UNIVERSITY OF CALIFORNIA, SAN DIEGO

Essays in Behavioral Economics and Ethics

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

in

Management

by

Silvia Saccardo

Committee in charge:

Professor Uri Gneezy, Chair
Professor Ayelet Gneezy, Co-Chair
Professor James Andreoni
Professor Yuval Rottenstreich
Professor Sally Sadoff

2015

The Dissertation of Silvia Saccardo is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2015

DEDICATION

I want to dedicate this dissertation to my family, who has consistently been a firm believer in my dreams, and has always encouraged me to pursue them. Thank you for your unconditional love and support throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I want to thank my advisor, Uri Gneezy, who has offered me invaluable mentorship and support throughout the years. There are so many things to thank him for. Uri took a chance on me despite my unconventional background, and relentlessly guided me along my path in research. From him I learned that the questions worth pursuing are those that make you excited, passionate, and impatient about learning their answers. I am grateful he encouraged me to work with other people, which is one of the things I love the most about academia. Working with Uri has been an incredibly enriching experience, and has made every day of work fun. And, thanks for the paddle board!

Another huge thanks goes to my co-advisor, Ayelet Gneezy. She was the one who gave me the opportunity to begin this journey by opening the doors to the fascinating world of behavioral research. Ever since, she has played a critical role in my growth as a researcher. She challenged me to think critically about research questions and inspired me with her remarkable drive. From her I learned that by working hard anything can be made possible. I thank Ayelet for her constant support and encouragement, and for always being there in the crucial moments!

I would also like to thank the rest of my committee. Sally Sadoff, who has always kept her door open to talk and discuss, and made being a TA very enjoyable. Jim Andreoni, who introduced me to a different approach on experimental research, both through his Ph.D. class and with his valuable feedback on my research project.

Yuval Rottenstreich, who has been truly helpful and insightful throughout the job market process, never hesitating to give me honest advice.

This body of work would not have been possible without my collaborators, from who I learned a lot. Roel van Velduhizen, with whom I started a fruitful collaboration that evolved into Chapters 1 and 2. He is always helpful and critical, and from him I learned that attention to details is never too much. Marta Serra-Garcia, a co-author of Chapter 2, from whom I learned to be comprehensive but yet right to the point. Christina Gravert and Franziska Tausch, with whom I shared the fun and stress of working on Chapter 3. Christina is enthusiastic and perseverant, and from her I learned a lot about networking and selling ideas to the outside world. Franzi's optimism and cheerfulness are contagious, and collaborating with her has truly been a pleasure. Anastasia Danilov is a special co-author to me. Working on Chapter 4 has been exciting and has made our friendship grow stronger. I also thank Aniela Pietraz for her work on chapter 5.

I am grateful to several other faculty members at Rady, who have provided me with support and constant feedback. A big thanks to Karsten Hansen, Charlie Sprenger, Pam Smith, Chris Oveis, Wendy Liu, On Amir, and Nora Williams.

The support from a big community of peers has been key for reaching this milestone. I would like to thank Alex Imas, for being such an inspiring and helpful colleague, Elizabeth Keenan, for sharing with me the ups and down of graduate life, and all the other Rady Ph.D. students.

Over the years, my life at Rady has been enriched, both academically and socially, by the presence of several researchers—The Visitors—who spent some time

with our group. The presence of a group of dedicated junior researchers around me has broadened my horizons, making the research environment even more stimulating. Together, we have built a nice community, and I am excited to see how our careers will evolve. A big thanks to them all, and especially to Orsola, my first friend in academia, Lea, for our time at summer camp and for "just reporting", Katharina, for making my last months in San Diego unforgettable, Agne, for the great partying and for listening so many of my talks, Marina, for offering unconditional help, and Janna, for sharing with me the struggles of running field experiments.

And because life is not only about work, I want to thank those who had a crucial role in making life outside academia fun. My family and friends in San Diego, Emanuele, Liz, Riccardo, Danieletto, and Alice. Without them my time here wouldn't have been the same! My friends from home, who traveled all the way to California to see me, and make me feel as if I had never left every time I go back: Marta, Sandro, Gabriele and Matteo. Gianni, Angela, and Nonna Milvia, who have welcomed me in their family with tons of great food. Giacomo and Virginia, for enthusiastically following my adventures via Skype and for giving me the joy of becoming aunt.

