niederle and Vesterlund (2007)[1] do not find significant evidence that tournament entry decisions are the result of a difference in performance for those subjects opting into the tournament and those selecting piece-rate across genders. They do report a marginal result for a difference in tournament performance between men selecting into the tournament and men selecting piece-rate. Similarly, for the arithmetic task, I do not find significant evidence that a difference in performance is responsible for the tournament selection decision. I find two marginally-significant results. One is the difference in tournament performance for women who select into the tournament and women who select the piece-rate (p=0.080). A second is the difference in the increase in performance from Task 1 to Task 2 for men who enter the tournament versus men

who select piece-rate (p=0.066). In the facial emotion task, there is no evidence that performance differences drive tournament entry differences for either gender.

Finding 6: While gender is a significant predictor of the tournament entry decision in the arithmetic task, gender does not significantly effect the tournament entry decision in the facial emotion task.

In a probit regression, Niederle and Vesterlund (2007)[1] find that gender was the only significant predictor of the tournament entry decision, controlling for tournament performance and the increase in performance from piece-rate to tournament. Table 1 presents my probit regression results. In the arithmetic task, I find very similar results, with again, only the gender dummy variable yielding a significant coefficient. However, in the facial emotion task, none of the three variables are significant, including gender.

Table 1.1: Probit of Task 3 Tournament Choice

	Arithmetic Task	Facial Emotions Task
	Coefficient (p-value)	Coefficient (p-value)
Female	-0.691**	0.356
	(0.023)	(0.245)
Tournament Performance	0.051	-0.027
	(0.261)	(0.788)
Tournament - Piece Rate Performance	0.101	0.127
	(0.140)	(0.156)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Additionally, while past performance does not appear to be driving the difference in tournament entry, future performance also does not correlate with the tournament entry decision. There is no evidence of a difference in Task 3 performance for men and women in either task regardless of the tournament entry decision (p=0.165, p=0.196, p=0.633, p=0.620).

Finding 7: Gender differences in confidence across tasks appear to help explain the difference in tournament entry rates across genders.

In regards to beliefs about relative performance, while Niederle and Vesterlund (2007)[1] find that both men are women are overconfident, specifically that men are more overconfident than women, I find that only men are overconfident in both tasks using a Fisher's exact test of independence between the distribution of guess rank and actual rank (p=0.000 and p=0.030). A Fisher's exact test of the independence between the distribution of guessed rank and actual rank for women in the arithmetic task and the facial emotion task yield p=0.226 and p=0.847, respectively. It does not appear that women are overconfident in either task. Women appear to be calibrating their beliefs about their performance correctly, while men fail to calibrate their beliefs correctly in either tasks.

Niederle and Vesterlund (2007)[1] observe that the guesses of men and women differ significantly. Using a Fisher's exact test of independence of the distributions for men and women, I also observe that the guesses of own rank by men and women differ significantly at the 10% level (p=0.070) in the arithmetic task, but the guesses of men and women do not differ significantly in the facial emotion task (p=0.376). Additionally, while there is little evidence that the distribution of guesses of women differs across tasks (p=0.131), there is mild evidence that the distribution of guesses for men differs across tasks (p=0.093). Specifically, men have lower confidence in the facial emotion task. While 50% of men guess the top rank in the arithmetic task, only 31% of men selected the top rank in the facial emotion task. A one-sided test of a difference in these proportions yields a significant result (p=0.044). In the arithmetic task, 24% of women guess the top rank while 42% of women guess the top rank in

the facial emotion task. A one-sided test of a difference in proportions is significant (p=0.049). There does appear to be some evidence that men are losing confidence while women are gaining confidence in the facial emotion task.

Table 2 presents the results from an ordered probit regression of guessed tournament rank (guessed ranks of 1, 2, or 3, excluding guesses of 4) as a function of gender, tournament performance, and the increase in performance from piece-rate to tournament. Both both tasks, I find an insignificant coefficient on the gender dummy variable, suggesting that once controlling for performance, confidence is similar across genders for either task. Furthermore, when running an ordered probit of guessed tournament rank conditioning on gender and controlling for both tournament performance and the increase in performance from Task 1 to Task 2, as well as including a task dummy variable and an interaction between female and the task dummy variable, I find evidence of a significant difference in confidence across tasks. There does appear to be a significant treatment effect for women. In particular, the coefficient on the interaction term is significant at the 10% level (p=0.075). Figure 4 and Figure 5 provide additional qualitative evidence that task affects both male and female confidence. Specifically, male confidence falls in the facial emotion task while female confidence increases.

Table 1.2: Ordered Probit of Task 2 Tournament Guessed Ranks

	Arithmetic Task	Facial Emotions Task	Pooled
	Coefficient (p-value)	Coefficient (p-value)	Coefficient (p-value)
Female	0.337	-0.417	-0.440
	(0.300)	(0.153)	(0.122)
Tournament Performance	-0.118**	-0.287**	-0.150***
	(0.031)	(0.21)	(0.001)
Tournament - Piece Rate Performance	-0.190**	-0.212***	-0.224***
	(0.043)	(0.008)	(0.000)
Treatment			-0.491
			(0.104)
Treatment*Female			0.725*
			(0.075)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 3 provides separate probit regressions predicting tournament selection as a function of gender, tournament performance, the increase in performance from the piece-rate to the tournament, and guessed tournament rank, a proxy for confidence. While it appears that the gender gap in tournament selection cannot be explain by beliefs about relative performance in the arithmetic task, once controlling for confidence, there exists no gender gap in tournament entry in the facial emotion task. Due to the finding that there is a treatment effect on confidence across tasks, I tested a pooled probit regression allowing for the treatment effect as well as an interaction between the treatment and guessed tournament rank. With this pooled regression, the gender coefficient is no longer significant (p=0.184), therefore the results suggest that the difference in tournament entry rates across tasks are driven by a difference in confidence across tasks.

Table 1.3: Probit of Task 3 Tournament Choice

	Arithmetic Task		Facial Emotions Task		Pooled
	Coefficient (p-value)		Coefficient (p-value)		Coefficient (p-value)
	(1)	(2)	(1)	(2)	(1)
Female	-0.692**	-0.674**	0.253	0.057	-0.308
	(0.033)	(0.040)	(0.417)	(0.864)	(0.184)
Tournament Performance	0.027	0.021	-0.007	-0.15	0.008
	(0.589)	(0.685)	(0.948)	(0.238)	(0.864)
Tournament - Piece Rate Performance	0.059*	0.045	0.105	0.031	0.017
	(0.051)	(0.619)	(0.246)	(0.754)	(0.789)
Guessed Tournament Rank		-0.142		-0.774***	-0.665***
		(0.592)		(0.003)	(0.003)
Treatment					-0.364
					(0.554)
Treatment*Guessed Tournament Rank					0.419
					(0.193)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

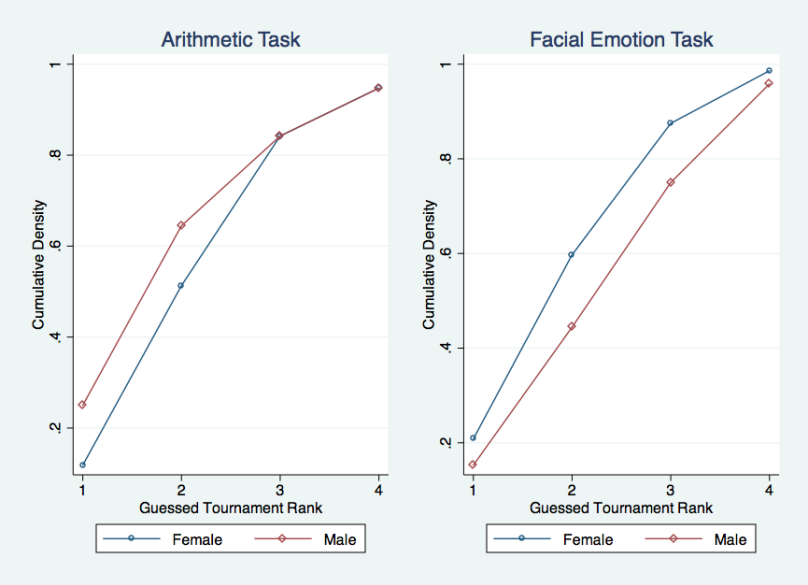

Figure 1.5: While the guess distribution for males lies above that of females in the arithmetic task, it appears that the guess distribution for females always lies above the distribution for males in the facial emotion task.

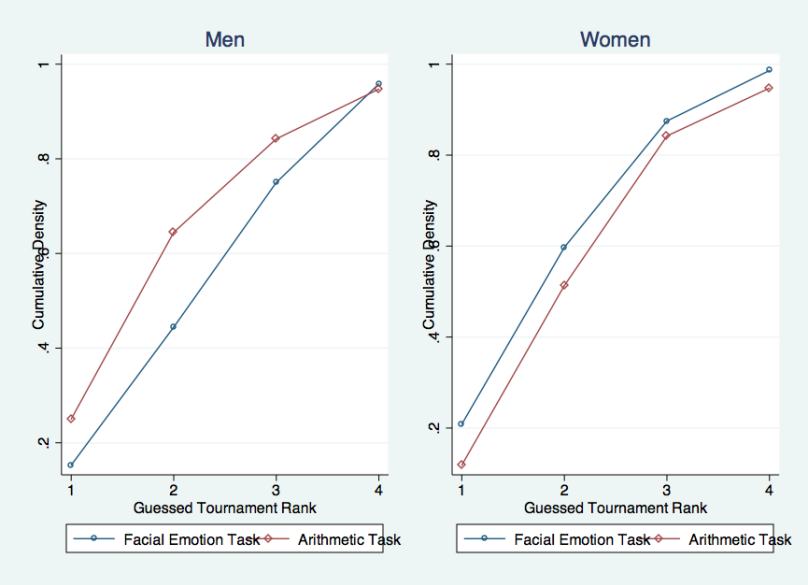

Figure 1.6: The guess distribution for males in the arithmetic task lies above that of males in the facial emotion task, while it appears that the guess distribution for females in the facial emotion task lies above that of females in the arithmetic task.

Finding 8: Controlling for risk preference eliminates the gender

difference in tournament entry rates in the arithmetic task.

Entering a tournament requires a subject to undertake some level of risk. Research suggests that levels of risk aversion differ for men and women. Subsequently, the differing tournament entry rates may be due to a difference in risk preferences. Women invest significantly less than men. While men invest 69% of their endowment, women invest roughly 51% of their endowment (p=0.000). This difference is also significant using a Wilcoxon Rank-Sum test (p=0.000).

The probit regressions of Table 4 suggest the observed difference in tournament entry rates in the arithmetic task mostly vanish when controlling for risk preferences. A one-sided hypothesis test for the coefficient on female suggests the gender gap still partially remains (p=0.057). Additionally, I cannot reject the null hypothesis that there the investment percentage among men who choose the tournament and men who select piece-rate for either task is equal (p=0.473 and p=0.153). Similarly, I cannot reject the null hypothesis that the investment percentage among women who choose the tournament and women who select piece-rate for either task is equal (p=0.672 and p=0.104). Because of these results, risk preference does not seem to be the main driving force behind the difference in tournament entry rates across tasks.

Table 1.4: Probit of Task 3 Tournament Choice Controlling for Risk Preference

	Arithmetic Task			Facial Emotions Task		
	Coefficient (p-value)			Coefficient (p-value)		
	(1)	(2)	(3)	(1)	(2)	(3)
Female	-0.694**	-0.581	-0.581	0.254	0.389	0.196
	(0.033)	(0.113)	(0.113)	(0.417)	(0.239)	(0.578)
Tournament Performance	0.027	0.022	0.018	-0.007	0.029	-0.097
	(0.589)	(0.674)	(0.734)	(0.948)	(0.801)	(0.459)
Tournament - Piece Rate Performance	0.059	0.065	0.055	0.105	0.084	0.012
	(0.506)	(0.464)	(0.559)	(0.246)	(0.375)	(0.905)
Percent Invested		0.471	0.406		1.525**	1.349**
		(0.499)	(0.573)		(0.014)	(0.040)
Guessed Tournament Rank			-0.105			-0.721***
			(0.700)			(0.007)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Finding 9: While there is a significant differences in tournament selection in Task 4 in the arithmetic task, there is difference in tournament selection in Task 4 in the facial emotion task.

In Task 4, subjects can select whether to submit their Task 1 piece-rate performance to a tournament or to a piece-rate compensation scheme in the event that Task 4 is the task randomly selected for payment. Subjects do not need to complete any additional tasks in Task 4 so the choice does not involve the prospect of additional effort. In the arithmetic task, 45% of men select the tournament while only 21% of women select the tournament. In the facial emotion task, 25% of both men and women select the tournament. Using a Fisher's exact test, men and women differ in their Task 4 choice in the arithmetic task (p=0.050) but not in the facial emotion task (p=1.000).

Consistent with Niederle and Vesterlund (2007)[1], there is no difference in piece-rate performance for women who select the tournament versus women who select piece-rate in the arithmetic task (9.125 and 8.43, p=0.468). However, there is a difference in piece-rate performance for men who select the tournament versus men who select the piece-rate in the arithmetic task (11.76 and 7.76, p=0.004). In the facial emotion task, there is no difference in performance for individuals who select the tournament versus the piece-rate compensation scheme for either gender (p=0.692 and p=0.864). These results provide evidence that performance is not driving the tournament selection decision.

Finding 10: Differences in confidence across genders can explain the difference in tournament entry rates in Task 4.

Beliefs about relative performance in Task 1 are consistent with beliefs about relative performance in Task 2. This suggests that the competitive environment is

not affecting beliefs about relative performance. A Fisher's exact test again shows that while men are overconfident in both tasks (p=0.003 and p=0.020), women are overconfident in neither task (p=0.225 and p=0.193). Again, women are calibrating correctly while men fail to accurately calibrate beliefs about relative performance. I observe that the distribution of guesses of men and women differ significantly in the arithmetic task (p=0.000), but the guesses of men and women do not differ significantly in the facial emotion task (p=0.326). However, there is evidence that the distribution of guesses for men differs across tasks (p=0.000) while the distributions of guess for women does not differ across tasks (p=0.880). While it appears that men are overconfident in both tasks, I do not observe the difference in tournament entry rates for men and women across both tasks.

Table 5 presents results from an ordered probit regression of guessed piece-rate rank (guessed ranks of 1, 2, or 3, excluding guesses of 4) on gender and piece-rate performance. I find a significant coefficient on the gender dummy variable, suggesting that despite controlling for performance, there remains a gender difference in confidence for both tasks. In particular, women guess significantly lower ranks than men in the arithmetic task, while women guess significant higher ranks than men in the facial emotion task. Furthermore, running an ordered probit regression of guessed piece-rate rank, conditioning on gender and controlling for piece-rate performance, I find evidence of a difference in male confidence across tasks but not a difference in female confidence across tasks. Specifically, the highly significant coefficient on the treatment variable in column 3 of Table 5 illustrates that men are significantly less confident in the facial emotion task than the arithmetic task.

Table 6 presents probit regression results that suggest controlling for confidence eliminates the observed gender gap in tournament selection.

Table 1.5: Ordered Probit of Task 1 Piece Rate Guessed Ranks

	Arithmetic Task	Facial Emotions Task	Males	Females
	Coefficient (p-value)	Coefficient (p-value)	Coefficient (p-value)	Coefficient (p-value)
Female	1.050***	-0.545*		
	(0.000)	(0.075)		
Piece Rate Performance	-0.201***	-0.278***	-0.274***	-0.156**
	(0.000)	(0.000)	(0.000)	(0.019)
Treatment			-1.299***	0.204
			(0.000)	(0.469)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 1.6: Probit of Task 4 Tournament Choice

	Arithmetic Task		Facial Emotions Task	
	Coefficient (p-value)		Coefficient (p-value)	
	(1)	(2)	(1)	(2)
Female	-0.567*	-0.323	0.001	-0.231
	(0.076)	(0.354)	(0.997)	(0.503)
Piece-Rate Performance	0.138***	0.077	0.018	-0.086
	(0.005)	(0.178)	(0.831)	(0.380)
Guessed Piece-Rate Rank		-0.567**		-0.574**
		(0.033)		(0.029)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Despite controlling for confidence, risk, and feedback aversion through the decision to submit one's piece-rate score to a tournament and beliefs about relative performance in the tournament, Table 7 illustrates that there is still a gender gap in tournament performance in the arithmetic task, significant at the 10% level. While Niederle and Vesterlund (2007)[1] take this as evidence that women and men differ in their preference for competition, this statement needs clarification. This gender gap is not observed in the facial emotion task and is eliminated when controlling for the treatment effect on confidence. A restatement of Niederle and Vesterlund (2007)[1] is that a difference in preference for competition across genders is only part of the story. A difference in beliefs about future performance appears to significantly affect

tournament selection decisions across genders.

Table 1.7: Probit of Task 3 Tournament Choice

	Arithmetic Task			Facial Emotions Task		
	Coefficient (p-value)			Coefficient (p-value)		
	(1)	(2)	(3)	(1)	(2)	(3)
Female	-0.694**	-0.674**	-0.622*	0.254	0.057	0.083
	(0.033)	(0.040)	(0.062)	(0.417)	(0.864)	(0.806)
Tournament Performance	0.027	0.021	0.013	-0.007	-0.150	-0.128
	(0.589)	(0.685)	(0.814)	(0.948)	(0.238)	(0.319)
Tournament - Piece Rate Performance	0.059	0.045	0.079	0.105	0.031	0.049
	(0.506)	(0.619)	(0.429)	(0.246)	(0.754)	(0.635)
Guessed Tournament Rank		-0.142	-0.068		-0.774***	-0.750***
		(0.592)	(0.341)		(0.003)	(0.004)
Submitting the Piece Rate			0.359			0.443
			(0.341)			(0.242)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

1.6 Discussion

My data confirm that altering subject beliefs through the perception of the task significantly affects tournament entry rates. While more men than women enter the tournament with the male stereotype task, more women than men enter the tournament with the female stereotype task, though this difference is not significant. This observed difference seems to be working through the confidence channel. My results provide evidence that male confidence decreases while female confidence increases in the female stereotype task leading to similar tournament entry rates across genders. I find evidence that a gender difference in beliefs about future performance may be driving the observed gender gap in tournament selection as opposed to a gender difference in pure preference for competition.

Performance in both the arithmetic task and the facial emotion task are consistent across genders regardless of compensation scheme. Because we do not observe a significant difference in performance across genders for either task, this observed difference in tournament entry rates does not appear to be driven by a performance

difference, further supporting my results.

One might be concerned that while men were more confident than women in the arithmetic task as expected, there was no difference in male and female confidence in the facial emotion task. Why do we not observed overconfidence for women in the facial emotion task? As Lundeberg et al. (1994)[2] succinctly put it, "...the problem may not be that women necessarily lack confidence, but that in some cases men have too much confidence especially when they are wrong!" Additionally, Lundeberg et al. (1994)[2] note that women are much more able to calibrate their confidence than men. Because men are unable to fully calibrate their confidence, it may look like they are just as confident as women, but the real result lies in the increase in the confidence of women through correct calibration.