Most importantly, I will be eternally grateful to my family: Mamma e Papà, to whom I owe much of what I learned about life. Throughout their journey, they have shown me how to believe in something and make it happen, how to lay out our own path and follow it, how to never stop in front of all difficulties that try to slow us down, and how to always look ahead to a bright future.

Lastly I want to thank Lorenzo, who has been my rock from the beginning and has made life exciting day after day. He encouraged me to apply for a Ph.D., he

tirelessly gave me feedback on projects, papers, ideas, and data analysis, and showed his love and support in every possible way. I am thrilled to have him by my side, and I am looking forward to keep pursuing our dreams together. Without him and Bimino none of this would have been possible.

Chapter 1, in full, has been submitted for publication of the material as it may appear in the Review of Economic Studies, 2015. Gneezy, Uri, Silvia Saccardo, Roel van Veldhuizen, "Bribery: Greed versus Reciprocity." The dissertation author was the co-primary investigator and author of this paper.

Chapter 2, in part, is currently being prepared for submission for publication of the material. Uri Gneezy, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen, "Motivated Self-Deception, Identity, and Unethical Behavior." The dissertation author was the co-primary investigator and author of this paper.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Gneezy, Uri, Christina Gravert, Silvia Saccardo, and Franziska Tausch, "A Must Lie Situation: Avoiding Giving Negative Feedback." The dissertation author was the co-primary investigator and author of this paper.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Danilov, Anastasia and Silvia Saccardo, "Discrimination in Disguise." The dissertation author was the co-primary investigator and author of this paper.

Chapter 5, in full, has been submitted for publication of the material as it may appear in Management Science, 2015, Gneezy, Uri, Aniela Pietraz, and Silvia Saccardo. "On the Size of the Gender Difference in Competitiveness." The dissertation author was the co-primary investigator and author of this paper.

VITA

2007	Laurea Triennale in Psychology, summa cum laude, University of Padova, Italy
2010	Laurea Specialistica in Psychology, summa cum laude, University of Padova, Italy
2015	Doctor of Philosophy in Management, Rady School of Management, University of California San Diego

ABSTRACT OF THE DISSERTATION

Essays in Behavioral Economics and Ethics

by

Silvia Saccardo

Doctor of Philosophy in Management

University of California, San Diego, 2015

Professor Uri Gneezy, Chair

Professor Ayelet Gneezy, Co-Chair

My dissertation examines the behavioral factors that affect the emergence of unethical behaviors and inequalities in today's society. Using insights from behavioral economics and experimental methods, I investigate the drivers of phenomena such as corruption, dishonesty, ethnic-discrimination, and gender-based differences in preferences.

Chapter 1 explores the mechanism through which receiving bribes leads evaluators to distort choices. In both a laboratory experiment in the US and an experiment in a market in India, evaluators receive bribes that distort their quality recommendations. We show that the driver of distortion is greed and not reciprocity.

Chapter 2 examines how self-deception affects judgment distortion in the presence of incentives. We show that when evaluators can convince themselves that they are behaving ethically, they are more likely to distort their judgment. When self-deception is not possible, recommendations are more honest. This shows that in some cases people are able to behave unethically without suffering from feelings of guilt or shame, by convincing themselves that they are ethical.

Chapter 3 explores individuals' unwillingness to provide negative feedback to others, which results in a "must lie situation". By asking experimental subjects to evaluate others' attractiveness, we show that individuals prefer to lie rather than tell an undesirable truth, even if lying comes at a monetary cost to both the person who gives the feedback and the person who receives it.

Chapter 4 studies prejudice-based ethnic discrimination, and shows that individuals are more likely to discriminate against others when discrimination can be disguised. We show that individuals do not discriminate in contexts where discrimination cannot be plausibly justified. However, discrimination emerges in contexts in which discriminatory behavior can be attributed to conformity to social or moral norms.