Ideally, one would hope women would enter a tournament more than men when the task is in their favor, but as Lundeberg et al. (1994)[2] note, men are consistently overconfident when they have no reason to be. The facial emotion task carries a very strong female gender stereotype, but it may take a task with an even stronger female stereotype for female tournament entry rates to exceed that of males. While it might be possible to find a more female-stereotyped task, any such task would be most useful if there is equal performance across genders.

1.7 Conclusion

Much research suggests that women avoid competition even when it can be beneficial and many researchers conclude that women differ in their preference for competition compared to men. My study explores the potential that another channel may be yielding the observed gender gap in tournament selection: a gender difference in beliefs about future performance.

My findings suggest that the observed difference in behavioral responses to competition among men and women is not due to a difference in preference for competition, but rather a difference in beliefs about future performance, evidence through the response to changing the task gender stereotype. Women are opting out of competitions when the selected task carries a connotation of male dominance, an environment where women would carry lower beliefs about future performance. I find that when the task carries a female stereotype, more women than men choose to enter the tournament. It appears an increase in female confidence and decrease in male confidence is driving this result.

Despite the fact that women earn roughly 57% of awarded bachelor's degrees, women occupy only 21 of the CEO positions for S&P 500 companies. While women are clearly capable of top positions, we are still not observing women seeking out these positions at rate similar to men. This study suggests that this observed difference may result from the gender stereotype of CEO positions. A Google image search of the word "CEO" produces only three images of females in the first 100 results.

The same thinking follows for the lack of women in STEM fields. Research suggests men can outnumber women as much as four to one in computer science classes while in non-STEM courses, women generally outnumber men. As discussed above, math and science are predominately viewed as male fields and given my results, it is again not surprising to observe this disparity. This study suggests that policy interventions regarding gender-stereotyped tasks might indeed be useful and that more research is needed.

Chapter 2

Gender, Emotions, and Tournament Performance in the Laboratory

2.1 Introduction

Many employees face competition within the workplace. Job promotions, awards such as employee of the month, or incentive pay for top performers are a few of the common tournament structures employed in many workplaces.[1] Additionally, employee interaction outside of the workplace is increasingly common. A study recently published by *Millennial Branding* found that individuals between the ages of 18 to 29

[1]Using a survey of 15,000 employed Americans, re-weighted to match the census tracts, Bo Cowgill (2015)[38] finds that roughly 77% of Americans face intra-worker competition that is a significant component of wage determination. Cowgill points out that a significant number of large firms are known to use some form of tournament for promotions. A few of the listed firms include Adobe, AIG, Amazon, American Express, Cisco Systems, Conoco, Dow Chemical, Enron, Expedia, Facebook, Ford, General Electric, GlaxoSmithKline, Goldman Sachs, Goodyear Tire, Google, Hewlett-Packard, IBM, Intel, LendingTree, Lucent, Microsoft, Motorola, Sun Microsystems Valve, and Yahoo.

are friends with an average of 16 co-workers on Facebook.[2] Garcia et al. (2013)[40] cite personal history as a potential influence on competitive behavior and note that situational factors can come into play where comparison and competitiveness are found within what the authors label "Social Category Fault Lines.

I use a controlled laboratory setting to enhance understanding of how emotions towards other individuals affect tournament performance as well as how the behavioral findings may differ across genders. Behavioral responses to emotions, tournament performance, and gender differences have been studied separately but there is little research that combines all three of these variables. My study adds to the body of research in that it aims to assess emotion, tournament performance, and gender differences simultaneously.

While individual emotions can be significantly dynamic and complex outside the laboratory, the anonymity of the laboratory allows me to observe the specific impact of laboratory induced emotions on individual behavior in the presence of competition, and how the impact on behavior differs across genders. Specifically, I test how negative interpersonal experiences affect performance in competitive situations, while controlling for the level of negativity and gender. My study provides a setting in which women are found to respond positively to the presence of a tournament while men are unaffected.

I conducted a two-stage laboratory experiment. In the first stage, a four-person public goods game was used to generate a range of emotions. Subjects observed the individual contribution decisions of all group members and then I asked them to state their feeling about each group member on a scale of one to five, where one indicated strongly negative feelings and five reflected strongly positive feelings. The

[2]Furthermore, *CBS News*[39] noted that "work and life are a mishmash now."

second stage consisted of a one-on-one tournament using a real-effort task where the assigned opponent is a group member from the public goods game. The opponent remained the same for all five rounds of the tournament, and subjects were informed about the selected opponent's identity prior to the start of the tournament. Between tournament rounds, subjects were informed whether their number of completed tasks exceeded the number of completed tasks of their opponent but not how many tasks the opponent completed.

By comparing individual performance in the tournament across the reported feelings, I determined if and how emotions affect tournament performance. Specifically, the experiment investigated if an individual who was matched with an opponent he rated as strongly negative responded with increased performance, compared to an individual matched with a less negatively rated opponent. Additionally, because gender may affect tournament behavior, the study tested how behavior differs between men and women, particularly while controlling for the level of emotion. I found that strongly negative emotions positively affect performance for women but not for men. This observed difference is not due to a difference in the intensity of the generated emotion across genders.

In previous studies, Fehr and Gächter (2000a, 2002)[41][42] show that strong negative emotions bring about a desire to punish free-riding in public goods games and, if given the opportunity, subjects choose to undertake costly punishment a significant portion of the time, even when there is no direct monetary benefit to the subject. In a survey on observed reciprocity in the literature, Fehr and Gächter (2000b)[43] mention numerous other studies also suggesting that negative reciprocity arises out of a desire to punish "hostile intentions" (e.g., Rabin, 1993; Blount, 1995; Dufwenberg and Kirchsteiger, 1999; and Falk and Fischbacher, 1999). In my study, punishment

takes the form of increased effort in a real effort task so as to increase the probability of winning the tournament and therefore the other person losing the tournament.

While emotions can influence individual behavior, the effect of this influence may differ between men and women. Bettencourt and Miller (1996)[44] show through a meta-analysis that while unprovoked males are more aggressive than unprovoked females, this difference is significantly reduced when both genders are provoked. In a review of the literature on gender differences, Croson and Gneezy (2009)[45] note many observed differences between men and women that are consistent across the data; however, the data on gender differences in public good contributions is mixed. They state that some studies show that women are more pro-social in a public goods game than men (e.g., Seguino, Stevens, and Lutz, 1996), whereas other studies find the opposite (e.g, Brown-Kruse and Hummels, 1993; Sell and Wilson, 1991; and Solow and Kirkwood, 2002). The authors note that psychological research indicates that women are more sensitive to social cues and therefore may respond differently depending on the experimental design. In my study, I look at how emotions affect behavioral response to a tournament environment, specifically when these emotions are connected to the other individuals in the tournament.

Combining gender differences and tournament performance, Gneezy, Niederle, and Rustichini (2003)[10] demonstrate that males respond more strongly than females to a tournament environment by increasing effort while female effort remains unchanged or decreases across piece-rate versus tournament environments. Additionally, Niederle and Vesterlund (2007)[1] show that, when given the option of compensation from either a piece-rate scheme or a tournament, males choose the tournament environment significantly more than females. The authors show that even when the tournament may be beneficial for high performing females, women tend to avoid competitive tasks.

Similarly, in a study of elementary school children, Gneezy and Rustichini (2004a)[11] assess the running speed of boys and girls in both competitive and noncompetitive environments. They find no difference in running speed between boys and girls in the noncompetitive environment, but do find a significant increase in the running speed of the boys when presented as a competition, while the running speed of the girls did not change significantly, thus creating a significant gender gap in competition performance. In contrast, a recent study by Cassar, Wordofa, and Zhang (2016)[23] finds that the gender gap in tournament selection is erased when incentives benefit one's child such as workplace daycare as opposed to monetary. This recent study provides evidence that there may not be an actual difference in preference for competition between males and females but rather that in the right environment and with properly aligned incentives, women can be enticed to compete as vigorously as men. Additionally, a difference in beliefs about future performance may be responsible for the observed difference in tournament entry rates across genders rather than a difference in preference for competition. Using a female stereotyped task, I find no difference in tournament entry rates across genders while I do find a difference in tournament entry rates across genders using a male stereotyped task (Halladay, 2016)[24]. This suggests a difference in beliefs about future performance is the channel which differences in tournament entry rates across genders is operating.

To the best of my knowledge, Gneezy and Imas (2014)[46] is the only experiment combining emotion and tournament performance. The authors find that with a strength-based task, anger improves performance. However, their all-male subject pool does not allow for the analysis of gender differences.

The paper proceeds as follows. Section 2 outlines my experimental design while section 3 explores my hypotheses and predictions. I present my results in section 4

and discuss the implications of these results in section 5. Section 6 concludes.

2.2 Experimental Design

The experiment takes place in two stages. The first stage presents a situation that can trigger feelings in the laboratory, both positive and negative, while the second stage is a tournament.[3] Participants do not learn about the nature of the second stage until after the conclusion of the first stage.

In the first stage, experimental subjects are randomly placed into groups of four to play a one-shot public goods game with voluntary contributions. Prior to the allocation stage, subjects learn how to calculate payoffs through a series of examples. Also, I require that subjects successfully calculate payoffs for two hypothetical scenarios on their own before the first stage begins. Subjects start with $7.00, total contributions are multiplied by 1.6, and then distributed equally to all group members. Additionally, subjects are told that all contribution decisions will be revealed to the group with identification by subject ID numbers. After allocation choices, subjects learn their own payoffs and the contributions and payoffs of the other three group members. Subsequently, subjects provide feedback about their feelings regarding the other three group members using a five-point scale. A rating of one indicates strongly negative feelings, and a rating of five indicates strongly positive feelings. A rating of three denotes neutral feelings, neither positive nor negative. I include the neutral feelings rating option in the event that a subject does not feel they can rate a group member. A screen shot of this zTree screen can be found in the appendix.

For the second stage, I match each participant with one other participant from

[3]Previous research provides evidence that the public goods game environment successfully generates emotions (Fehr & Gächter, 2002)[42].

their original group of four for five rounds of a slider task using a tournament payment scheme (Gill & Prowse, 2009)[47]. Participants are informed about the subject ID (e.g., Person 1) of the selected individual on the instructions screen for the second stage. I followed essentially the same procedures as Charness and Villeval (2009, *AER*)[48]. This design is also similar to many studies that use a matching groups protocol (e.g., Charness, 2000[49]; Greiner and Vittoria Levati, 2005[50]; Charness, Fréchette, and Qin, 2007[51]).

After reading the instructions, but before beginning the tournament, I remind subjects of the results of the first stage and the ratings they assigned to each group member. In each of the five rounds, participants have 90 seconds to complete as many slider tasks as possible. A slider bar is complete if the subject slides the marker exactly to the halfway position (50). Initially, subjects see a screen displaying 48 slider bars. If the subject completes all 48 slider bars within the 90 seconds, I give subjects an additional 48 slider bars to complete to ensure performance is not constrained. As the round progresses, the screen displays how many sliders the subject has successfully completed. A screen shot depicting this zTree screen can be found in the appendix. A participant wins the tournament round if he completed more slider tasks than his opponent. At the end of each round, each participant learns whether he won the tournament but not how many tasks the matched person completed. Participants were told that one of the five tournament rounds would be randomly selected for payment in addition to the public goods game payoff. If the participant won the selected round, he receives $2.50. Total payment consisted of the payoff from the first stage and the result of the tournament. Subjects earned $8.36 on average from the public goods game.

My design follows a between-subjects design. I conducted three treatments: *posi-*

tive feelings (PF), negative feelings (NF), and median feelings (MF). In the *PF* treatment, individuals were matched with the group member they rated most positively, while in the *NF* treatment, individuals were matched with the group member they rated most negatively. In the *MF* treatment, individuals were matched with the group member they rated intermediately compared to the other two group members.[4] Treatments did not vary within sessions.

This experiment was programmed in z-Tree (Fischbacher, 2007)[33]. All sessions took place in the Experimental and Behavioral Economics Laboratory at the University of California, Santa Barbara. I used the Universitys ORSEE system to recruit subjects, and all were current students. A total of 180 students participated. I ran a total of 12 sessions with each session having either 12 or 16 subjects (three or four groups). 64 subjects participated in the PF treatment, 72 participated in the NF treatment, and 44 participated in the MF treatment. No subject participated in more than one session or more than one treatment. Average earnings were $9.84 and each session lasted approximately 45 minutes. A set of instructions can be found in the Appendix.

2.3 Hypotheses

This experiment addresses three main hypotheses about how individuals behave in a competitive environment with someone with whom they have some recent experience. Rational economic theory suggests that effort in the competition should be unaffected by the results of the first stage unless first stage contributions to the pub-

[4]For example, I rate my three group members 1, 4, and 5. In the PF treatment, I would be matched with the individual I rated a 5. In the NF treatment, I would be matched with the individual I rated a 1. In the MF treatment, I would be matched with the individual I rated a 4. If two individuals received the same rating, one was randomly assigned as the opponent.

lic goods game are informative about the opponent's second stage performance/effort choice. If, for example, individuals who are viewed as strongly negative are also more likely to exert higher levels of effort, opponents might respond to the higher expected effort levels by also increasing their own effort.

Will the opportunity to compete with an individual who has generated negative feelings for a subject lead to increased performance? This kind of behavior would be evidence of negative reciprocity. Previous studies show that in public goods games, individuals will punish free riders if given the chance, even when punishment is costly (Fehr & Gächter, 2000a)[41]. Though this experiment does not allow for direct punishment, subjects can increase performance as a means of punishing a non-cooperator. While previous work has focused on punishment through reducing one's own payoff, another form of punishment is increasing one's probability of winning a tournament and thus reducing the payoff of the other individual through increased real effort. In another experiment allowing for sanctions, Fehr and Ficshbacher (2004)[52] find that negative emotions drive sanctioning decisions that promote more pro-social behavior. Additionally, Kahneman, Knetsch, and Thaler (1986)[53] find evidence of punishment through indirect reciprocity when subjects chose to forgo a larger payoff for themselves to punish an individual who had previously acted unfairly. As seen in Garcia et al. (2013)[40], the history between two parties can significantly affect competitive behavior. The first stage of this experiment may indirectly draw social category fault lines dividing cooperators and free-riders. It is plausible to consider that given an individuals negative feelings, these emotions may 'light a fire', so that he may seek to 'let off steam', or desire to reestablish dominance, all resulting in increased performance. However, I will be unable to distinguish if this observed increase in performance arises in part from being previously hurt financially by someone, or in

part from being matched with someone who harms you by violating a social norm. I present this explicitly now:

> **Hypothesis 1: The number of completed slider tasks will be *higher* for individuals if and only if they are matched in the tournament with someone about whom they reported strongly negative feelings.**

Previous research demonstrates that it is much easier to find evidence of negative reciprocity than evidence of positive reciprocity. The lack of evidence of positive reciprocity in their data led Charness and Rabin (2002)[54] to not even include positive reciprocity as an explanation for behavior in their model. In an experimental labor market, Charness (2004)[55] finds strong evidence of negative reciprocity when employer-assigned wages are low, but no significant evidence of positive reciprocity when employer-assigned wages are high. Additionally, Offerman (2002)[56] provides evidence that individuals respond more strongly to negative intentions as opposed to positive intentions as the result of a self-serving bias. Individuals tend to view positive outcomes as a positive reflection of themselves while they tend to view negative outcomes as a negative reflection of others. Fehr & Gächter (2000a)[41] demonstrate through punishment in a public goods game that "...there is a large drop in punishments if an individuals contribution is close to the averageThus, the more an individuals contribution falls short of the average the more she gets punished." If low contributions are viewed more negatively confirming Hypothesis 1, consistent with Fehr & Gächter, I hypothesize that I should only observe the increase in performance for individuals competing with someone toward whom they have strongly negative feelings. Succinctly:

Hypothesis 2: There will be no difference in the the number of completed slider tasks across all other reported feelings other than strongly negative feelings.

Gneezy and Rustichini (2004)[11], find no gender differences in speed when children run alone but do find that boys outperform girls when running in mixed-gender pairs providing further evidence that males tend to be more responsive to competitive environments. Additionally, Buser and Dreber (2014)[57] show that under a piece-rate payment scheme, men significantly outperform women in the slider task. Therefore, with the combination of the male dominant task and the competition driven male performance, I hypothesize that men will outperform women independent of emotion. To summarize:

Hypothesis 3: The number of completed slider tasks will be higher for men than for women, holding reported emotion constant.

Lastly, previous research yields inconsistent evidence regarding which gender will be most affected by the emotions. Eckel and Grossman (2005)[58] find that women are more likely to punish unfair behavior than men. In contrast, Christensen et al. (1983)[59] find evidence that in romantic relationships, males exhibit a stronger self-serving bias, and they may be more apt to engage in negative reciprocity. Hence, there is no clear prediction about behavior in this respect.

2.4 Results

My subject pool was 43.64% male. Subjects contributed an average of $2.25 (32% of the endowment) during the public goods game. The average rating of feelings

towards the *matched* opponent over all treatments was 2.9 and the average assigned rating across *all* group members was 2.89.[5] Table 1 presents the breakdown of opponent ratings by treatment while Table 2 presents the breakdown of opponent ratings by gender. Average tournament performance was 25.07 tasks across all five rounds.

Table 2.1: Opponent Rating by Treatment

	Opponent Rating					
Treatment	1	2	3	4	5	Overall
Negative Feelings	27	19	22	3	1	72
Middle Feelings	4	6	16	13	3	42
Positive Feelings	4	4	19	15	22	64
Overall	29	57	31	26	178	

Table 2.2: Opponent Rating by Gender

	Opponent Rating					
Gender	1	2	3	4	5	Overall
Female	19	17	31	18	15	100
Male	16	12	26	13	11	78

Finding 1: Higher (lower) contributions do lead to higher (lower) ratings.

My data confirm that individuals view low contributions negatively. There is a clear positive relationship between contribution and average rating assigned as shown in Figure 1. A test of correlation between contribution and rating assignment affirms the relationship that subjects negatively view low contributions ($\rho = 0.5435$, $p<0.0000$). Using Figure 2, I check for average contribution given a rating across

[5]This slight difference is due to unequal treatment sizes. I had 72 subjects in the negative feelings treatment, 64 subjects in the positive feelings treatment, and 42 subjects in the middle feelings treatment. My goal was to have the same number of subjects in the negative feelings treatment as the positive feelings treatment, but based on no-show subjects and the need for groups of 4 for the public goods game, I ended up having 2 extra groups in the negative feelings treatment compared to the positive feelings treatment. Because changes in behavior tend to observable at the extremes, I planned on less subjects in the middle feelings treatment.

genders to determine if there is a gender difference when assigning ratings based on contribution levels. Confirmed by pair-wise t-tests, there is no evidence that women and men have different thresholds for assigning a specific rating. Figure 3 suggests that contributions that fall below the group average are viewed negatively, while contributions matching or exceed the group average are viewed positively, further evidence that lower contributions lead to lower ratings.