Chapter 5 explores gender differences in preferences for competitiveness, which have been suggested to partly account for the relative lack of success of women

in many sectors of the labor market. We introduce a novel measure that captures the extent of competitiveness. We find that the gender gap in competitiveness is larger than what had been documented before, with strikingly lower ratios of women at the top of the competitiveness distribution.

Introduction

Our choices are governed by norms of behavior, conventions, social customs and inertia forces. These forces are an (incomplete) prescription of behavior for different circumstances that are common to groups of individuals, and therefore are labeled "social norms." Simple observations from our daily life make clear that many social norms affect how we behave. There are norms of fair behavior, norms that specify behavior towards family members, norms that restrict the type of food we eat, and norms that govern our social activity—how to behave in meetings, or in restaurants, whom to invite for dinner, or how much to tip. There are norms that specify the ownership of goods, basic civil rights, contribution to a common cause or what negotiation in good faith is.

When norms do not contradict self-interest, following them does not create any dilemma. For example, a norm (or a convention) that specify the side of the road in which we drive does not contradict self-interest. But norms that advocate fair behavior or that encourage donating to charity, or that condemn discrimination, dishonesty or bribing in a business transaction, may contradict private self-interest. These type of norms give rise to a conflict between following the norm and profit maximizing behavior. This dissertation advances our understanding of the behavioral factors that affect choices and judgments when ethical rules or social norms conflict with profit maximizing motives.

Chapter 1 (coauthored with Uri Gneezy and Roel van Veldhuizen) focuses on bribery, a widespread and economically important phenomenon, with over a trillion dollars exchanged in bribes every year around the world (Kauffman, 2005). In this chapter we experimentally investigate how receiving bribes affect individuals' decision making. In a novel design, two participants (the workers) perform a task, competing for a prize. A third participant (the referee) is asked to select the best performer, who receives a monetary prize. Workers have the opportunity to bribe the referee. This experimental paradigm captures an important outcome of bribery: the distortion of judgment and facts that occurs when decision-makers base their decisions on bribes rather than performance or quality, and therefore behave unethically. We investigate whether individuals are willing to engage in bribery and whether bribes distort outcomes, inducing the referee to award the prize to the worker who sent the higher bribe. When the referee is allowed to only keep the winners' bribe we find that individuals are willing to make decisions that distort the true ranking between the workers, awarding the prize to the one who sent the higher bribe. When the referee is allowed to keep any bribe received regardless of her choice of a winner, we find that referees do not distort their judgment, as decisions are largely based on performance. We replicate this finding in an extra-lab experiment conducted in India, in the market of the city of Shillong. In the experiment, we ask shoppers in the market to taste two pineapples from two different vendors, and select the tastier. Shoppers know we would then buy a pineapple from the seller they indicate. Before the beginning of the experiment, vendors agreed to pay some money to each shopper who recommends their pineapple. In line with the laboratory experiment, we find that shoppers distort

their judgment only when their choices directly affect their earnings. Taken together, the results from both studies suggest that the mechanism by which bribes work in our setting is greed (i.e. profit maximization) and not reciprocity.

Chapter 2 (co-authored with Uri Gneezy, Marta Serra-Garcia, and Roel van Veldhuizen) delves into the psychology of judgment distortion, investigating the relationship between self-deception and unethical behavior. While some people have no psychological costs associated with behaving unethically and do so whenever it is profit maximizing, for others, distorting ethical judgment comes with a cost to self-image. When facing the opportunity to distort ethical judgments for a financial gain, these people face a tension between maintaining their self-image as a moral person and the desire to increase material goals. This tension may be attenuated if individuals engage in self-deception, convincing themselves that their behavior is ethical. We explore this hypothesis in an experiment in which decision-makers are asked to evaluate two subjective options and recommend one based on quality. In the experiment, an advisor recommends one of two investment choices to a client. The two options differ in risk and expected return; no option strictly dominates the other. The advisor has a monetary incentive to recommend a specific investment option. We explore whether evaluators are more likely to distort their judgment in favor of the incentivized option when they can convince themselves that their choice is ethical. For this purpose, we contrast two timelines of decision-making that manipulate individuals' scope for self-deception. In one case, the evaluator is informed about an incentive associated to one of the options before she is provided with information about the quality of the options. In the other case, the evaluator receives the