Figure 4 illustrates that there is also no gender difference in contributions to the first stage public goods game. A Kolmogorov-Smirnov test used to detect a difference in the distribution of public goods game contributions for men and women is not significant (p=0.147) and therefore there is no evidence that there exists a gender difference in the distribution of contributions. Also, a test of the equality of means for average public goods game contributions can also not be rejected (p=0.170) such that there is no evidence that average contributions diff across genders.

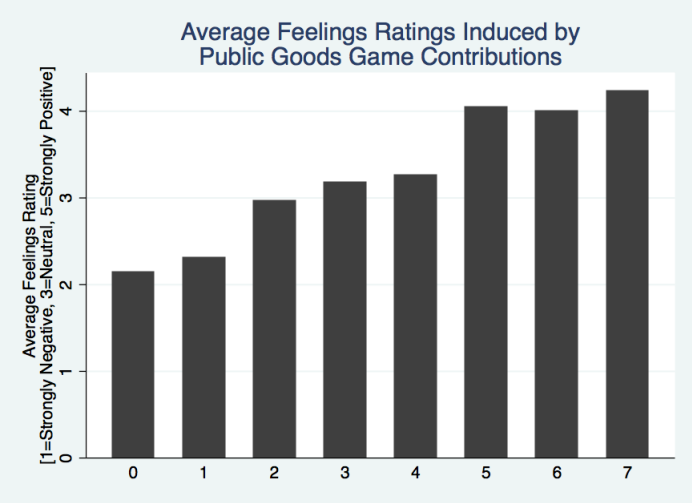

Figure 2.1: Persons who contributed more in the public goods game generated more positive emotions in others.

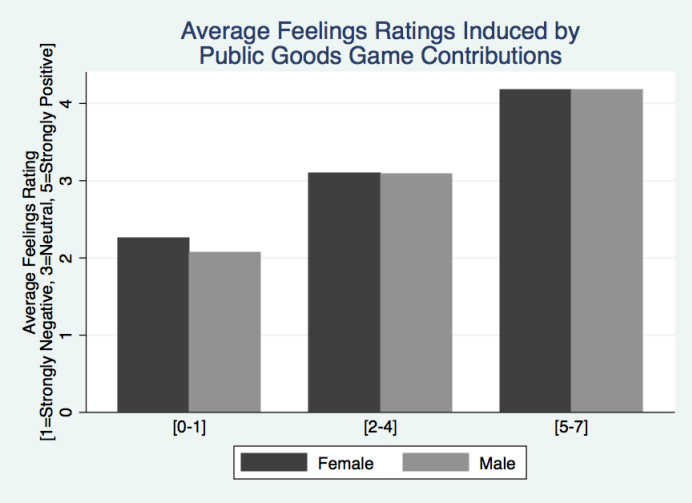

Figure 2.2: Persons who contributed more in the public goods game generated more positive emotions in both males and females. There is no difference in ratings assigned by males and females conditional on opponent contribution to the public goods game.

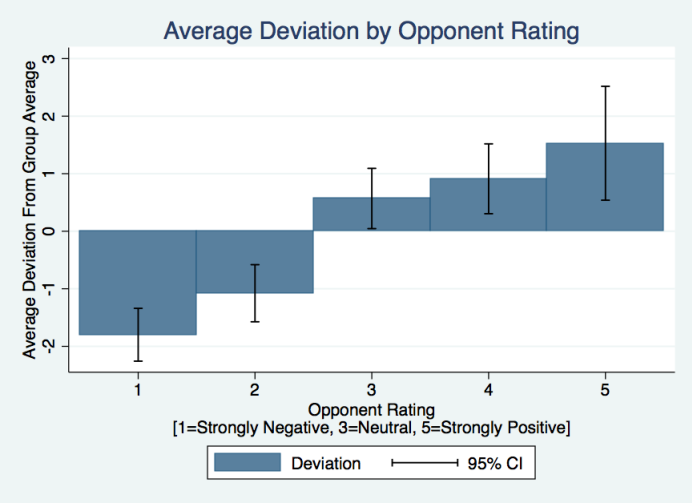

Figure 2.3: Negative deviations from the average group contribution generated negative emotions while positive deviations from the average group contribution generated neutral or positive emotions

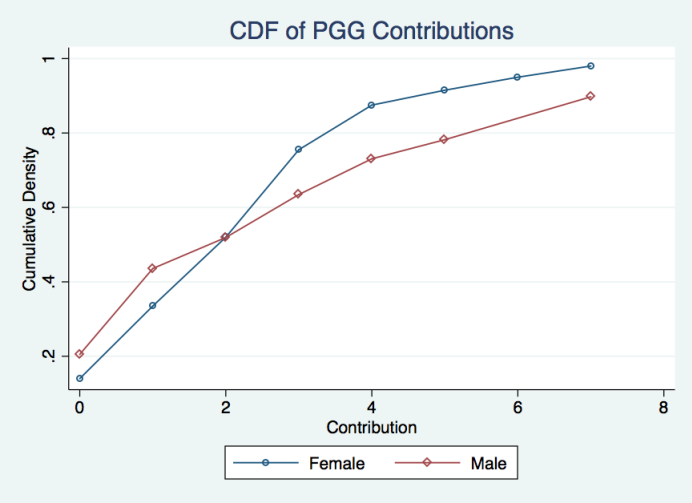

Figure 2.4: There is no difference in average male and female public goods game contributions nor is there a difference in the distribution of male and female public goods game contributions.

Because subjects receive feedback between tournament rounds, the forthcoming analysis will proceed using only data from the first period in order to avoid learning and feedback effects. For example, Gneezy and Rustichini (2004)[11] demonstrate that in the presence of relative performance feedback, the performance of boys increased while the performance of girls was unaffected. Using International Tennis Federation data, David Wozniak (2012)[60] finds that while males are influenced by performance over many periods and their behavior seems to reflect a belief in a "hot hand", females are influenced by their most recent tournament performance. Additionally, previous work in psychology has shown that women view their success as the result of good luck while men view their success as the result of their own ability. Further, research as shown that emotional responses tend to die out over time, and therefore the all period analysis may fail to capture to behavioral responses tied to the reported emotions. Grimm and Mengel (2011)[61] find that low ultimatum game offers are accepted 60-80% of the time when subjects are given a ten minute delay prior to the rejection

decision, whereas these low offers are only accepted 20% of the time without a delay. An additional concern about the confounds of the all period analysis is that it will be impossible to disentangle the behavioral responses due to the first stage induced emotions and the behavioral responses due to the feedback induced emotions. Not only will the emotions from the first stage diminish, but it will be unclear when and to what extent the feedback emotions take over. While the results using only period one and all five periods are relatively consistent, it is clear that the results from all five rounds are potentially confounded due to the between round feedback and the time elapsed. The analysis using all five rounds is available in the appendix.

While Table 3 illustrates there is an effect when running the regression analysis using the randomly assigned treatment, the effects are weaker because, as shown in Table 1, there are still subjects matched in the tournament with an individual assigned the strongly negative rating despite being in the *MF* or *PF* treatments. Therefore, using the treatment variable is a much noisier, however still significant, signal of reported feelings.[6] The average opponent rating in the *NF* treatment was 2.06, 3.12 in the *MF* treatment, and 3.73 in the *PF* treatment. All of three of these pair-wise tests of the equality of means are significant ($p \leq 0.008$). It does appear that the treatment captured a difference in opponent ratings but is clearly a noisier signal. The remaining analysis will group data into bins by assigned opponent rating as opposed to treatment.

Finding 2: Performance is higher for individuals competing with someone rated as strongly negative.

[6]Suppose a subject is in a group with individuals who all contribute their entire endowment. This subject is most likely going to report positive feelings for all group members. Therefore, even if this subject is in the negative feelings treatment, they did not experience negative feelings and should therefore not be grouped with other individuals who did.

Table 2.3: OLS Regression Results

	Completed Slider Tasks
Subject Contribution	0.832*
	(0.425)
Male	4.302**
	(1.732)
Negative Feelings Treatment	1.167**
	(0.410)
Positive Feelings Treatment	-2.070**
	(0.780)
Constant	6.958***
	(1.880)
R^2	0.1284
N	178
Session Dummies	Yes
Demographics	Yes

*p<0.10, **p<0.05. P-values are two-sided. Standard errors in parentheses. Standard errors are clustered by session.

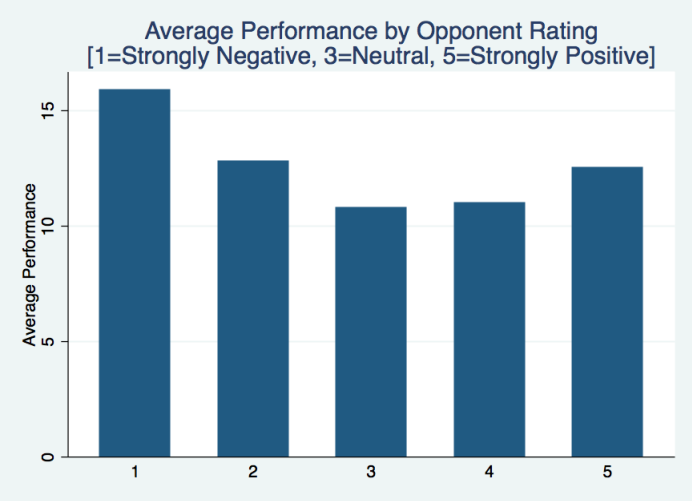

Figure 2.5: Average performance increases significantly when one strongly dislikes their opponent (assigned a rating of one)

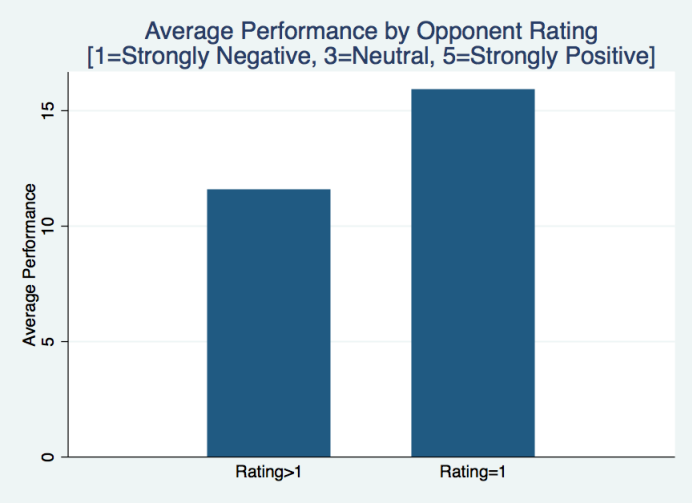

Figure 2.6: Average performance increases significantly when one strongly dislikes their opponent (assigned a rating of one)

Not controlling for gender, average performance is higher for individuals whose opponents received ratings of one (strongly negative feelings) compared to ratings two through five. For the individuals matched with an opponent rated one, average performance was 15.91. For opponents rated two through five, average performance was 12.79, 10.79, 11.00, and 12.54, respectively. This sharp increase in performance is illustrated in Figure 5, and a two-sided t-test comparing the difference in performance for subjects matched with an opponent who received a rating of one versus grouping subjects matched with opponents who received ratings greater than one (Figure 6) is significant (p=0.0169).[7] I have clear evidence that strongly negative personal history significantly increases competitive behavior.

These results are in line with Fehr and Gächter (2000a)[41] where punishment and negative emotions both intensify the larger the negative deviation from the group average. In Figure 3, I demonstrated that ratings of one and two pertained to

[7]Regressing performance on opponent rating, I find a significant and negative trend using a one-sided test (p=0.041), further demonstrating the increase in performance when competing with an opponent one rated as strongly negative.

below-average contributions while ratings three through five reflected above-average contributions. Keeping with Fehr and Gächter (2000a)[41], I would expect to see punishment when the contributions are below average, but not necessarily when the contribution is close to the group average. The average group deviation for opponents rated strongly negative (a rating of one) was -$1.80 and the deviation for opponents rated somewhat negative (a rating of two) was -$1.08. Both of these values are significantly different from zero and negative, however, in terms of potential contribution values, because subjects were restricted to contributing whole numbers, a deviation of one from the group average does not seem to indicate a significant deviation. Therefore similar to Fehr and Gächter (2000a)[41], I would not expect to see an increase in effort for individuals matched with subjects rated somewhat negative. However, it seems plausible that because the range of potential contributions was only [0,7] inclusive, a deviation of roughly two from the group average would be viewed much more significantly. As with Fehr and Gächter (2002)[42], this is where I would expect to see the increase in punishment. These results support Hypothesis 1 that performance will increase when competing with someone about whom you feel strongly negative.

Finding 3: Performance only increases significantly when the behavior of the opponent is particularly flagrant.

Using a KruskalWallis test for a difference in median performance among individuals competing with subjects rated two through five, I find that there is no difference in effort among these four groups (p=0.4168). Additionally, a Kruskal-Wallis test for a difference in median performance across all ratings is marginally significant, suggesting at least one of the medians differs (p=0.0813). It must be then that this difference lies in the median performance of individuals competing with an opponent viewed strongly negative. One-sided pairwise t-tests tests produce relatively consis-

tent results with the mean performance of individuals competing with a subject rated one being significantly different than subjects rated three, four or five (p=0.0115, p=0.0277, p=0.0891, respectively). The p-value on the one-sided t-test comparing average performance of individuals with opponents rated one and opponents rated two is very close to marginally significant at 0.1031. All t-tests comparing opponents of ratings two through five could not reject the null hypothesis of no difference in average performance. Behavior is unaffected when the emotions involved are not strongly negative, confirming Hypothesis 2.

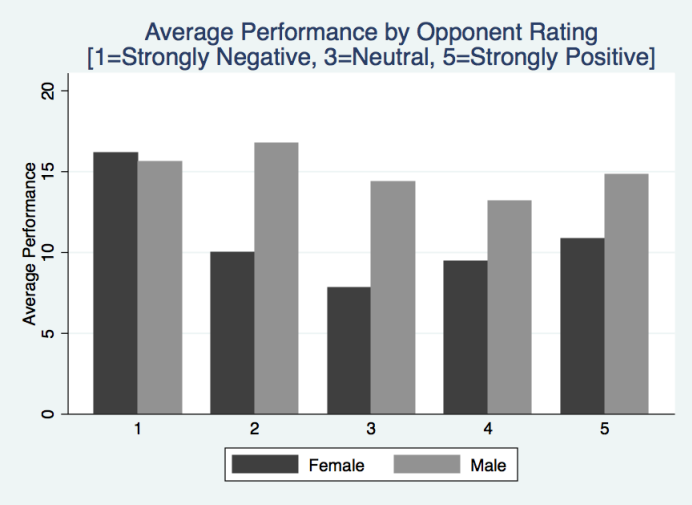

Figure 2.7: Average performance increases significantly when one strongly dislikes their opponent (assigned a rating of one)

Finding 4: Male performance is higher than female performance.

Looking at average effort by gender, my data supports the hypothesis that men compete more and have higher performance in the tournament. Average male performance is 14.85 tasks while average female performance is 10.52 tasks. This difference is significant using a one-sided t-test (p=0.0015). The creators of the slider task, Gill and Prowse, provide evidence that within their subject pool, male and female be-

havior was not significantly different. By the final round, men completed 25.75 tasks on average while women complete 26.83 tasks on average. Figure 7 corroborates this finding, illustrating that male performance is above the performance of females for every reported emotion other than strongly negative emotions.

Finding 5: Women respond more strongly to negative emotions.

Figure 8 allows a comparison in performance for opponent ratings of one versus ratings above one by gender. The t-tests in Table 4 illustrate five results. All p-values presented in Table 4 are two-sided unless otherwise specified. Male performance rises by 0.98 tasks (p=0.7430) on average when the emotions are strongly negative while female performance rises by 6.96 tasks (p=0.0012) on average in response to the strongly negative emotions. In the absence of strongly negative emotions, males perform 5.45 tasks (p=0.0004) more than females, however, with strongly negative emotions, this difference falls to 0.53 tasks (p=0.8850). All of these results hold as well using a Fisher's exact test on the difference in medians. The gender gap in competition performance is eliminated in the presence of this strongly negative emotional stimulus. Women seem to be responding to the emotion more significantly than men. This notion is confirmed by the difference-in-differences estimate provided in the regression results of column 4 on Table 5 (p=0.098). The observed increase in performance when competing with a negatively viewed opponent appears to be purely driven by the female response as male performance is unaffected. Though men are more competitive across the board, negative emotions appear to evoke a "competitive fire in women. A similar analysis comparing performance across subjects who were "badly wronged" in the public goods game can be found in the appendix.

Lastly, it is worth noting that aggregate performance is significantly higher for pairs with one member viewed strongly negative. In pairs with the opponent assigned

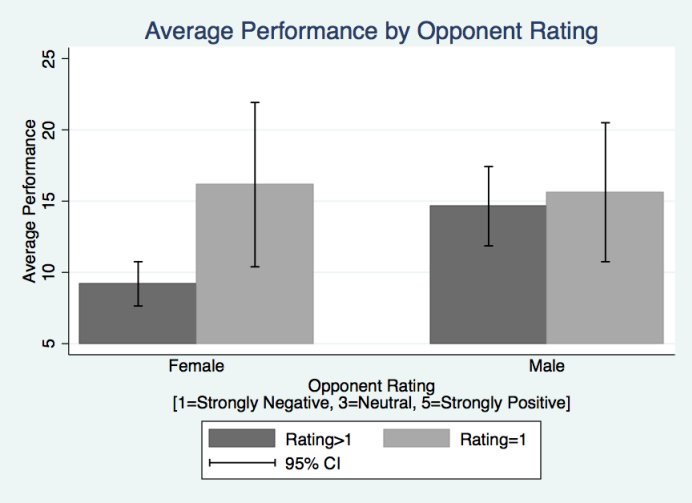

Figure 2.8: The increase in female performance in the presence of the strongly negative emotions is driving the overall observed performance increase.

Table 2.4: Difference in Means

Group 1	Group 1 Mean Perf.	Group 2	Group 2 Mean Perf.	Difference in Means (p-value)	Difference in Medians (p-value)
All Subjects (Rating>1)	11.56	All Subjects (Rating=1)	15.91	4.35**	4.00**
				(0.0169)	(0.023)
Males (Rating>1)	14.65	Males (Rating=1)	15.63	0.98	2.00
				(0.7430)	(0.401)
Females (Rating>1)	9.20	Females (Rating=1)	16.16	6.96***	4.00*
				(0.0012)	(0.073)
Females (Rating=1)	16.16	Males (Rating=1)	15.63	0.53	1.00
				(0.8850)	(1.00)
Females (Rating>1)	9.20	Males (Rating>1)	14.65	5.45***	3.00***
				(0.0004)	(0.004)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

a rating of one, aggregate performance is 4.4 tasks higher than in pairs without. This difference is significant using a one-sided t-test (p=0.0423).

Regressions in Table 5 support the above results. All regressions included session dummy variables and demographic variables including whether a subject had been in an economics experiment previously, if the subject was an economics/accounting major, and standard errors were clustered on the session level. By including both session dummy variables and clustering on the session level I am essentially producing the first stage of the Donald & Lang (2007)[62] two-step correction in the presence of a small number of clusters. I do not need to worry about the second stage because

the second stage yields coefficient estimates for coefficients of interest that are session specific of which I have none.

Table 2.5: OLS Regression Results

	(1) Completed Slider Tasks	(2) Completed Slider Tasks	(3) Completed Slider Tasks	(4) Completed Slider Tasks
Subject Contribution	0.679*	0.607	0.647	0.621
	(0.357)	(0.370)	(0.386)	(0.375)
Male	4.353**	4.230**	4.234**	5.461**
	(1.778)	(1.833)	(1.799)	(1.614)
Negative Feeling Dummy	2.072			
	(1.853)			
Positive Feeling Dummy	1.419			
	(1.208)			
Strongly Negative		3.828	3.282	5.977*
		(2.154)	(1.928)	(3.207)
Somewhat Negative		0.528		
		(1.979)		
Somewhat Positive		0.801		
		(2.010)		
Strongly Positive		2.263		
		(2.251)		
Strongly Negative*Male				-5.940*
				(3.289)
Constant	6.968**	6.926**	7.352**	7.051**
	(2.852)	(2.878)	(2.175)	(2.196)
R^2	0.1347	0.1456	0.1413	0.1551
N	178	178	178	178
Session Dummies	Yes	Yes	Yes	Yes
Demographics	Yes	Yes	Yes	Yes

*p<0.10, **p<0.05. P-values are two-sided. Standard errors in parentheses. Standard errors are clustered by session.

In model (1), I created a dummy variable for opponents rated one or two (strongly negative and somewhat negative, respectively), and another dummy variable for opponents rated four or five (somewhat positive and strongly positive, respectively). I left those with an opponent rated three (neither positive nor negative feelings) as the reference group. Neither of these coefficients were significant which is not surprising given the graphical evidence above demonstrating that this performance increase is found only for those individuals competing with someone rated strongly negative.

I used model (2) to include dummy variables for all the rating categories, leaving the neither positive nor negative category as the reference group. The strongly negative category was the only close to significant coefficient (p=0.103) when compared to the reference group using a two-sided hypothesis test, and is significant at the 10% level using a one-sided test. Subjects competing with a strongly disliked (rating of one) individual completed roughly four additional tasks compared to subjects competing with an individual with a rating of three.

This confirms the pattern in Figure 5 where the only spike in performance was evident with individuals whose opponents were viewed as strongly negative. The positive coefficient for strongly negative feelings is marginally insignificant using a two-sided test (p=0.117), however with a one-sided hypothesis test like the one of interest, I would obtain a significant coefficient at the 10% level. It appears that something is masking the observed increase in performance such as a gender difference driven by an increase in female performance. A simple difference-in-differences should clear this up.

I included an interaction term between the dummy variable for male and the dummy variable for strongly negative feelings to test if the reaction to emotional stimulus was different for men and women. The negative and significant coefficient

provides evidence that given an individual is competing with someone viewed as strongly negative, women increase performance by roughly six tasks more than men in the presence of the emotional stimulus. This result corroborates Figure 8 and the t-tests presented in Table 4 that females are driving the increase in performance and are much more responsive to the negative emotions. Interestingly, as shown in Figure 8 as well, male performance in unaffected by the negative emotional stimulus. This can be demonstrated by the summation of the coefficients on the strongly negative dummy variable and the new interaction variable. These two coefficients essentially cancel each other out, illustrating that males' performance is not altered by the presence of these negative emotions. Using a Wald test for the summation of these two coefficients, the hypothesis that the coefficients on the strongly negative dummy and the interaction of the strongly negative dummy with the gender dummy cancel each other out cannot be rejected (p=0.9802).

2.5 Discussion

My data confirm that when competing with a person towards which one has strongly negative emotions, individual performance increases substantially . Specifically, an increase in female performance drives the observed overall performance increase. This suggests that individuals use the tournament environment as an opportunity to retaliate with matched individuals who they previously experienced negatively. Individuals are willing to undertake costly effort in order to increase their own performance as a means of increasing the probability of a win, suggesting the presence of negative reciprocity.

One possible explanation for the increase in demonstrated performance may be a rational response of individuals who expect negatively rated opponents to compete

more vigorously. Average opponent effort by ratings one through five was 12.23, 11.83, 10.54, 14.9, and 12.92, respectively. There is no difference in the distribution of opponent performance levels by opponent rating using a Kruskal-Wallis test (p=0.6507), therefore there does not seem to be evidence that the increase in effort is driven by a rational response to work harder with a poorly rated individual. Additionally, using an ANOVA, we cannot reject that there is no difference in average opponent performance across opponent ratings (p=0.4715). While the figure of interest is the performance of individuals rated one, both tests show that there is no evidence that the effort of opponents rated one (strongly negative) is significantly higher than opponents of any other rating (two through five). Because effort levels do not differ by opponent rating, this suggests that the emotional channel is driving the increase in effort.

On the other hand, the increase in performance could be driven by something unique about subjects willing to rate another individual as strongly negative. While 39% of subjects assigned at least one other participant a "strongly negative" rating, not all of these subjects were matched with the specific individual assigned the low rating. Because I observe these subjects across all opponent ratings but do not observe the increase in performance across all opponent ratings, my results do not appear due to something special about subjects who give ratings of strongly negative. Additionally, a test of a correlation between an indicator for whether subjects assigned at least one strongly negative rating and the subject's own effort is significant ($\rho = 0.155$, p=value=0.0385). However, when I restrict this test to only subjects who never competed with the individual assigned the strongly negative rating, the significance disappears ($\rho = 0.066$, p-value=0.4370). If my results were driven by a concern that there is something unique about individuals willing to assign low ratings, this

correlation test would have remained significant even for individuals not competing with the subject with the low rating.

Another possible explanation is that individuals with low profit from the first stage may increase effort in order to increase the probability of winning the tournament thus increasing total payoff. There is no difference in average first stage profit between individuals competing with a strongly negative opponent (rating=1) and those competing with a somewhat negative opponent (rating=2) (p=0.9570). Consequently, if effort were driven by a concern for profit, one would expect to observe an increase in effort in both of these two groups. However, as showed above, the increase in effort was only evident for the individuals competing with a strongly negative opponent. Therefore, there is no evidence that the increase in effort is driven by a concern for increasing ones profit, again, providing evidence that emotions are driving the result.

My difference-in-differences estimate is robust to multiple specifications. Including opponent's public goods game contribution as a control does little to my difference-in-differences estimate and it remains significant (0.093), even increasing slightly in significance. Controlling for opponent's public goods game contributions allows me to say that the reported emotions have predictive power over and above parodying for opponent contribution. Replacing subject contribution with subject first stage profit and including opponent's public goods game contribution as a control, again does little to my difference-in-differences estimate, again, slightly increasing it's significance (p=0.087). By including subject's first stage profit, I can be more sure that income effects are not driving my results.

Regardless of the behavioral motivation, my results strongly support that the motivation is largely extrinsic due to the significant increase in performance when competing with an individual rated strongly negative. Though there may be some

component of effort that is intrinsic in nature, if the motivation was purely intrinsic such as if individuals viewed the slider task as fun, I would not expect to observe a significant difference in behavior among individuals or a significant treatment effect.

Researchers may argue that a four-person public goods game may make it difficult to establish feelings based on direct intentions. For example, a player may have contributed zero, but the other group members rate this person positively because the choice was seen as the smart thing to do. However, given my results and the correlation between contribution and rating, that possibility does not seem to have affected my results. Alternatively, the other group members may rate this person negatively because they are envious of not making the same decision rather than angry. I cannot differentiate between envious and angry individuals beyond that both emotions have a negative connotation. Eliciting intensities of specific emotions with both positive and negative affect would be an interesting extension.

2.6 Conclusion

Competitive environments are found throughout our everyday lives, and are present in many workplace settings. Examples include "Employee of the Month rewards or a ranking system that sets a company hierarchy used for determining pay raises. My experiment examines if and how emotions play a role in competitive exchanges.

My findings indicate that performance increases substantially when an individual competes with a strongly disliked group member, as the mean performance in the case of strongly negative reported feelings s higher than all pairings. I have also shown that women are more responsive to the presence of negative feelings in a competitive setting. Given that women are generally more averse to competition, workplace programs that take into account interpersonal relationships may be able

to counteract this demonstrated aversion towards competition.

Future work on this topic would benefit from determining the true nature of the generated feelings such as anger, envy, remorse, etc. to help underpin what is truly driving the increase in performance including whether this increase in arising from being wronged financially or from observing an individual violate a social norm. It would also be of interest to determine whether this increase would exist if the violation of the social norm or financial wronging was realized indirectly. Also, observing a closing of the gender gap in this unique environment further emphasizes the need for additional work analyzing other environments that may also close this gap.

I have shown that negative emotions can increase effort in a tournament environment. Additionally, I have shown that the increase in effort cannot be driven by a rational increase in effort due to individuals expecting to encounter increased opponent effort. Also, women increase their work effort in response to negative emotions and actually embrace the tournament environment suggesting that under the right circumstances, women an be induced to compete as hard as men.

While it seems intuitive that a positive emotional environment is good for performance, my results in this paper provide an example where the opposite is true: negative feelings towards co-workers raise performance. Combining this tendency to respond with increased effort in the presence of negative emotions and the result that women embrace competition, employers need not worry as much about the potential for soured relationships outside of the workplace negatively impacting productivity inside the workplace.

Chapter 3

Experimental methods: Pay one or pay all

3.1 Introduction

What distinguishes laboratory experiments in economics from most experiments in other social sciences is the use of financial incentives. The working assumption is that providing incentives for choices and outcomes leads to more meaningful and reliable choice behavior. Specifically, incentives that encourage subjects to make honest and non-arbitrary choices will accurately reveal characteristics of their preferences.

The simplest and cleanest approach is to have only a single experimental choice. Cross-task contamination is not an issue with single-choice experiments and there is no way for subjects to hedge decisions. However, there are also a number of limitations to this approach. While the incentive structure in single-choice experiments is extremely salient to the subjects, this type of experiment only allows for between-subject comparisons, requiring a significant amount of observations and relatively high cost. In addition, it is impossible to study learning or to examine more set-

tled behavior in relation to equilibrium predictions in an experiment involving some complexity. Thus, it is common in experiments for participants to make multiple decisions or one decision for each of a number of periods. This approach is useful for gathering a considerable amount of incentivized data over the course of a compact session, allowing the researcher to understand individuals fixed effects and to observe learning.

The traditional and conservative approach in experiments with multiple decisions is to pay for the outcome from every decision made ("pay all). An alternative approach is to pay for only one of the choices made, drawn at random at the end of the session ("pay one). When using pay one, the amount at stake per choice is multiplied to compensate for the decreased likelihood of that choices outcome being drawn for payoff. A related approach to pay one is to pay for a selected number of periods rather than just one. Doing so helps to even out the earnings made by participants, which can be useful in terms of subject-pool maintenance. For example, one might pay for one period of every 10, so that five periods are chosen for payment in a 50-period session.

In terms of current practice, none of the 15 experiments with multiple decisions reported in the top-five economics journals in 2011 used pay one, and it appears that, overall, pay one is used less than half as frequently as pay all (Azrieli, Chambers, and Healy, 2015)[63]. Given this current practice, we consider pay all to be the baseline approach in this paper. Yet, while it appears that the majority of well-published multi-period experiments are implementing pay all, some recent research suggests a theoretical advantage of pay one (Azrieli et al, 2015)[63]. In this paper we discuss the two approaches, present the challenges and benefits of each, and discuss how the relative advantage changes depending on the particular laboratory setting.

Experimenters must weigh a variety of considerations when choosing the incentive structure of their projects.

Paying for all decisions makes wealth and portfolio effects possible, as well as cross-task contamination. In experiments where losses are possible, this approach also runs the risk of having bankruptcy issues distort incentives: If a participant knows that she has a negative aggregate payoff during a session and knows that losses cannot be enforced, it is in her interest to take risks in the hope of attaining positive payoffs. In multi-period experiments where bankruptcy issues are feasible, paying for a period or periods determined after the experiment ameliorates or eliminates this issue.[1]

Consider instead a multi-decision experiment where subjects earn money for a single randomly-selected decision.[2] This incentive structure eliminates the opportunity for wealth and portfolio effects and eliminates hedging opportunities (Bardsley, Cubitt, Loomes, Moffatt, Starmer and Sugden, 2009)[66]. If one holds the expected payment for a decision the same across pay one and pay all, this means that the nominal payment will be correspondingly larger (e.g., 10 times as large if one of 10 decisions will be paid).[3]

Another form of this question is whether paying only a subset of the people (rather

[1]For example, Charness and Levin (2009)[64] examine the winners curse (in the Acquire-a-Company task) in 60-period sessions. Bidding positive amounts can generate losses even larger than the endowment given. To mitigate the bankruptcy issue, smooth the overall individual payoffs to some extent, and maintain the subjects attention throughout a session, one period from each 10-period range was chosen for actual payoff at the end of the session.

[2]Cubitt, Starmer, and Sugden (2001)[65] note that if randomization is involved, it is imperative that subjects understand the process by which this is done. A related potential concern with pay one is the introduction of background risk, discussed in section 3.

[3]It is an open empirical question if and when having a smaller-but-certain nominal payoff leads to more attentiveness than a larger-but-less likely nominal payoff. Anecdotal evidence from Cabrales, Charness, and Corchn (2003)[67] shows that in a task requiring considerable cognitive resources, when the nominal payoff was 50, 100, or 150 and each period was paid, some people made arbitrary responses (i.e., choosing the same color in all 10 periods, despite different circumstances). However, when the nominal payoff was instead 500, 1000, or 1500 (with only a 10% chance of that period being selected for payment), no such behavior was observed. This suggests that the higher nominal payoff is more salient than the reduction in the likelihood of payment for a particular period.

than choices) in an experiment induces different choice behavior than paying every participant. Paying only a subset can help with logistical considerations, for example with online experiments and experiments conducted in large classes, since far fewer monetary transactions need be made in this case. Can this effort-saving approach be used without distorting choices compared to those made with a pay-every-person approach?

In this article we explore the issues involving each of these incentive structures. We present evidence for and against each incentive option in diverse laboratory settings, and mention some recent and related ideas. We aim at providing a reference and guide for experimental economists. We present evidence regarding pay one and pay all in section 2 and present evidence regarding paying all subjects or a subset in section 3. We conclude in section 4.

3.2 Pay one or pay all: Evidence

The theoretical literature regarding pay one or pay all in experiments delivers predictions that depend on assumptions about a specific utility theory and the environment. See for example Holt (1986)[68] and Karni and Safra (1987)[69] regarding risk and expected utility, Chandrasekhar and Xandri (2014)[70] regarding stochastic termination, and Baillon, Halevy, and Li (2015)[71] who show theoretically that the order of the randomization and the event is important in ambiguity experiments. In a recent paper, Azrieli et al. (2015)[63] conclude that in a rather general environments that include strategic choices paying for only one period (and in fact, not having a show-up fee paid) is the best mechanism.

The general conclusion from this theoretical literature is that the benchmark prediction is sensitive to the assumptions made. Hence, for theoretically-motivated ex-

periments, the incentive structure should be part of the benchmark prediction of the model, and the equivalence between the two payments should not be assumed. However, the focus of this paper is not on the theoretical aspect of the comparison between pay all and pay one, but rather on practical considerations in the laboratory. We now proceed to empirical evidence regarding each incentive structure, specifying the laboratory setting.

Table 1 summarizes the results of each study discussed.

Author(s)	**Experimental Task**	**Summary of Results**
Andersen et al. (2014)[72]	Discounting task	The effect of paying only a subset of subjects by exogenously varying the probability of payment from 10% to 100%. Do not find an effect on subject behavior.
Baltussen, Post, van den Assem and Wakker (2012)[73]	Deal or No Deal (dynamic setting)	Risk preferences were consistent on average across the pay one and single-choice treatments. However, in the treatment where only 10% of subjects were selected at random for payment, observe a significant decrease in risk aversion.

Bardsley, Cubitt, Loomes, Moffatt, Starmer andSugden (2009)[66]	Book Chapter about incentive mechanisms	Indicate that the pay one incentive structure eliminated the opportunity for wealth and portfolio effects and eliminates hedging opportunities. They find little evidence of bias when using the random lottery incentive scheme, and conclude there is more support than opposition for this payment scheme.
Beaud and Willinger (2015)[74]	Risk elicitation	Observe no difference in risk vulnerability whether only 10% of subjects are selected at random for payment or all subjects are paid.
Blanco, Englemann, Koch and Normann (2010)[75]	Cooperation game with belief elicitation	Do not find evidence of hedging when paying for both the game outcome and the belief elicitation. Additionally, stated beliefs did not differ whether subjects were paid for both the game outcome and the belief elicitation or one selected at random.
Bolle (1990)[76]	Social Preferences	No difference in behavior across the treatment where subjects are paid for all decisions (pay all) and the treatment where two out of twenty subjects are randomly selected for payment.
Brokesova, Deck, and Peliova (2015)[77]	Risk elicitation	Observe no difference in behavior on a risk-taking task between a single choice, a task selected at random for payment, and a subject selected at random for payment.

Casari, Ham and Kagel (2007)[78]	First price sealed-bid common value auctions	Inexperienced bidders with larger cash balances bid somewhat less aggressively than those with smaller balances. Argue that the effect on behavior is not economically significant but they find even less influence on behavior with experienced bidders.
Clot, Grolleau and Ibanez (2015)[79]	Social Preferences	Find no difference in behavior when all participants are paid versus when only some participants are paid.
Cox (2009)[80]	Social Preferences	Stronger social context (repeated interaction with pay one) resulted in more generous behavior by dictators, second movers in the investment game but not first movers in the investment game than with weak social context (single-choice).
Cox, Sadiraj, and Sadiraj (2008)[81]	Social Preferences	Conduct a test of trust, fear, and reciprocity, with both a single-choice treatment and a pay one treatment and do not observe differences in subject behavior across treatments.

Cox, Sadiraj and Schmidt (2014)[82]	Lottery Choice	Subjects choose the safe option more frequently when listed next to an asymmetrically dominated alternative than when presented in isolation, resulting in a preference reversal that creates concern about using pay one.
Cox, Sadiraj and Schmidt (2015)[83]	Lottery Choice	Test eight different incentive mechanisms and observe large differences in revealed risk preferences across the eight options.
Cubitt, Starmer and Sugden (1998)[84]	Lottery Choice	Find no evidence of a contamination effect and cannot reject the independence axiom therefore behavior appears unbiased with pay one. Additionally, there is no evidence that behavior under the pay one incentive structure is less risk averse.
Erat and Gneezy (2012)[85]	Deception game	Some people prefer not to lie even when lying increases theirs and the receiver's payoff. One out of twenty subjects were paid.
Garvin and Kagel (1994)[86]	First price sealed-bid common value auctions	Find that the coefficient on the cash balance to be statistically significant and negative, indicating evidence that paying all can have a significant effect on subject bidding behavior.