information about the incentive only after she privately evaluates the quality of the options. In the first case, the evaluator can easily convince herself that recommending the incentivized option is ethical while in the second case there is less scope for self-deception. In line with our hypothesis, we find recommendations to be biased in the direction of the incentive in the former case, but not in the latter. In an additional experiment, we limit the scope for self-deception by introducing strict dominance of one of the lotteries that the advisor evaluates over the other one. Strict dominance implies that advisors cannot easily convince themselves that recommending the dominated lottery is ethical. In this context, we observe no effect of the timing manipulation. We term this behavior "motivated self-deception", since advisors engage in self-deception to protect their identity when motivated to do so by incentives.

Chapter 3 (co-authored with Uri Gneezy, Christina Gravert, and Franziska Tausch) explores settings in which individuals prefer to lie rather than telling a negative truth to another person, referred to as the "must lie situation". We report experimental results showing that individuals are reluctant to give honest negative feedback to others, even if honesty could help the receiver achieve better outcomes. We choose to use attractiveness as the subject of feedback, as a proxy for traits individuals care about, and for which objective information is hard to obtain. In a novel experimental design, we ask individuals to rank other participants in term of attractiveness. In the first stage of the experiment, subjects (ten men and ten women) provide a ranking of all individuals of the opposite sex. In the second stage, we ask individuals to guess their own position in the ranking. We find that individuals are

largely overconfident about their own attractiveness ranking. We then randomly assign subjects to treatments. In one case, we ask individuals to confidentially guess the ranking of another same-sex participant. We find that despite inflating the guesses about their own ranking, individuals are able to accurately guess the attractiveness rank of another individual of the same sex. In another treatment, we ask them to report their guess via face to face feedback to the other participant. That is, each individual sends an open message to another same-sex participant, indicating a suggested rank for that participant. After receiving the message, participants have a chance to update the guess about their own ranking, increasing their earnings. In line with our hypothesis, people give accurate face-to-face feedback to attractive individuals, but avoid doing so to the less-attractive ones. In our experiment this comes at a monetary cost to both the person who gives the feedback and the one receiving it. We also find that, surprisingly, a substantial increase of these costs—through a raise in the price of providing biased feedback from $10 to $50 for both parties—does not increase honesty. Finally, to shed light on the mechanism driving the reluctance to provide negative feedback, we compare face-to-face to anonymous feedback. We find that when the identity of the feedback provider is not revealed, feedback towards the less-attractive individuals is more honest. Our results suggest that the inflated face-to-face-feedback we identify in our experiment is not driven by individuals' unwillingness to hurt the recipient.

Chapter 4, (co-authored with Anastasia Danilov), explores whether the opportunity of disguising behavior behind adherence to moral or social norms makes individuals more likely to engage in discrimination toward minorities. Modern

societies have made substantial advances in establishing policies and social norms against ethnic discrimination. Today, societies are more diverse and multicultural than ever before, and open expression of prejudice has declined. Yet, inequalities on the grounds of race and ethnicity persist. In a series of experiments, we show that when individuals can plausibly rationalize their behavior as non-discriminatory and therefore maintain a positive image, ethnic discrimination emerges. When disguising discrimination is not possible, however, individuals act similarly toward their own or a different ethnicity. We study subjects' behavior toward individuals of either their own or of a different ethnicity by asking subjects to make choices that affect both their payoffs and the payoffs of another participant. In a dictator game, subjects determine the earnings of both participants by choosing between two options. We study the rate of prosocial choices as a function of the receiver's ethnicity. In this context, individuals cannot easily attribute lack of prosocial behavior toward the other participant to something other than prejudice. Hence, we expected to find no difference in the rate of prosocial choices. In line with our hypothesis, we find no evidence of discrimination. In a second experiment, subjects could reach the same prosocial payoff allocation of the dictator game by telling an altruistic lie to their counterpart. In this context, the possibility of complying with the norm of honesty might provide subjects with a reason not to favor the altruistic payoff allocation. We find that individuals are less likely to tell a white lie when lying helps a member of a different ethnicity than when it benefits an individual of their own ethnicity. These results provide evidence of taste-based discrimination disguised behind honesty. In an additional experiment in which individuals have scope for using fairness norms self-

servingly, we again find evidence of ethnic discrimination. Taken together, our findings suggest that prejudice-driven discrimination still persists and arises in situations in which individuals can plausibly justify their behavior.