Ham, Kagel and Lehrer (2005)[87]	First price sealed-bid private value auctions	Accumulated cash balances significantly affect behavior and subjects seem to have income targets. Subjects bid high initially to try and win the auction but as cash balances rise and subjects approach what appears to be an income target, bids fall. This suggests that here pay all may have a problematic effect on behavior.
Harrison et al. (2007)[88]	MPL	Do not observe a difference in subject behavior when paying only 10% of the subjects versus paying all subjects.
Harrison, Martinez-Correa and Swarthout (2013)[89]	Lottery Choice	Facing multiple lottery choices and pay one, subjects' choices show a significant shift towards risk-neutral behavior, in line with Expected Utility Theory.
Harrison and Swarthout (2014)[90]	Lottery Choice	When implementing pay one, compared to a single-choice experimental design, there is no difference in behavior under the assumption of Expected Utility Theory.
Laury (2006)[91]	MPL	Finds risk aversion to be highest under high stakes but does not observe a difference in behavior when comparing pay all and pay one under low stakes, therefore, concludes that pay one is not causing subjects to scale down incentives to account for the random selection.

Lee (2008)[92]	Change improving decision model	The introduction of background risk due to the pay one incentive structure makes subjects behave more risk aversely. However, pay one controls for possible wealth effects compared to pay all, and the authors find evidence that pay one is better under repetition in individual decision making experiments.
List, Haigh and Nerlove (2005)[93]	Lottery Choice	Using both students and experienced traders, show that while students' choices do not reflect evidence of the reduction of compound lotteries principle, the choices of the professional traders do.
Loomes, Starmer and Sugden (1991)[94]	Lottery Choice	Subjects preferences exhibit cyclical patterns and the direction of the patterns is consistent with regret theory and a violation of transitivity, therefore the cycles are not due to an issue with pay one.
Schmidt and Hewig (2015)[95]	Lottery Choice	Using a within-subjects design, show that subjects make more risky decisions with pay all than with pay one.
Sefton (1992)[96]	Social Preferences	When the probability of payment for each subject is only 25% because only two out of eight pairs are paid, dictator behavior is more generous than when all subjects are paid.

Sherstyuk, Tarui and Saijo (2013)[97]	Infinite-horizon Prisoners dilemma game	Find that pay one causes players to discount the future more heavily than pay all. Additionally, subjects are more myopic and time inconsistent under pay one. Cooperation rates do not differ whether paying for the last round or all rounds but are significantly lower when paying for one round randomly. Subjects use the "Always Defect strategy significantly more under pay one than under paying for only the last round or paying for all rounds. More subjects use the "Tit-for-Tat and "Trigger-with-Reversion strategies under paying for only the last round or paying for all rounds.
Smeets et al. (2015)[98]	Social Preferences	People with over one million Euro in their bank account play dictator game for $100. One out of ten subjects is paid.

Stahl and Haruvy (2006)[99]	Social Preferences	Pay one misrepresents behavior as egalitarian in the form of "path-dependent egalitarian warm glow. Because subjects choices are only implemented with small probabilities, the cost of an egalitarian option is very low but the benefits of warm glow are felt regardless of the implemented choice. Larger group sizes can minimize other-regarding preferences.
Starmer and Sugden (1991)[100]	Lottery Choice	Compare behavior of subjects facing two pairs of lotteries with one randomly selected for payment and subjects facing a single choice. Behavior is consistent across treatments thus there is no evidence that subjects reduce compound lotteries, suggesting that pay one was not problematic.

3.2.1 Binary-choice lotteries

In the binary-choice lottery paradigm, participants face pairs of lotteries and in each pair they select which lottery they would like to play, and a randomization device then determines the result of the chosen lottery. If a researcher chooses pay all, subjects observe the outcome of each selected lottery and are then paid for each realized outcome. With pay one, though the subjects have submitted decisions for many pairs of lotteries, they are only paid for the outcome of the lottery pair that has been selected randomly. Note that binary-choice-lottery experimental designs already include risk and choosing pay one

introduces an additional layer of risk.

Option A	Option B
1/10 of \$2.00, 9/10 of \$1.60	1/10 of \$3.85, 9/10 of \$0.10
2/10 of \$2.00, 8/10 of \$1.60	2/10 of \$3.85, 8/10 of \$0.10
3/10 of \$2.00, 7/10 of \$1.60	3/10 of \$3.85, 7/10 of \$0.10
4/10 of \$2.00, 6/10 of \$1.60	4/10 of \$3.85, 6/10 of \$0.10
5/10 of \$2.00, 5/10 of \$1.60	5/10 of \$3.85, 5/10 of \$0.10
6/10 of \$2.00, 4/10 of \$1.60	6/10 of \$3.85, 4/10 of \$0.10
7/10 of \$2.00, 3/10 of \$1.60	7/10 of \$3.85, 3/10 of \$0.10
8/10 of \$2.00, 2/10 of \$1.60	8/10 of \$3.85, 2/10 of \$0.10
9/10 of \$2.00, 1/10 of \$1.60	9/10 of \$3.85, 1/10 of \$0.10
10/10 of \$2.00, 0/10 of \$1.60	10/10 of \$3.85, 0/10 of \$0.10

Figure 3.1: Holt and Laury (2002) mechanism

Consider a binary-choice-lottery experimental design where subjects face two decisions. Perhaps the most well-known example is the risk-elicitation multiple price list mechanism used in Holt and Laury (2002; see Figure 1)[101]. This involves an array of 10 decision tasks presented in rows and featuring a higher-variance option (risky) and a lower-variance option (safe). The higher-variance option becomes more and more attractive as one descends down the rows, until choosing this option in the bottom row is the dominant strategy. Participants make a choice in each of the 10 rows; one is chosen at random for actual payoff.

The likelihood of the high (low) payoff increases (decreases) as one moves down the Table, so that Option B is increasingly attractive in lower and lower rows; a risk-neutral person will choose Option A in the top four rows and Option B in the bottom six rows. The row in which one switches from choosing Option A to choosing Option B depends on ones risk preferences. In the original study, one row was selected randomly, with the decision made for that row implemented.

Evidence that pay one and pay all lead to similar results

Laury (2006)[91] addresses the concern of diluted incentives with pay one in a multiple-price-list experiment on risk preferences, with the same format as Holt and Laury (2002)[101].

In one treatment she pays one of the ten choices at low stakes, while in another all ten choices are paid at low stakes. A third treatment is pay one at high stakes. Laury (2006)[91] finds risk aversion to be highest under high stakes but does not observe a difference in behavior when comparing pay all and pay one under low stakes. Because subjects behave the same in these treatments, there does not appear to be evidence that subjects are scaling down the incentives to account for the random selection.

Starmer and Sugden (1991)[100] present subjects with two pairs of lotteries with one pair randomly chosen for payment, comparing behavior in a single-choice treatment and a pay-one treatment in order to test for whether people reduce compound lotteries to simple ones, as prescribed by expected utility. The authors find evidence supporting the usefulness of pay one. Additionally, Cubitt et al. (1998)[84] find no evidence of cross-task contamination with pay one and no evidence that subjects choices in pay one exhibit less risk aversion than in the single-choice treatment.

Harrison and Swarthout (2014)[90] show that when implementing pay one, compared to a single-choice experimental design, there is no difference in behavior under the assumption of Expected Utility Theory (EUT). Indeed, Harrison, Martinez-Correa, and Swarthout (2013)[89] find that even when faced with multiple pairs of lotteries in pay one, subjects choices show a significant shift towards risk-neutral behavior from more risk-averse behavior. In contrast, Schmidt and Hewig (2015)[95] show that subjects make more risky decisions with pay all than with pay one, using a within-subjects design. Since expressed risk preferences differ in the two treatments, this creates concern about using pay one. However, within-subject designs and between-subjects designs may lead to different results; see Charness, Gneezy, and Imas (2013)[102] for more discussion.[4]

In experiments using the binary-choice lottery design where subjects face at least 20

[4]Charness, Gneezy, and Imas (2013)[102] consider the advantages and disadvantages of each type of design. A main point is that within-subject designs may lead to spurious correlations and that between-subjects designs seem more conservative. While between-subject designs lack the statistical power of within-subject designs, significant effects found with this method are more reliably present (Type I error versus Type II error).

pairs of lotteries, one can observe cycles in subjects choices. Loomes, Starmer, and Sugden (1991)[94] investigate if these cycles are the result of a problem with pay one, finding that the direction of the cycles is consistent with regret theory and that they cannot be due to random error. Additionally, they determine that the cycles are the result of a violation of transitivity, rather than reflecting any issue with pay one.

List, Haigh, and Nerlove (2005)[93] use an Allais-paradox experiment to test whether the independence axiom is violated in a study with professional traders from the Chicago Board of Trade. The traders choices do not offer evidence of difficulties with compound lotteries, so that pay one does not seem problematic in this case. Brokesova, Deck, and Peliova (2015)[77] compare behavior on a risk-taking task where it is either the only task and payment is assured, where it is one of several similar tasks of which one will be randomly selected for payment, and whether it is the only task but there is only a small probability of receiving payment. They find no difference in risk-taking behavior when subjects decide between a risky payment and a safe payment across single-choice and pay-one treatments.

Evidence that pay one and pay all lead to different results

There are also studies that show differences in behavior across payment approaches. Cox, Sadiraj, and Schmidt (2014)[82] presented subjects with pairs of lotteries that include asymmetrically-dominated alternatives: An alternative that is inferior in all respects to one option; but, in comparison to the other option, it is inferior in some respects and superior in others (Huber, Payne and Puto, 1982)[103]. The authors find that subjects choose the safe option more frequently when listed next to an asymmetrically dominated alternative than when presented in isolation. The preference reversal calls into question the usefulness of pay one in experiments with this design feature. In addition, Cox, Sadiraj, and Schmidt (2015)[83] run a thorough and exhaustive test of eight different incentive structures and find large differences in revealed risk preferences across the eight options.

The List, Haigh, and Nerlove (2005)[93] study referenced above also has a treatment involving undergraduate students. In contrast to the results with professional traders, the

students do appear to have some difficulty with compound lotteries, creating concern about using pay one.

Nevertheless, our sense is that the overall empirical evidence supports the usefulness of pay one approach in the preponderance of lottery choice experiments, but that concerns arise when the environment deviates from the simple pairwise lottery selection into more complex decisions or when subjects are less sophisticated in terms of complexity and compound lotteries.

3.2.2 Social preferences

Cox, Sadiraj, and Sadiraj (2008)[81] conducted a test of trust, fear, and reciprocity using both triadic and dyadic designs. Subjects participated in either a single-choice treatment or a pay-one treatment, and there was no difference in subject behavior across the two treatments. In this specific social-preference environment, pay one and single-choice approaches produce similar results.

Stahl and Haruvy (2006)[99] find that pay one distorts behavior in dictator games of varying group sizes, where subjects know the selected dictator prior to making allocation choices, leading them to state "Selecting one decision out of many to determine monetary payoffs can drastically reduce the importance of terminal payoffs relative to path-dependent effects. Clearly, when the dictator is announced after subjects have made their choices and the group size is larger, the probability of ones choice being implemented for actual payment is small. In that situation, subjects appear to behave more pro-socially.

Cox (2009)[80] uses game triads to investigate differences in behavior across social contexts. He finds significant changes in behavior given the strong social context generated by repeated interaction and pay one compared to a single-choice weak social context.

On balance, we do not have a clear recommendation regarding using pay one in social-preference experiments. We find that there is reason for caution, but that there is also evidence that using pay one may be relatively innocuous. Whether the experimenter chooses

to implement pay one or pay all in this environment may come down to logistical considerations. For example, Charness and Rabin (2002)[54] considered a large number of simple experimental games and decision tasks in a pen-and-paper experiment. To amass a sufficient number of observations for each game or task, it was necessary to have subjects make choices in a number of different tasks. The authors chose to pay for one of four (or two of eight) decisions, randomly-drawn to ease the process of calculating payoffs and making payments in real time.[5]

3.2.3 Belief elicitation

Another issue involves incentivizing belief elicitation, an important issue for belief-based models of behavior. Concerns can arise when paying for both the game outcome as well as the incentivized belief elicitation. Subjects may find an opportunity to hedge by, for example, betting against their success in the task. One alternative is to pay only for action choices or beliefs, with the choice of which of these two made randomly. Another is to simply not incentivize beliefs. A third option is to make the payoffs from beliefs relatively small; this approach can substantially reduce incentives to hedge, although of course these are not completely eliminated. A fourth option (used by Armentier and Treich, 2009[104], and Danz, Fehr, and Kübler, 2012[105]) is to pay a subject for a stated belief based on the choice of a subject other than the one with whom she is paired.

Blanco, Engelmann, Koch, and Normann (2010)[75] conduct experiments with the Prisoners Dilemma and a coordination game, testing whether behavior differs when people are paid for each task or for only one task drawn at random. They find no difference in stated beliefs or choices in the Prisoners Dilemma, when beliefs are measured using the quadratic-scoring rule. However, they do find a difference in the coordination game using a linear belief-elicitation mechanism. They argue that hedging possibilities are more transparent

[5] An additional issue was that the payoffs in each task might have seemed miniscule if subjects were paid for each task. The evidence from Cabrales, Charness, and Corchn (2003)[67] mentioned earlier also influenced the choice.

in their latter case. Overall, their conclusion is that "Hedging can indeed be a problem in belief-elicitation experiments. However at least according to our results, this seems to be the case only if incentives to hedge are strong and prominent.

So what should an experimenter do when faced with the necessity of eliciting beliefs in a multi-round experiment? The first approach of simply not incentivizing beliefs may be unsatisfactory. Erev, Bornstein, and Wallsten (1993)[106] find differences in behavior across incentivized and non-incentivized beliefs in a public-goods game and Croson (2000)[107] also finds differences in a Prisoners Dilemma. On the other hand, Nyarko & Schotter (2002)[108], Rutstrm and Wilcox (2009)[109] and Gächter and Renner (2010)[110] do not find evidence of differences in behavior according to whether or not beliefs are incentivized. In any case, we are reluctant to recommend not incentivizing beliefs when this is feasible, given that incentivization has long been considered a strong suit of experimental economics. Regarding paying for either actions or beliefs (but not both), the Blanco et al. (2010)[75] results suggest that this may not be necessary if hedging opportunities are not obvious.

Overall, we favor one of the other two approaches, although we feel that these do need more study. Paying for beliefs about another persons action may not always be feasible, but seems attractive when it can be implemented without confusing the subjects. Alternatively, one could simply make the belief payoffs small relative to the choice payoffs, thereby greatly reducing the scope for hedging. In this case, we do not recommend paying only for either beliefs or choices drawn randomly.

3.2.4 Auctions

In both first-price sealed-bid private-value auctions (each potential bidder has a separate draw from the range of possible values) and common-value auctions (the value is the same for all bidders), the accumulation of cash balances from earnings significantly affects bidding behavior including evidence of income targeting (Garvin & Kagel, 1994[86]; Ham, Kagel, and Lehrer, 2005[87]; Casari, Ham, and Kagel, 2007[78]). In the beginning, subjects appear

to bid high in order to win the auction, but as their cash balance grows, bidding falls as they reach their income target (Ham et al., 2005)[87]. This trend in bidding behavior appears more prevalent in inexperienced bidders and diminishes as bidders gain experience in additional sessions (Casari et al., 2007)[78].

Some argue that this issue can be addressed with econometric techniques. For example, in a first-price sealed-bid private value auction, cash balances are endogenous and because only the winners cash balance adjusts each period, there is little variation in cash balances. Ham et al. (2005)[87] tackle the endogeneity problem in two ways: By adjusting the experimental design to allow for exogenous variation in cash balances, and by using instrumental variables in their estimation of the cash-balance effect based on the new exogenous variation in cash balances, including cash balances as an explanatory variable. Still, after correcting for endogeneity, they find significant evidence of cash balances affecting bidding behavior, as mentioned above.

Overall, our sense is that the concerns about cash-balance accumulation and bankruptcy with pay all support using pay one in auction settings.

3.2.5 Dynamic-choice and Infinite-horizon settings

Comparing behavior in a pay-one treatment and a single-choice treatment in a dynamic-choice setting ("Deal or No Deal), Baltussen, Post, van den Assem, and Wakker (2012)[73] find, on average, that risk preferences were consistent across the single-choice and pay-one treatments. This result supports the approach of paying each subject for one of her 10 decisions in dynamic-choice experiments, allowing the researcher to obtain more data and make within-subject comparisons.

Sherstyuk, Tarui, and Saijo (2013)[97] assess behavior in an infinite-horizon prisoners-dilemma game; the infinite horizon is implemented using a stochastic termination rule. Paying subjects for a randomly-selected round as opposed to all rounds resulted in lower cooperation rates, more myopic behavior, and a present period bias. Subjects showed a

higher probability of adopting the "Always Defect strategy rather than the "Tit-For-Tat as found with pay all. While paying all subjects in an infinitely-repeated game can get costly, especially when the incentives must be large enough to ensure subject motivation, it appears that pay one changes behavior relative to pay all.

Sherstyuk et al. (2013)[97] also investigate if paying for only the last round might provide similar results to paying for all rounds in the infinite-horizon environment. The authors find the same cooperation rates and the same strategy profiles in these two treatments. Since paying for only the last round avoids wealth and portfolio effects, this approach (with infinite-horizon environments) seems useful; Chandrasekhar and Xandri (2014)[70] develop theory indicating that this is the best approach.

3.3 Paying only a subset of the participants: Evidence

In some experiments, such as classroom or online, paying only a subset of the participants is desirable from a logistics perspective. Paying only one out of many participants combines the issues discussed above in relation to making the actual pay for a given action probabilistic. In addition, there is potentially a fundamental psychological difference between being definitely paid a positive amount and only having a chance to receive payment, suggesting caution concerned paying a random subset of the experimental subjects.

In general, the results in the literature are encouraging because they show little difference between the two methods. In an early experiment comparing the pay-all to the pay-subset mechanism, Bolle (1990)[76] conducted an ultimatum game experiment comparing the pay-all treatment to a treatment in which only two out of twenty subjects were paid. There was no significant difference in behavior between the treatments.

Sefton (1992)[96] tested the effect of paying all participants versus paying 2 out of 8 pairs in a $5 dictator game. He reports that participants are more generous when only one

in four is paid. Note that the dictator game is particularly sensitive to the actual amount of payment (Forsythe, Horowitz, Savin, and Sefton, 1994[111]) compared for example with payments in the ultimatum game (Slonim and Roth, 1998[112]; Andersen, Ertac, Gneezy, Hoffman, and List, 2011[113]). In contrast with Sefton (1992)[96], Clot, Grolleau and Ibanez (2015)[79] compare the results of a dictator game treatment in which all participants were paid with one in which only a subset was paid, and report no significant difference between the treatments.