Finally, Chapter 5 (co-authored with Uri Gneezy and Aniela Pietraz) turns to the investigation of gender differences in competitiveness. Although the number of women in leadership position has increased over the past years, the gender gap in labor markets is still large. Several factors can contribute to this gap, such as discrimination or individuals' preferences. In this chapter, we focus on gender differences in competitiveness. Previous experimental literature has argued that women are less competitive than men, and that this difference can contribute in explaining wage gaps in the labor market. These papers have both shown that women perform worse than men in competitive settings (e.g. Gneezy, Niederle and Rustichini, 2003), and that women are less likely to select into competitive environments (e.g., Niederle and Vesterlund, 2007). The papers that study selection consistently find that about a third of women and two thirds of men select into competitive environments. In this chapter, we investigate a new dimension of competitiveness—the extent of competitiveness. We introduce a new measure that allows us to observe 101 levels of competitiveness. In particular, we ask subjects to perform a task and choose what percentage of their compensation they would prefer to be derived from a tournament scheme and which percentage they would prefer to be derived from a piece rate scheme. Hence, this measure allows us to measure the intensive margin of competitiveness. We find that the evidence for gender difference in competitiveness is much stronger than that revealed by previous experimental paradigms studying

competitiveness on the extensive margin. Our results reveal that the ratio of women to men in the distribution of competitiveness decreases as the degree of competitiveness increases. In the upper tail of the distribution of competitiveness, we find that the women to men ratio is substantially smaller than the ratio detected by previous experimental paradigms. Of all the participants in the top 25 percent of the competitiveness, only 5 percent are women. All the participants on the top 10 percent are men. These insights can help explaining the lack of women in top corporate positions. If successful careers in some segments of the labor market require high levels of competitiveness, we can reasonably project that a weaker preference for competition may lead fewer women to commit to such a career path.

1. Bribery: Greed versus Reciprocity

Abstract

It is estimated that over a trillion dollars are exchanged in bribes around the world, distorting justice and economic efficiency. Better understanding of the reasons for bribery can help the effort to reduce it. We designed an experiment in which two participants compete for a prize. A third participant acts as a referee and picks the winner out of the two. Participants are allowed to send a bribe to the referee. When the referee can keep only the winner's bribe, we find substantial bribery, and in 86% of the cases, the participant who bribes more wins. However, when the referee keeps the bribes regardless of her choice of a winner, participants bribe less and referees are significantly less likely to ignore quality and award the prize to the worker with the higher bribe. We find similar results using an extra-laboratory experiment in a market in India. Hence, our participants are easy to corrupt, and the mechanism by which bribes work in our experiment is greed and not reciprocity.

1.1 Introduction

Bribery affects economic activities around the world. Because it is illegal in most places, getting good empirical data about these activities is difficult. However, the existing data show bribery is likely widespread. The World Bank estimates that $1 trillion exchanges hands in bribes annually (Kaufmann, 2005), and many companies report having to pay bribes to win business—from 15% to 20% in industrialized countries, to 40% in China, Russia, and Mexico (Transparency International, 2011). In some places, these kinds of activities are a major source of income. For example, bribes are estimated to amount to 20% of Russia's GDP in 2005 (INDEM, 2005).

But why do bribes "work?" In particular, if one of the sides in a bribery case does not fulfill his part, the other side cannot take him to court or use traditional enforcement mechanisms. What prevents one, for example, from accepting a payment but then not providing the good? If receiving the bribe is credibly contingent on success (e.g., winning a contest or in the case of repeated interactions), traditional economic models with selfish agents can explain behavior. In other one-shot cases in which receiving a bribe is not contingent on delivering the desired outcome, traditional economic assumptions may not be sufficient. In these cases, social preferences may be able to explain the success of bribery. People might be engaged in reciprocal behavior in which one side gives a "gift" and the other reciprocates (Akerlof, 1982; Rabin, 1993; Fehr and Gächter, 2000).

Consider the case of Rod Blagojevich, the former governor of Illinois. When Barack Obama was elected president in 2008, he had to give up his seat in the Senate,

and governor Blagojevich was in charge of finding a temporary replacement. Though the duration of such temporary appointments varies across states and situations, Blagojevich had the ability to appoint a new senator until the next general elections, which took place two years later. This situation was a unique, non-repeated instance. Instead of choosing the best candidate, the governor tried to "sell the senate seat," as US District Attorney at the time Patrick Fitzgerald said, to the highest bidder. Among the things Blagojevich asked for were a large salary at a labor union, a paid position for his wife on corporate boards, and promises for campaign funds (Fitzgerald, 2008). If legislation can reduce the possibility of contingent awards, for example, by preventing donations to public officials and their campaign funds, the success of bribes in one-shot situations will have to depend to a larger extent on trust. After all, once the position is filled, a candidate who bribed the governor cannot complain to court that he did not get the job in return. In this case, for bribes to succeed, reciprocity will be important.

To reduce bribery, it is important to understand what drives it. In the process of understanding the motivation for bribery, experiments are an important tool because they can help us isolate key aspects of the relevant behavior. Our paper distinguishes between the two rival motivations for bribery discussed above: reciprocity and greed (i.e., payoff maximization). This distinction is important from a public policy perspective. If reciprocity drives bribery, policy interventions should focus on reducing social ties and making reciprocity more difficult, for example, by decreasing personal contact through anonymity and staff rotation. If greed drives behavior and

individuals only care about maximizing their profit, such policy interventions will not prevent people from engaging in bribery, and traditional anti-corruption methods based on auditing and sanctions may be more effective (Becker and Stigler, 1974; Olken, 2007).

Our ability to distinguish between different motives for bribery comes from the novel game we study. The game captures an important feature that distinguishes bribery from other transactions: a distortionary effect. This kind of distortion is a key element in bribery and occurs when a decision maker uses bribes rather than other objective criteria such as merit, performance, or quality to determine who receives a particular outcome. As a result, public resources may go to the more corrupt people, not necessarily the most talented ones (Pareto, 1896; Goldsmith, 1999; Del Monte & Papagni, 2001). A large empirical literature has shown that such outcomes have detrimental effects on efficiency (see, e.g., Mauro, 1995; Reinikka and Svennson, 2004; Bertrand, Djankov, Hanna, and Mullainathan, 2007; Sequeira and Djankov, 2014; or see Olken and Pande, 2012, for a review).

We are not the first to study bribery using experiments: the existing experimental literature examines different elements of bribing behavior, from the effect of staff rotation (Abbink, 2004) and asymmetric liability (Abbink, Dasgupta, Gangadharan, and Jain, 2014) to culture (Barr and Serra, 2010; Cameron et al., 2009) and the influence of wages (Abbink, 2005; Armantier and Boly, 2013; Van Veldhuizen, 2013). See Abbink and Serra (2012) for a comprehensive survey of these experiments. However, in this literature, participants are asked to choose between

different monetary allocations. These decisions may include negative externalities on a third party, but they do not include a distortion of facts or judgment. The ability to study the effect of distortion of judgment is, as we show, critical to understanding some bribery behaviors.[1]

To capture this key element, we introduce a new bribery game in which two participants ("the workers") compete on a task. A third participant, the referee, then chooses the winner, who gets a prize. Importantly, the judgment of the quality of the task is subjective. Apart from working on the task, the two workers can also choose to send bribes to the referee. We use this basic design to test whether workers actually send money in an attempt to influence the referee. When workers choose to send money, we investigate whether these bribes distort the referee's judgment. We also vary whether the referee can keep both bribes or only the winner's bribe. This allows us to test whether the distortion is driven by reciprocity or greed, because whenever the referee is able to keep both bribes regardless of her decision, greed cannot influence her choice and only social preferences may drive behavior.