Beaud and Willinger (2015)[74] conduct risk-elicitation experiments, using an adaptation of the Gneezy and Potters (1997)[8] risk-elicitation method, while varying different aspects of methodology. In Experiment 1, participants could earn up to 250 Euro, but only 10% of the participants were randomly selected at the end of the experiment to actually be paid. In Experiment 2, all participants were paid according to their earnings in the experiment, but stakes were much lower than in experiment 1 (the maximum was 50 Euro). The main finding is that there is no difference in risk vulnerability (reflecting sensitivity to background risk) across these two designs.[6] This was true despite of the fact that many other features were varied across these treatments.[7] Brokesova et al. (2015)[77] tested how a small probability of receiving payment affected subjects risk-taking behavior compared to a single-choice treatment. In the experiment, subjects faced a choice between a safe and a risky payment that mirrored the structure of a local bank promotion. Again, they find no difference in subjects choices.

Harrison, Lau, and Rutstrm (2007)[88] tested the effect of increasing the payoffs in the

[6]A related potential concern with pay one is the introduction of background risk, described as any risk that is committed but unresolved (see also Lee, 2008)[92]. A randomization device that determines subjects payments after all decisions are made and is not resolved before subjects make their decisions introduces this issue. If the introduction of such background risk alters subjects behavior, pay one is a less attractive approach.

[7]Experiment 1 was a pen-and-paper experiment, while Experiment 2 was computerized. Participants in Experiment 1 were first-year masters students in economics, while participants in Experiment 2 were selected from a subject pool from various disciplines. Finally participants in Experiment 1 received an endowment from the experimenter, while participants in Experiment 2 had to work in a preliminary task to earn a monetary endowment.

Holt and Laury (2002)[101] design discussed above, using payoffs that are roughly 150 times the base payoffs in Holt and Laury (2002)[101], but paid only 10% of their participants. They report an additional experiment that directly examine the hypothesis that paying only a subset of the participants generates the same responses as paying all of them the large amount. Harrison et al (2007)[88] report that they could not reject the null hypothesis that paying all the participants and paying only 10% of them results in the same behavior. Similarly, Andersen, Harrison, Lau, and Rutstrm (2014)[72] explicitly tested the effect of paying only a subset of the participants by exogenously varying the probability of payment for their discounting task from 10% to 100%. They conclude: "the effect of probabilistic discounting is non-existent or negligible in our sample, and for the specifications considered here.

Baltussen et al. (2012)[73] is an exception to the above, demonstrating concerns about paying a random subject, particularly in a complex and dynamic environment. They use the game show "Deal or No Deal to test this type of dynamic-choice environment. In the experimental design, they used a baseline treatment where subjects played the game once for payment. As discussed earlier, one treatment of interest included subjects playing 10 rounds with one of these rounds randomly selected for payment. In another treatment subjects played a single round and 10% of subjects were randomly selected to receive payment.[8] While the paying-a-random-subject treatment and baseline treatment avoid any cross-task contamination concerns, the authors observe a significant decrease in risk aversion in the random-subject treatment.

Overall, the majority of comparisons of paying all the participants versus only a subset of them indicate that the loss of motivation is smallmuch smaller than the implied reduction in actual payment.

An important issue relates to moral choices in cases where pay only one out of N participants is used. In many experiments, the decision made by the participant includes

[8]The face value of prizes was held constant across treatments (guaranteed payment, random round, and random subject).

some moral value. In such situations, participants may care about signaling to others that they are moral people. They might also be concerned about identity, and hence would use their choices to self-signal how moral they are. In situations where signaling is important, paying only some of the participants may be problematic because the moral act is chosen with certainty, while the material payoff is paid only in a fraction of the cases. This problem could affect moral and immoral choices.

Consider for example Smeets, Bauer, and Gneezy (2015)[98], who conduct a dictator game with people that have more than one million Euro in a bank account. These participants receive 100 Euro, and are asked whether they wish to give some of this money to another participant. However, only one out of ten participants is paid according to the decision he/she made. If a participant decides to send a large sum of money, say the entire 100 Euro, then he/she can feel good about the action, and gain the value of positive signaling. The positive signaling value from sending 100 Euro could be large. Yet, the expected cost of this decision is only 10 Euros, because only one out of ten participants is being paid. For an early example of studying this issue, see Stahl and Haruvy (2006)[99].

A similar case, but with a negative action, relates to lying. Participants in Erat and Gneezy (2012)[85] play a deception game in which one party sends a message advising the second party about which action pays the receiver the most. If the advice given is followed, in most cases lying increases the payoff of the sender but decreases the payoff of the receiver, while in some cases lying increases the payoffs for both the sender and receiver. In the experiment, only one out of twenty senders was paid, and the conclusion was that some people prefer not to lie even when lying increases theirs and the receivers payoff. However, as in the dictator game, this conclusion might be inflated because a sender who chooses to lie is a liarregardless of whether he or she subsequently receives payment. The signaling value is certain, while the monetary payoff is not.

Future research could investigate whether this difference between the signal value and its cost is indeed larger when only some of the participants are actually paid. If it is, then

the conclusion drawn from these studies is inflated in the direction of considering people to be moral. If the difference is not important, then this is a very efficient way of collecting data.

3.4 Conclusion

Experimental research in economics differs from many disciplines in that researchers generally incentivize subjects decisions. When selecting an incentive structure, researchers need to assess various characteristics including risk of cross-task contamination, bankruptcy, income effects, and data analysis concerns. While designing an experiment around a single choice is the cleanest approach, such experiments can be expensive and only allow for between-subject data analyses. Gathering data from multiple periods is considerably more efficient than a single-choice design. In addition, more complex decisions may require multiple periods to ensure comprehension.

Yet having subjects complete multiple decisions and paying for each one comes with its own set of issues including the potential for wealth and portfolio effects, cross-task contamination, hedging opportunities, and bankruptcy issues. One alternative approach that we find is generally useful is to have subjects make multiple decisions and pay for a randomly-selected one. But using pay one raises its own concerns: while this eliminates problems associated with wealth and portfolio effects or bankruptcy, there is the potential for diluted incentives and the introduction of background risk. Another technique, which minimizes transactions costs, is to be pay only a subset of the participants. The evidence we have presented suggests that that there may not be a substantial loss of motivation when this device is used.

Table 2 summarizes our recommendations in the different types of experiments that we have discussed.

Environment	**Recommendation**

Lottery Choice/Multiple Price Lists	Empirical evidence supports the usefulness pay one in the great majority of lottery choice experiments but concerns arise when the environment deviates from the simple pairwise lottery selection into more complex decisions.
Social Preferences	Because social preferences span such a variety of environments and the evidence is mixed, we do not make a strong recommendation. It is not clear that pay one leads to different results, so in principle either approach could be employed. Logistical issues may affect the decision to use pay one or pay all.
Belief Elicitation	If hedging opportunities are not obvious, pay one is a useful method. A more cautious approach is to pay for both the game outcome and the belief elicitation but to make the belief payoffs small relative to the choice payoffs, thereby greatly reducing the scope for hedging.
Auctions	The concerns about cash balance accumulation and bankruptcy with a pay all incentive structure support pay one over pay all in auction settings.
Background Risk	Pay one is not problematic if the background risk is resolved prior to the subject making any decisions. For example, handing the subject a sealed envelope containing the randomly-selected round for payment.

Dynamic Settings	Only one study of which we are aware. Pay one and pay all lead to similar behavior in a "Deal or No Deal setting and so we endorse pay one here. However, paying only 10% of subjects led to a significant decrease in risk aversion, so we do not recommend this approach in this setting. More research is needed.
Infinite Horizon	Only one study of which we are aware. Shown both theoretically and empirically in infinite-horizon settings with random termination, subject behavior is consistent whether paying for all rounds or when paying for the last round. More research is needed.

We have detailed existing evidence for and against each incentive structure, providing examples of experiments comparing subjects behavior under different incentive structures. In general, while in some cases the pay-one (and pay-a-subset) method may distort behavior, the data suggest that in the majority of cases it is either equal to (or sometimes superior to) the pay-all method. In general, we feel that both pay one and pay all are useful methods, although there are instances (such as when bankruptcy is possible) where practical considerations suggest using pay one (as is indicated by the theoretical work of Azrieli et al., 2015). We hope to further the discussion about how to best choose an incentive structure when designing an experiment.

3.5 Permissions and Attributions

1. The content of chapter 3 is the result of a collaboration with Gary Charness and Uri Gneezy, and has previously appeared in the Journal of Economic Behavior & Organization, Volume 131, pages 141-150. It is reproduced here with the permission

of Elsevier.

Appendix A

Appendix

A.1 Chapter 1 Appendix

A.1.1 Math Task Treatment Instructions

WELCOME

In the experiment today you will be asked to complete four different tasks. None of these will take more than 5 minutes. At the end of the experiment we will randomly select one of the tasks and pay you based on your performance in that task. Once you have completed the four tasks we determine which task counts for payment by drawing a number between 1 and 4. The method we use to determine your earnings varies across tasks. Before each task we will describe in detail how your payment is determined.

Your total earnings from the experiment are the sum of your payment for the randomly selected task and a $5 show up fee. At the end of the experiment you will

be paid in private.

TASK 1 Piece Rate

For Task 1 you will be asked to calculate the sum of five randomly chosen two-digit numbers. You will be given 5 minutes to calculate the correct sum of a series of these problems. You cannot use a calculator to determine this sum, however you are welcome to write the numbers down and make use of the provided scratch paper. You submit an answer by clicking the submit button with your mouse. When you enter an answer the computer will immediately tell you whether your answer is correct or not. Your answers to the problems are anonymous.

If Task 1 is the one randomly selected for payment, then you get 50 cents per problem you solve correctly in the 5 minutes allotted for this task. Your payment does not decrease if you provide an incorrect answer to a problem. We refer to this payment as the piece rate payment.

Please do not talk with one another for the duration of the experiment. If you have any questions, please raise your hand.

TASK 2 - Tournament

As in Task 1 you will be given 5 minutes to calculate the correct sum of a series of five 2-digit numbers. However for this task your payment depends on your performance relative to that of a group of other participants. Each group consists of

four people, the three other members of your group are located in the same row as you.

If Task 2 is the one randomly selected for payment, then your earnings depend on the number of problems you solve compared to the three other people in your group. The individual who correctly solves the largest number of problems will receive $2 per correct problem, while the other participants receive no payment. We refer to this as the tournament payment. You will not be informed of how you did in the tournament until the end of the experiment. If there are ties the winner will be randomly determined.

Please do not talk with one another. If you have any questions, please raise your hand.

TASK 3 - Choice

As in the previous two tasks you will be given 5 minutes to calculate the correct sum of a series of five 2-digit numbers. However you will now get to choose which of the two previous payment schemes you prefer to apply to your performance on the third task.

If Task 3 is the one randomly selected for payment, then your earnings for this task are determined as follows.

If you choose the piece rate you receive 50 cents per problem you solve correctly. If you choose the tournament your performance will be evaluated relative to the per-

formance of the other three participants of your group in the previous task (Task 2-tournament). The Task 2-tournament is the one you just completed. If you correctly solve more problems than they did in Task 2, then you receive four times the payment from the piece rate, which is $2 per correct problem. You will receive no earnings for this task if you choose the tournament and do not solve more problems correctly now, than the others in your group did in the Task-2 tournament. You will not be informed of how you did in the tournament until the end of the experiment. If there are ties the winner will be randomly determined.

Please choose whether you want the piece rate or the tournament applied to your performance. You will then be given 5 minutes to calculate the correct sum of a series of five randomly chosen two-digit numbers.

Please do not talk with one another. If you have any questions, please raise your hand.

TASK 4 Submit Piece Rate

You do not have to add any numbers for the fourth task of the experiment. Instead, your score from Task 1 - Piece Rate will be used according to your choice for payment. You can either choose to be paid according to the piece rate, or according to the tournament.

If the fourth task is the one selected for payment, then your earnings for this task are determined as follows. If you choose the piece rate you receive 50 cents per

problem you solved in Task 1.

If you choose the tournament your performance will be evaluated relative to the performance of the other three participants of your group in the Task 1-piece rate. If you correctly solved more problems in Task 1 than they did then you receive four times the earnings of the piece rate, which is equivalent to $2 per correct problem. You will receive no earnings for this task if you choose the tournament and did not solve more problems correctly in Task 1 than the other members of your group.

Above, we remind you how many problems you correctly solved in Task 1. Please choose whether you want the piece rate or the tournament applied to your performance.

Please do not talk with one another. If you have any questions, please raise your hand.

A.1.2 Facial Emotion Task Treatment Instructions

WELCOME

In the experiment today you will be asked to complete four different tasks. None of these will take more than 5 minutes. At the end of the experiment we will randomly select one of the tasks and pay you based on your performance in that task. Once you have completed the four tasks we determine which task counts for payment by drawing a number between 1 and 4. The method we use to determine your earnings

varies across tasks. Before each task we will describe in detail how your payment is determined.

Your total earnings from the experiment are the sum of your payment for the randomly selected task and a $5 show up fee. At the end of the experiment you will be paid in private.

TASK 1 Piece Rate

For Task 1 you will be shown 15 photographs depicting individual's faces. For each image, you will be asked to identify the emotion depicted on the individual's face. The emotions in the images have been professionally classified by psychologists doing research in this field. The images will be projected on your computer screen for a very short period of time (2 seconds). After the image is shown, you will be given four options from which to select the correctly displayed emotion. You will have 20 seconds to submit your answer. You submit an answer by clicking the submit button with your mouse. When you enter an answer the computer will immediately tell you whether your answer is correct or not. Your answers to the problems are anonymous. While you will never see the same image twice, the same emotion may be repeated.

If Task 1 is the one randomly selected for payment, then you get 50 cents per emotion you correctly identify in Task 1. Your payment does not decrease if you provide an incorrect answer to an image. We refer to this payment as the piece rate payment.

Please do not talk with one another for the duration of the experiment. If you have any questions, please raise your hand.

TASK 2 Tournament

As in Task 1 you will be shown 15 photographs of individual's faces. You will again be trying to correctly identify the emotions depicted on the faces. For this task your payment depends on your performance relative to that of a group of other participants. Each group consists of four people. The three other members of your group are located in the same row as you.

If Task 2 is the one randomly selected for payment, then your earnings depend on the number of emotions you correctly identify compared to the three other people in your group. The individual who correctly identifies the largest number of emotions will receive $2 per correct emotion, while the other participants receive no payment. We refer to this as the tournament payment. You will not be informed of how you did in the tournament until the end of the experiment. If there are ties the winner will be randomly determined.

Please do not talk with one another. If you have any questions, please raise your hand.

TASK 3 Choice

As in the previous two tasks, you will again be shown 15 photographs of individ-

ual's faces and asked to identify the depicted emotions. However you will now get to choose which of the two previous payment schemes you prefer to apply to your performance on the third task.

If Task 3 is the one randomly selected for payment, then your earnings for this task are determined as follows.

If you choose the piece rate you receive 50 cents per correctly identified emotion. If you choose the tournament your performance will be evaluated relative to the performance of the other three participants of your group in the previous task (Task 2-tournament). The Task 2-tournament is the one you just completed. If you correctly identify more emotions than they did in Task 2, then you receive four times the payment from the piece rate, which is $2 per correct answer. You will receive no earnings for this task if you choose the tournament and do not correctly identify more emotions now, than the others in your group did in the Task-2 tournament. You will not be informed of how you did in the tournament until the end of the experiment. If there are ties the winner will be randomly determined.

Please choose whether you want the piece rate or the tournament applied to your performance. The images will begin on the screen once everyone makes their selection.

Please do not talk with one another. If you have any questions, please raise your hand.

TASK 4 Submit Piece Rate

You do not have to identify any emotions for the fourth task of the experiment. Instead, your score from Task 1 - Piece Rate will be used according to your choice for payment. You can either choose to be paid according to the piece rate, or according to the tournament.

If the fourth task is the one selected for payment, then your earnings for this task are determined as follows. If you choose the piece rate you receive 50 cents per correct answer in Task 1.

If you choose the tournament your performance will be evaluated relative to the performance of the other three participants of your group in the Task 1-piece rate. If you correctly identified more emotions in Task 1 than they did then you receive four times the earnings of the piece rate, which is equivalent to $2 per correct answer. You will receive no earnings for this task if you choose the tournament and did not correctly identify more emotions in Task 1 than the other members of your group.

Above, we remind you how many problems you correctly solved in Task 1. Please choose whether you want the piece rate or the tournament applied to your performance.

Please do not talk with one another. If you have any questions, please raise your hand.

A.2 Chapter 2 Appendix

A.2.1 Effect of being "badly wronged"

While the previous analysis has centered on the impact of the subjects' feelings towards their opponent on their tournament performance, the following analysis will focus on the impact of being "badly wronged" on tournament performance in the first period. I will define a subject as being badly wronged if he contributed at least four more in the public goods game than his opponent. I chose this cutoff because it represents the 90th percentile of differences between own and opponent public goods game contributions. Essentially, those who are "badly wronged" consist of the 10% of subjects who faired worst in the public goods game. This seems like a practical and intuition threshold for categorizing subjects as "badly wronged."

Table 6 presents t-tests as well as Fisher's exact test of medians comparing performance between subjects who were and were not badly wronged in the public goods game. All p-values presented in Table 6 are two-sided unless otherwise specified. Overall, there is a significantly difference in tournament performance between individuals who were and were not badly wronged as evident in the test of medians (p=0.039). Male performance appears unaffected by being badly wronged as the number of completed tasks rises by 1.03 (p=0.7676) on average and the median number of completed tasks rises by 2.00 (p=1.00). The effect of being badly wronged on female tournament performance is marginally significant when looking at the test for a difference in medians. Badly wronged females complete 4.59 more tasks on average than not badly wronged females and the difference in medians is 5.00 (p=0.102). If we look at the one-sided p-value instead because we are interested in an increase in

performance, the difference in means is significant at the 10% level (p=0.085).[1] There is no difference in male and female tournament performance for those subjects who were badly wronged, however there is a significant difference between male and female tournament performance for subjects who were not badly wronged. Males who were badly wronged complete 0.90 more tasks than badly wronged females on average (p=0.7941) while there is no difference in the median number of completed tasks for badly wronged males and females. Males who were not badly wronged complete 4.46 more tasks than badly wronged females on average (p=0.0047) while the difference in the median number of completed tasks for not badly wronged males and females is 3.00 (p=0.010).

While the results using reported feelings and being badly wronged do not exactly mirror each other, the signs of the statistics remain consistent and tell a consistent story that the increase in female performance is driving the overall increase and the gender gap in tournament performance is eliminated.

Table A.1: Difference in Means - Subject was badly wronged (contributed at least 4 more than assigned opponent)

Group 1	Group 1 Mean Perf.	Group 2	Group 2 Mean Perf.	Diff. in Means (p-value)	Diff. in Medians (p-value)
All Subjects	12.10	All Subjects	15.41	3.31	4.00**
(Not Badly Wronged)		(Badly Wronged)		(0.1814)	(0.039)
Males	14.70	Males	15.73	1.03	2.00
(Not Badly Wronged)		(Badly Wronged)		(0.7676)	(1.00)
Females	10.24	Females	14.83	4.59	5.00
(Not Badly Wronged)		(Badly Wronged)		(0.2049)	(0.102)
Females	14.83	Males	15.73	0.90	0.00
(Badly Wronged)		(Badly Wronged)		(0.7941)	(1.00)
Females	10.24	Males	14.70	4.46***	3.00***
(Not Badly Wronged)		(Not Badly Wronged)		(0.0047)	(0.010)

P-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01

[1]There were only six females who were categorized as "badly wronged" and this small sample size could be what is responsible for the large p-values.

A.2.2 All Period Analysis

Between tournament rounds, subjects were informed whether or not they won the previous round. While they were not informed about how many tasks their opponent completed, informing subjects about the results prior to additional round may influence behavior.

This difference in information processing appears to muddle the true effect of the emotions on tournament performance, however, the general trends in male and female performance are maintained. Table 7 shows that over all five periods, subjects competing with an opponent who was rated as strongly negative complete 5.82 tasks more than subjects who completed again a differently rated individual. This difference is highly significant (p=0.0000). Consistent with the first period of data analysis, females respond strongly to the emotion through a significant increase in performance of 6.95 tasks (p=0.0000). While male and female performance differs significantly when competing with an individual not viewed as strongly negative (p=0.0000), this gender gap in tournament performance is almost entirely eliminated in the presence of strongly negative emotions and is only significant at the 10% level. (p=0.0711). While male performance appeared unaffected by the emotions in the period one analysis, male performance appears to differ significantly at the 5% level when assessing all five periods (p=0.0409). The last two results differ slightly from the period one analysis and may be the result of the varying influences of relative performance feedback for men and women. There may even be an interaction between emotions and the processing of relative performance that are confounding the results using all five periods of data. Positive feedback may spur men to compete hard for multiple periods while positive feedback for women may be significantly more transitory. Women may view a successful round as a "fluke", while men will see it as a direct result of their

efforts. However, the general results still hold using either analysis but it is easy to see that an analysis with only the first period of data is much cleaner and without potential confounds.

Table A.2: Difference in Means - All Periods

Group 1	Group 1 Mean Perf.	Group 2	Group 2 Mean Perf.	Difference in Means (p-value)
All Subjects (Rating> 1)	23.90	All Subjects (Rating=1)	29.73	5.82*** (0.0000)
Males (Rating>1)	28.04	Males (Rating=1)	32.21	4.17* (0.0409)
Females (Rating>1)	20.75	Females (Rating=1)	27.70	6.95** (0.0000)
Females (Rating=1)	27.70	Males (Rating=1)	32.21	4.51 (0.0711)
Females (Rating>1)	20.75	Males (Rating>1)	28.04	7.29** (0.0000)

P-values in parentheses. * p<0.05, ** p<0.01

Table 8 presents the regression analyses using all five periods of data. As with the previous analysis, all regressions included session dummy variables and demographic variables. Additionally, I included period dummy variables to account for any learning across periods. Standard errors were clustered on the session level.

The majority of the period one analysis results hold including the significant effect of gender and opponents rated strongly negative on tournament performance. The difference-in-differences estimates are no longer significant using all five periods of data but this is very possibly due to the potential autocorrelation in the standard errors due to the between round feedback. Because I am looking at the differing responses to the emotion across genders, the fact that men and women process and respond to feedback differently can significantly influence behavior and thus coefficient values as well as standard errors and the resulting hypothesis tests.

Additionally, if the regressions in Table 8 are run including a dummy variable equal to one if the individual won the previous tournament round and zero otherwise, in all specifications, the other significant predictors of subject effort are public goods

Table A.3: OLS Regression Results - All Periods

	(1) Completed Slider Tasks	(2) Completed Slider Tasks	(3) Completed Slider Tasks	(4) Completed Slider Tasks
Subject Contribution	0.838*	0.735	0.652	0.640
	(0.454)	(0.459)	(0.454)	(0.456)
Male	6.732**	6.684**	6.646**	7.162**
	(2.503)	(2.570)	(2.553)	(2.762)
Negative Feeling Dummy	0.634			
	(2.283)			
Positive Feeling Dummy	1.013			
	(1.663)			
Strongly Negative		3.165	3.633*	4.732
		(2.341)	(1.826)	(3.167)
Somewhat Negative		-1.641		
		(2.372)		
Somewhat Positive		0.532		
		(2.955)		
Strongly Positive		1.705		
		(3.601)		
Strongly Negative*Male				-2.463
				(4.285)
Constant	7.079**	6.969**	6.802**	6.686**
	(2.918)	(2.881)	(2.722)	(2.724)
R^2	0.2949	0.3027	0.3002	0.3011
N	889	889	889	889
Session Dummies	Yes	Yes	Yes	Yes
Period Dummies	Yes	Yes	Yes	Yes
Demographics	Yes	Yes	Yes	Yes

*p<0.10, **p<0.05. P-values are two-sided. Standard errors in parentheses. Standard errors are clustered by session.

game contribution, a dummy variable for whether the individual is a male, and the "win" dummy variable. More specifically, positive feedback after the previous round results in a significant increase in effort in the following round. These additional regressions provide more concrete evidence that the role of the feedback is eliminates the effect of the emotions and therefore the period one analysis is ideal for answering the current research question.

A.2.3 Instructions and Screen Shots

Welcome

Thank you for choosing to participate in this experiment on decision-making. You will be paid for your participation in cash, privately at the end of the experiment. What you earn depends partly on your decisions and the decisions of the other participants and partly on chance. Please turn off pagers, mp3 devices and cell phones now.

The entire experiment will take place through the computer and there will be no interaction with participants seated at other computers. Please do not talk or in any way try to communicate with other participants during the experiment. This experiment consists of two stages. I first describe Stage 1 and Ill give you instructions for Stage 2 once Stage 1 is over. I will review the instructions with you. Please follow along on your screen and click the Next button when told. If at any point you have a question, please do not hesitate to raise your hand, and I will be by to answer your question. Please refrain from communicating with any of the other participants.

Stage 1 Instructions

In the first stage you will be randomly placed in a group of four. In this stage you will each be given $7. You must decide how much of your $7 to allocate to the group account while keeping the rest for yourself. You can choose any whole number between zero and seven, inclusive. After each group member has chosen an allocation, the money in the group account will be multiplied by 1.6 and then distributed evenly among group members.

Your payoff from the first stage is the sum of what you kept from your $7 and your portion of the group account.

Your Payoff = $7 Your Contribution + 1.6*(Sum of all Contributions)/4

After the money is divided, each person's allocation decision will be shown to all group members listed by ID number (please note your ID number at the top of your screen in blue). Here are a few examples to make sure you understand.

Example 1

One person contributes $7 and the other three contribute nothing.

In this scenario, the individual who contributed everything will receive $2.80.

$7-$7+1.6*(7)/4 = $11.20/4 = $2.80

The other three who contributed nothing will receive $9.80.

$7-$0+1.6*(7)/4 = $7+$11.20/4 = $9.80

Example 2

Everyone contributes $7.

In this scenario, all individuals will receive $11.20.

$$\$7\text{-}\$7+1.6*(28)/4 = \$44.80/4 = \$11.20$$

Example 3

Two people contribute $7 and the other two contribute nothing.

In this scenario, the individuals who contributed everything will receive $5.60.

$$\$7\text{-}\$7+1.6*(14)/4 = \$22.40/4 = \$5.60$$

The other two who contributed nothing will receive $12.60.

$$\$7\text{-}\$0+1.6*(14)/4 = \$7+\$22.40/4 = \$12.60$$

Example 4

Everyone contributes $0.

In this scenario, all individuals receive their initial $7.00 endowment.

$$\$7\text{-}\$0+1.6*(0)/4 = \$7.00$$

Example 5

One person contributes $2, one contributes $5, one contributes $0, and one contributes $7.

In this scenario, the individual who contributed $2 will receive $10.60.

$$\$7\text{-}\$2+1.6*(14)/4 = \$5+\$22.40/4 = \$10.60$$

The individual who contributed $5 will receive $7.60.

$$\$7\text{-}\$5+1.6*(14)/4 = \$2+\$22.40/4 = \$7.60$$

The individual who contributed nothing will receive $12.60.

$$\$7\text{-}\$0+1.6*(14)/4 = \$7+\$22.40/4 = \$12.60$$

The individual who contributed everything will receive $5.60.

$$\$7\text{-}\$7+1.6*(14)/4 = \$22.40/4 = \$5.60$$

You will now be asked to calculate the payoffs for two other potential scenarios in order to check your understanding. Once you have completed the questions, please wait for the other participants to finish as well. Once all participants have finished, the first stage will begin.

Stage 2 Instructions

The second stage will consist of five 1.5-minute (90 seconds) rounds. In this stage, 48 slider bars ranging from 0 to 100 will appear on your screen. Your task is to successfully slide as many bars as you can to the half way mark (50) in the 1.5 minutes (90 seconds). If you complete all 48 slider bars before time is up, a screen

with an additional 48 slider bars will appear for you to work on. You can use any combination of the mouse and keyboard to complete the task. You will be paired with an individual from your group in the first stage, and you will know the ID number of the individual with whom you are now paired. You will be paired with this individual for all five rounds. The screen will display the number of successfully completed tasks as the period progresses. One of the five rounds will be selected at random for payment. The winner (the individual who completes the most tasks successfully among the pair) of the selected round will earn $2.50. In the case of a tie, the winner will be selected at random.

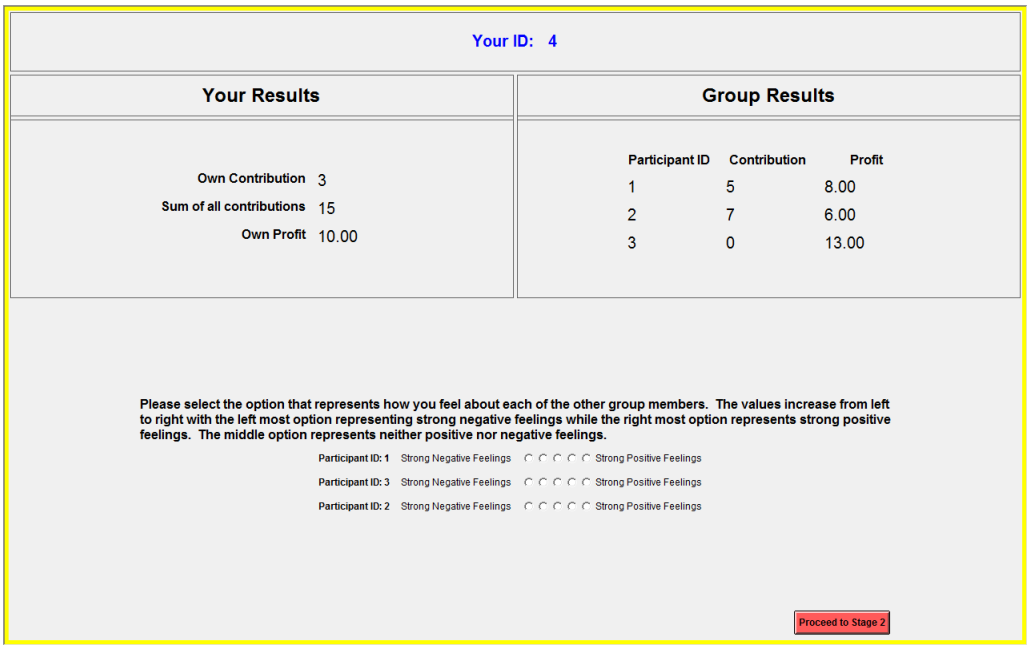

Figure A.1: Public Goods Game Results and Emotion Elicitation

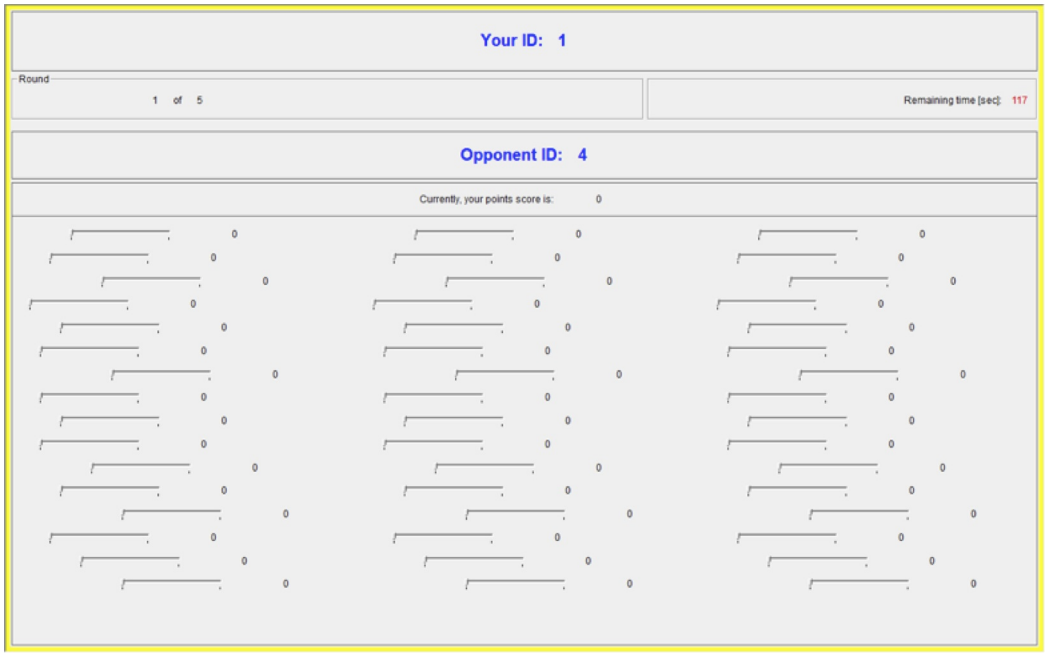

Figure A.2: Slider Task Screenshot

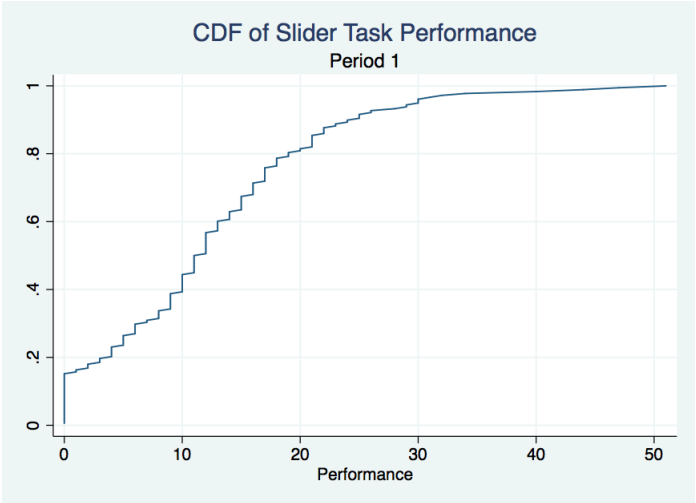

Figure A.3: Cumulative Density Function of Slider Task Performance in Period 1

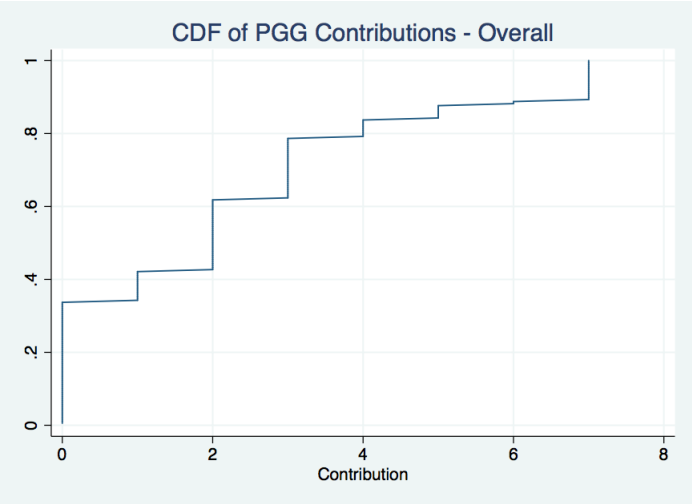

Figure A.4: Cumulative Density Function of Public Goods Game Contributions - All Subjects

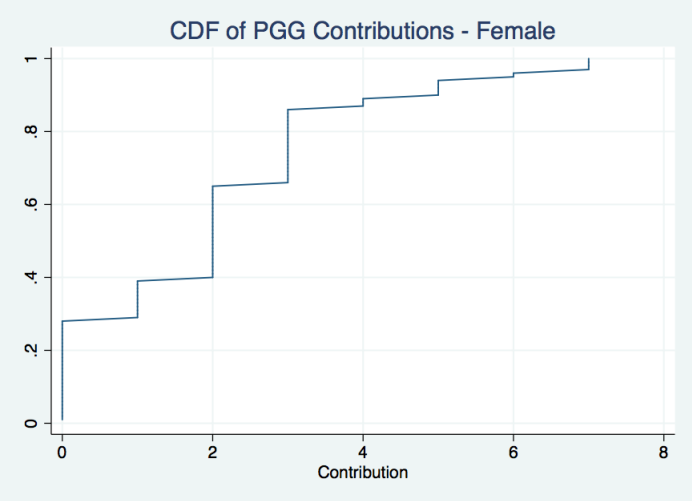

Figure A.5: Cumulative Density Function of Public Goods Game Contributions - Females

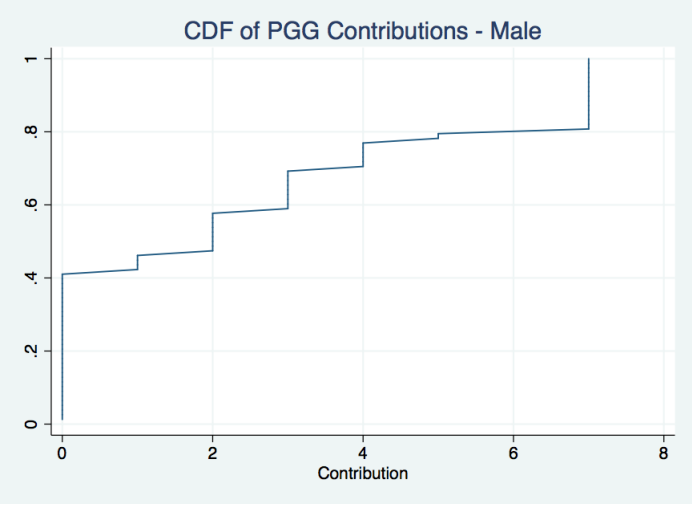

Figure A.6: Cumulative Density Function of Public Goods Game Contributions - Males

Bibliography

[1] M. Niederle and L. Vesterlund, *Do women shy away from competition? do men compete too much?*, *The Quarterly Journal of Economics* **122** (2007), no. 3 1067–1101.

[2] M. A. Lundeberg, P. W. Fox, and J. Punćcohaŕ, *Highly confident but wrong: Gender differences and similarities in confidence judgments*, *Journal of Educational Psychology* **86** (1994), no. 1 114–121.

[3] B. A. Nosek, M. R. Banaji, and A. G. Greenwald, *Math = male, me = female, therefore math ≠ me*, *Journal of Personality and Social Psychology* **83** (2002), no. 1 44–59.