In addition to the laboratory experiments conducted in the United States (San Diego), we also report the results of an extra-laboratory experiment from a market in the city of Shillong in India. The data from a market in a country where corruption is

---

[1] Previous literature has shown that choices and judgment can be distorted by social pressure (Asch, 1954; Bond and Smith, 1996), self-serving biases (Babcock et al., 1995; Lord et al., 1979; Kunda, 1990; Haisley and Weber, 2010, see also Bazerman et al., 2002), or through a conflict of interest (Cain et al., 2005; Moore and Loewenstein, 2010).

more spread allow us to investigate whether our results generalize beyond the scope of the original laboratory experiments.

1.2 The Bribery Game and Research Questions

1.2.1 The Bribery Game

Our bribery game involves three players: two workers and a referee. The workers compete against each other on a task and the referee is asked to determine a winner. The worker who wins gets a prize of p, and the other worker receives nothing. Additionally, workers can send a bribe ($b_i \in [0, \frac{1}{2}p]$) to the referee, with only integer amounts allowed.

Our main identification relies on two versions of the basic game. In treatment KeepWinner, referees keep the bribe of the winning worker; the other worker's bribe is returned. The referee's monetary payoff maximizing strategy is then to choose the worker who submits the higher bribe. Assuming the referee chooses this strategy, and given the restriction that $b_i \leq 0.5p$, the workers' monetary payoff maximizing strategy is to bribe \$1 more than the other worker. This strategy results in a unique Nash equilibrium in which both workers bribe the maximum $b_i = 0.5p$. The referee's equilibrium payoff under these assumptions is $\Pi_R = b_{i*}$, where i^* is the winner of the round. The monetary payoff of each worker i is given by

$$\Pi_i = \begin{cases} -b_i + p & \text{if i wins} \\ 0 & \text{if i loses} \end{cases}$$

In the second treatment we study ("KeepBoth"), the referee (R) keeps both bribes, and the payoff for the referee in each given round is therefore given by $\Pi_R = b_i + b_j$. The monetary payoff of each worker i is given by

$$\Pi_i = \begin{cases} -b_i + p & \text{if i wins} \\ -b_i & \text{if i loses} \end{cases}$$

A monetary payoff maximizing referee will then be financially indifferent between both workers, irrespective of the bribes. The workers' payoff-maximizing strategy depends on their beliefs regarding how the referee will reward bribes. In particular, whenever a worker's belief that referees will select the worker with the higher bribe as the winner (*pRef*) is low enough, the best response will be not to bribe. For the parameters used in the experiment, when $pRef<.6$, workers' optimal strategy is not to bribe. For $.6 \leq pRef<1$, a mixed equilibrium exists in which workers bribe with some probability. For $pRef=1$, a pure strategy equilibrium exists in which both workers bribe the maximum. For a more detailed analysis, see Appendix A8.

This game allows us to study whether bribes induce referees to distort the true ranking between workers, resulting in an allocation of the prize based on bribes rather than performance, and to investigate which motives drive distortion.

Note that in our experiment, we focus on investigating bribery in situations where the judgment of the best performer is subjective. Further, in our experiment the referee's payoff depends only on the bribes and not on worker performance. These features of our design are reflective of many real-world situations in which judgment is subjective and it is difficult to directly reward good decisions, as discussed below.

The fact that the referee was not rewarded based on good judgment also allows us to cleanly disentangle greed, reciprocity, and moral costs of distortion, as discussed below. The game we use in the paper can be extended to study what happens in games in which the referee's payoff does depend on the quality of her decision.

Finally, to isolate the effect of distortion, we deliberately did not introduce other elements often associated with bribery, such as monitoring, punishment, and third-party externalities. Future research could use our bribery game to incorporate these additional features.