[4] J. Hall, R. C. M. Philip, K. Marwick, H. C. Whalley, L. Romaniuk, A. M. McIntosh, I. Santos, R. Sprengelmeyer, E. C. Johnstone, A. C. Stanfield, A. W. Young, and S. M. Lawrie, *Social cognition, the male brain and the autism spectrum*, *PLOS ONE* **7** (12, 2012).

[5] S. Connor, "The myth of female intuition exploded by fake smile test." `http://www.independent.co.uk/news/science/the-myth-of-female-intuition-exploded-by-fake-smile-test-484585.html`. Accessed: 2015-07-07.

[6] E. R. Simon-Thomas, "Are women more empathic than men?." `http://greatergood.berkeley.edu/article/item/women_more_empathic_than_men`. Accessed: 2016-11-18.

[7] G. Gigerenzer, M. Galesic, and R. Garcia-Retamero, *Stereotypes about men's andwomen's intuitions: A study of two nations*, *Journal of Cross-Cultural Psychology* **45** (2013), no. 1 62–81.

[8] U. Gneezy and J. Potters, *An experiment on risk taking and evaluation periods*, *The Quarterly Journal of Economics* **112** (1997), no. 2 631–645.

[9] S. Frederick, *Cognitive reflection and decision making*, *Journal of Economic Perspectives* **19** (2005), no. 4 25–42.

[10] U. Gneezy, M. Niederle, and A. Rustichini, *Performance in competitive environments: Gender differences*, *The Quarterly Journal of Economics* **118** (2003), no. 3 1049–1074.

[11] U. Gneezy and A. Rustichini, *Gender and competition at a young age*, *American Economic Review* **94** (2004), no. 2 377–381.

[12] C. Günther, N. A. Ekinci, C. Schwieren, and M. Strobel, *Women cant jump?–an experiment on competitive attitudes and stereotype threat*, *Journal of Economic Behavior & Organization* **75** (2010), no. 3 395 – 401.

[13] A. Dreber, E. von Essen, and E. Ranehill, *Outrunning the gender gap–boys and girls compete equally*, *Experimental Economics* **14** (2011), no. 4 567–582.

[14] U. Gneezy and A. Rustichini, "Executives versus teachers." 2004.

[15] N. D. Gupta, A. Poulsen, and M. C. Villeval, *Gender matching and competitiveness: Experimental evidence*, *Economic Inquiry* **51** (2013), no. 1 816–835.

[16] M. Niederle and L. Vesterlund, *Gender and competition*, *Annual Review of Economics* **3** (2011) 601–630.

[17] M. Niederle and A. H. Yestrumskas, "Gender differences in seeking challenges: The role of institutions." 2, 2008.

[18] U. Gneezy, K. L. Leonard, and J. A. List, *Gender differences in competition: Evidence from a matrilineal and a patriarchal society*, *Econometrica* **77** (2009), no. 5 1637–1664.

[19] M.-P. Dargnies, *Men too sometimes shy away from competition: The case of team competition*, *Management Science* **58** (2012), no. 11 1982–2000.

[20] P. Kuhn and M. C. Villeval, *Are women more attracted to co-operation than men?*, *The Economic Journal* **125** (2015), no. 582 115–140.

[21] A. Booth and P. Nolen, *Choosing to compete: How different are boys and girls?*, *Journal of Economic Behavior & Organization* **81** (2012), no. 2 542–555.

[22] M. Niederle, C. Segal, and L. Vesterlund, *How costly is diversity? affirmative action in light of gender differences in competitiveness*, *Management Science* **59** (2013), no. 1 1–16.

[23] A. Cassar, F. Wordofa, and Y. J. Zhang, *Competing for the benefit of offspring eliminates the gender gap in competitiveness*, *PNAS* **113** (2016), no. 19 5201–5205.

[24] B. Halladay, "Gender, emotions, and tournament performance in the laboratory." 10, 2016.

[25] N. Grosse, G. Riener, and M. Dertwinkel-Kalt, "Explaining gender differences in competitiveness: Testing a theory on gender-task stereotypes." 1, 2015.

[26] A. Dreber, E. von Essen, and E. Ranehill, *Gender and competition in adolescence: task matters*, *Experimental Economics* **17** (2014), no. 1 154–172.

[27] J. S. Hyde and M. C. Linn, *Gender differences in verbal ability: A meta-analysis*, *Psychological Bulletin* **104** (1988), no. 1 53–69.

[28] J. S. Hyde, E. Fennema, M. Ryan, L. A. Frost, and C. Hopp, *Gender comparisons of mathematics attitudes and affect: A meta analysis*, *Psychology of Women Quarterly* **14** (1990), no. 3 299–324.

[29] J. L. Meece, J. E. Parsons, C. M. Kaczala, S. B. Goff, and R. Futterman, *Sex differences in math achievement: Toward a model of academic choice*, *Psychological Bulletin* **91** (1982), no. 2 324–348.

[30] S. Skaalvik and E. M. Skaalvik, *Gender differences in math and verbal self-concept, performance expectations, and motivation*, *Sex Roles* **50** (2004), no. 3 241–252.

[31] G. Charness and U. Gneezy, *Portfolio choice and risk attitudes: An experiment*, *Economic Inquiry* **48** (2010), no. 1 133–146.

[32] B. Gillen, E. Snowberg, and L. Yariv, "Experimenting with measurement error: Techniques with applications to the caltech cohort study." 12, 2016.

[33] U. Fischbacher, *z-tree: Zurich toolbox for ready-made economic experiments*, *Experimental Economics* **10** (2007), no. 2 171–178.

[34] H. Hoffmann, H. Kessler, T. Eppel, S. Rukavina, and H. C. Traue, *Expression intensity, gender and facial emotion recognition: Women recognize only subtle facial emotions better than men*, *Acta Psychologica* **135** (2010) 278–283.

[35] T. S. Mohr, "Why women dont apply for jobs unless theyre 100% qualified." `https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified`, 2014. Accessed: 2015-07-07.

[36] G. Charness, R. Cobo-Reyes, and Ángela Sánchez, *The effect of charitable giving on workers performance: Experimental evidence*, *Journal of Behavior & Organization* **131** (2016) 61–74.

[37] C. M. Steele, *A threat in the air: How stereotypes shape intellectual identity and performance*, *American Psychologist* **52** (1997), no. 6 613–629.

[38] B. Cowgill, "Competition and productivity in employee promotion contests." 2015.

[39] L. Vanderkam, "Can coworkers be facebook friends?." `http://www.cbsnews.com/news/can-coworkers-be-facebook-friends/`. Accessed: 2016-03-25.

[40] S. M. Garcia, A. Tor, and T. M. Schiff, *The psychology of competition: A social comparison perspective*, *Perspectives on Psychological Science* **8** (2013), no. 6 634–650.

[41] E. Fehr and S. Gächter, *Cooperation and punishment in public goods experiments*, *American Economic Review* **90** (2000), no. 4 980–994.

[42] E. Fehr and S. Gächter, *Altruistic punishment in humans*, *Nature* **415** (2002) 137–140.

[43] E. Fehr and S. Gächter, *Fairness and retaliation: The economics of reciprocity*, *The Journal of Economic Perspectives* **14** (2000), no. 3 159–181.

[44] B. A. Bettencourt and N. Miller, *Gender differences in aggression as a function of provocation: A meta-analysis*, *Psychological Bulletin* **119** (1996), no. 3 422–447.

[45] R. Croson and U. Gneezy, *Gender differences in preferences*, *Journal of Economic Literature* **47** (2009), no. 2 448–474.

[46] U. Gneezy and A. Imas, *Materazzi effect and the strategic use of anger in competitive interactions*, *PNAS* **111** (2014), no. 4 1334–1337.

[47] D. Gill and V. Prowse, "A novel computerized real effort task based on sliders." 2009.

[48] G. Charness and M. C. Villeval, *Cooperation and competition in intergenerational experiments in the field and in the laboratory*, *American Economic Review* **99** (2009), no. 3 956–978.

[49] G. Charness, *Self-serving cheaptalk: A test of aumann's conjecture*, *Games and Economic Behavior* **33** (2000), no. 2 177–194.

[50] B. Greiner and M. V. Levati, *Indirect reciprocity in cyclical networks: An experimental study*, *Journal of Economic Psychology* **26** (2005), no. 5 711–731.

[51] G. Charness, G. R. Fréchette, and C.-Z. Qin, *Endogenous transfers in the prisoner's dilemma game: An experimental test of cooperation and coordination*, *Games and Economic Behavior* **60** (2007), no. 2 287–306.

[52] E. Fehr and U. Fischbacher, *Third-party punishment and social norms*, *Evolution and Human Behavior* **25** (2004) 63–87.

[53] D. Kahneman and J. L. K. an Richard H. Thaler, *Fairness and the assumptions of economics*, *The Journal of Business* **59** (1986), no. 4 285–300.

[54] G. Charness and M. Rabin, *Understanding social preferences with simple tests*, *The Quarterly Journal of Economics* **117** (2002), no. 3 817–869.

[55] G. Charness, *Attribution and reciprocity in an experimental labor market*, *Journal of Labor Economics* **22** (2004), no. 3 665–688.

[56] T. Offerman, *Hurting hurts more than helping helps*, *European Economic Review* **46** (2002), no. 8 1423–1437.

[57] T. Buser and A. Dreber, *The flipside of comparative payment schemes*, *Management Science* **62** (2015), no. 9 2626–2638.

[58] C. C. Eckel and P. J. Grossman, *The relative price of fairness: gender differences in a punishment game*, *Journal of Economic Behavior & Organization* **30** (1996), no. 2 143–158.

[59] A. Christensen, M. Sullaway, and C. E. King, *Systematic error in behavioral reports of dyadic interaction: Egocentric bias and content effects*, *Behavioral Assessment* **5** (1983), no. 2 129–140.

[60] D. Wozniak, *Gender differences in a market with relative performance feedback: Professional tennis players*, *Journal of Economic Behavior & Organization* **83** (2012), no. 1 158–171.

[61] V. Grimm and F. Mengel, *Let me sleep on it: Delay reduces rejection rates in ultimatum games*, *Economics Letters* **111** (2011), no. 2 113–115.

[62] S. G. Donald and K. Lang, *Inference with difference-in-differences and other panel data*, *The Review of Economics and Statistics* **89** (2007), no. 2 221–233.

[63] Y. Azrieli, C. P. Chambers, and P. J. Healy, "Incentives in experiments: A theoretical analysis." 2015.

[64] G. Charness and D. Levin, *The origin of the winner's curse: A laboratory study*, *American Economic Journal: Microeconomics* **1** (2009), no. 1 207–236.

[65] R. P. Cubitt, C. Starmer, and R. Sugden, *Discovered preferences and the experimental evidence of violations of expected utility theory*, *Journal of Economic Methodology* **8** (2001), no. 3 385–414.

[66] N. Bardsley, R. Cubitt, G. Loomes, P. Moffat, C. Starmer, and R. Sugden, *Experimental Economics: Rethinking the Rules*. Princeton University Press, 2010.

[67] A. Cabrales, G. Charness, and L. C. Corchón, *An experiment on nash implementation*, *Journal of Economic Behavior & Organization* **51** (2003), no. 2 161–193.

[68] C. A. Holt, *Preference reversals and the independence axiom*, *American Economic Review* **76** (1986), no. 3 508–515.

[69] E. Karni and Z. Safra, *"preference reversal" and the observability of preferences by experimental methods*, *Econometrica* **55** (1987), no. 3 675–685.

[70] A. G. Chandrasekhar and J. P. Xandri, "A note on payments in the lab for infinite horizon dynamic games with discounting." 2014.

[71] A. Baillon, Y. Halevy, and C. Li, "Experimental elicitation of ambiguity attitude using the random incentive system." 2015.

[72] S. Andersen, G. W. Harrison, M. I. Lau, and E. E. Rutström, *Discounting behavior: A reconsideration*, *European Economic Review* **71** (2014) 15–33.

[73] G. Baltussen, G. T. Post, M. J. van den Assem, and P. P. Wakker, *Random incentive systems in a dynamic choice experiment*, *Experimental Economics* **15** (2012), no. 3 418–443.

[74] M. Beaud and M. Willinger, *Are people risk vulnerable?*, *Management Science* **61** (2015), no. 3 624–636.

[75] M. Blanco, D. Engelmann, A. K. Koch, and H.-T. Normann, *Belief elicitation in experiments: is there a hedging problem?*, *Experimental Economics* **13** (2010), no. 4 412–438.

[76] F. Bolle, *High reward experiments without high expenditure for the experimenter?*, *Journal of Economic Psychology* **11** (1990), no. 2 157–167.

[77] Z. Brokesova, C. Deck, and J. Peliova, "Bringing a natural experiment into the laboratory: the measurement of individual risk attitudes." 2015.

[78] M. Casari, J. C. Ham, and J. H. Kagel, *Selection bias, demographic effects, and ability effects in common value auction experiments*, *American Economic Review* **97** (2007), no. 4 1278–1304.

[79] S. Clot, G. Grolleau, and L. Ibanez, "Shall we pay all? an experimental test of random incentivized systems." 2015.

[80] J. C. Cox, *Trust and reciprocity: implications of game triads and social contexts*, *New Zealand Economic Papers* **43** (2009), no. 2 89–104.

[81] J. C. Cox, K. Sadiraj, and V. Sadiraj, *Implications of trust, fear, and reciprocity for modeling economic behavior*, *Experimental Economics* **11** (2008), no. 1 1–24.

[82] J. C. Cox, V. Sadiraj, and U. Schmidt, *Alternative payoff mechanisms for choice under risk*, *International Advances in Economic Research* **20** (2014), no. 2 239–240.

[83] J. C. Cox, V. Sadiraj, and U. Schmidt, *Paradoxes and mechanisms for choice under risk*, *Experimental Economics* **18** (2015), no. 2 215–240.

[84] R. P. Cubitt, C. Starmer, and R. Sugden, *On the validity of the random lottery incentive system*, *Experimental Economics* **1** (1998), no. 2 115–131.

[85] S. Erat and U. Gneezy, *White lies*, *Management Science* **58** (2012), no. 4 723–733.

[86] S. Garvin and J. H. Kagel, *Learning in common value auctions: Some initial observations*, *Journal of Economic Behavior & Organization* **25** (1994), no. 3 351–372.

[87] J. C. Ham, J. H. Kagel, and S. F. Lehrer, *Randomization, endogeneity and laboratory experiments: the role of cash balances in private value auctions*, *Journal of Econometrics* **125** (2005), no. 1 175–205.

[88] G. W. Harrison and M. I. L. an E. Elisabet Rutström, *Estimating risk attitudes in denmark: A field experiment*, *The Scandinavian Journal of Economics* **109** (2007), no. 2 341–368.

[89] G. W. Harrison, J. Martínez-Correa, and J. T. Swarthout, *Inducing risk neutral preferences with binary lotteries: A reconsideration*, *Journal of Economic Behavior & Organization* **94** (2013) 145–159.

[90] G. W. Harrison and J. T. Swarthout, *Experimental payment protocols and the bipolar behaviorist*, *Theory and Decision* **77** (2014), no. 3 423–438.

[91] S. K. Laury, "Pay one or pay all: Random selection of one choice for payment." 2006.

[92] J. Lee, *The effect of the background risk in a simple chance improving decision model*, *Journal of Risk and Uncertainty* **36** (2008), no. 1 19–41.

[93] J. A. List, M. S. Haigh, and M. Nerlove, *A simple test of expected utility theory using professional traders*, *PNAS* **102** (2005), no. 3 945–948.

[94] G. Loomes, C. Starmer, and R. Sugden, *Observing violations of transitivity by experimental methods*, *Econometrica* **59** (1991), no. 2 425–439.

[95] B. Schmidt and J. Hewig, *Paying out one or all trials: A behavioral economic evaluation of payment methods in a prototypical risky decision study*, *The Psychological Record* **65** (2015), no. 2 245–250.

[96] M. Sefton, *Incentives in simple bargaining games*, *Journal of Economic Psychology* **13** (1992), no. 2 263–276.

[97] K. Sherstyuk, N. Tarui, and T. Saijo, *Payment schemes in infinite-horizon experimental games*, *Experimental Economics* **16** (2013), no. 1 125–153.

[98] P. Smeets, R. Bauer, and U. Gneezy, *Giving behavior of millionaires*, *PNAS* **112** (2015), no. 34 10641–10644.

[99] D. O. Stahl and E. Haruvy, *Other-regarding preferences: Egalitarian warm glow, empathy, and group size*, *Journal of Economic Behavior & Organization* **61** (2006), no. 1 20–41.

[100] C. Starmer and R. Sugden, *Does the random-lottery incentive system elicit true preferences? an experimental investigation*, *American Economic Review* **81** (1991), no. 4 971–978.

[101] C. A. Holt and S. K. Laury, *Risk aversion and incentive effects*, *American Economic Review* **92** (2002), no. 5 1644–1655.

[102] G. Charness, U. Gneezy, and A. Imas, *Experimental methods: Eliciting risk preferences*, *Journal of Economic Behavior & Organization* **87** (2013) 43–51.

[103] J. Huber, J. W. Payne, and C. Puto, *Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis*, *The Journal of Consumer Research* **9** (1982), no. 1 90–98.

[104] O. Armentier and N. Treich, *Subjective probabilities in games: An application to the overbidding puzzle*, *International Economic Review* **50** (2009), no. 4 1079–1102.

[105] D. N. Danz, D. Fehr, and D. Kübler, *Information and beliefs in a repeated normal-form game*, *Experimental Economics* **15** (2012), no. 4 622–640.

[106] I. Erev, G. Bornstein, and T. S. Wallsten, *The negative effect of probability assessments on decision quality*, *Organizational Behavior and Human Decision Processes* **55** (1993), no. 1 78–94.

[107] R. Croson, *Thinking like a game theorist: factors affecting the frequency of equilibrium play*, *Journal of Economic Behavior & Organization* **41** (2000), no. 3 299–314.

[108] Y. Nyarko and A. Schotter, *An experimental study of belief learning using elicited beliefs*, *Econometrica* **70** (2002), no. 3 971–1005.

[109] E. E. Rutström and N. T. Wilcox, *Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test*, *Games and Economic Behavior* **67** (2009), no. 2 616–632.

[110] S. Gächter and E. Renner, *The effects of (incentivized) belief elicitation in public goods experiments*, *Experimental Economics* **13** (2010), no. 3 364–377.

[111] R. Forsythe, J. L. Horowitz, N. E. Savin, and M. Sefton, *Fairness in simple bargaining experiments*, *Games and Economic Behavior* **6** (1994), no. 3 347–369.

[112] R. Slonim and A. E. Roth, *Learning in high stakes ultimatum games: An experiment in the slovak republic*, *Econometrica* **66** (1998), no. 3 569–596.

[113] S. Andersen, S. Ertac, U. Gneezy, M. Hoffman, and J. A. List, *Stakes matter in ultimatum games*, *American Economic Review* **101** (2011), no. 7 3427–3439.