1.2.2 Research Questions

The two treatments described above help us in answering two important questions regarding bribery. First, we want to investigate workers' bribing behavior. Note that even if a worker believes that offering a high bribe pays in terms of monetary rewards, she may choose not to do so, because of some moral costs associated with unethical behavior. Studies have shown that such motives are important in related deception behavior (Gneezy, 2005; Dreber and Johannesson, 2008; Sutter, 2009; Erat and Gneezy, 2012).[2]

Second, we want to investigate whether bribing distorts the referee's judgment, and if so, how this distortion interacts with the treatments. In our game, the referee is asked to choose the winner based on the workers' performance on the task. Basing the

---

[2] See also Cappelen, Sørensen, and Tungodden (2013), Erat (2013), and Lightle (2013) for investigations of the factors that affect moral costs in deception behavior, and Belot and Schröder (2013), who examine the relationship between payment schemes and lying and theft in a principal agent setting.

decision instead on the size of the bribe leads to a distortion of the true ranking between workers. If individuals have some moral costs (e.g., lying costs) associated with distorting their judgment, they may choose to reward the better performer, and bribery would not influence their behavior.

Two important forces could explain why bribery affects judgment: reciprocity and greed. According to the reciprocity, or gift exchange, hypothesis (e.g., Abbink et al., 2002; Malmendier and Schmidt, 2012), if a worker sends money to the referee, the referee might want to reciprocate the favor by choosing to reward the worker who sent her (more) money. In this case, referees will choose the worker who sent the higher bribe, because they want to reciprocate the worker who was nicer to them, and not just because that bribe provides them with more money.[3] In contrast to the gift-exchange explanation, greed implies that referees choose the worker who bribes more only when doing so benefits them financially.

Comparing behavior in treatment KeepWinner with behavior in treatment KeepBoth allows us to test whether moral costs, gift exchange, or greed drive behavior. In treatment KeepWinner, a selfish payoff-maximizing referee would base her decision solely on the size of the bribes. Similarly, a referee who cares only about reciprocity will also choose the worker who sent the higher bribe. In treatment KeepBoth, the referee's choice of a winner does not affect her payment. Hence, a selfish payoff-maximizing referee will be indifferent between workers. A reciprocal

[3] When we refer to reciprocity in the paper, we refer exclusively to non-strategic, social-preference based reciprocity. Strategic reciprocity would require repeated interactions with feedback, which are not part of our design.

referee will still reward the higher bribe even when doing so does not affect her payoff.

If reciprocity drives the distortion of judgment, the distortionary effects of bribery will be similar in both treatments. If greed drives behavior, referees will distort their judgment in treatment KeepWinner, but in general not in treatment KeepBoth.

If referees also have moral costs, and—all else equal—prefer an allocation that does not require them to distort their judgment, these moral costs may outweigh greed and/or reciprocity concerns, preventing distortion of judgment in both treatments. However, if the power of greed or reciprocity in our experiment is large enough to outweigh moral costs, we expect to see some distortion. By comparing the two treatments, assuming moral costs do not change, we can rank the importance of greed and reciprocity.

A feature of our design is that although treatment is randomized and workers are randomly paired within sessions, bribes are not determined at random. To analyze referees' behavior and make treatment comparisons, referees in the two treatments must face similar combinations of bribes. That is, the distribution of the difference between the bribes the referees receive must be similar across both treatments. We will explore whether this is the case in the results section, where we also discuss how other possible differences in bribing behavior across the two treatments may affect referees' decisions.

1.2.3 Additional Treatments

Other than the two main versions of the basic game, we ran six additional treatments to provide additional support to our findings and rule out some alternative explanations. First, we take into account that, in many cases, the person being bribed (the referee in our game) may choose to reject the bribe. We consider the effect of such an option in treatments KeepWinnerReject and KeepBothReject.[4] These treatments are the same as treatment KeepWinner and treatment KeepBoth, respectively, except that referees also have the option to reject both bribes. Honest behavior may imply choosing a worker but rejecting his bribe. Adding the ability to reject even the winning bribe allows us to investigate how this option affects behavior. An additional interesting question is what would happen if we allowed referees to choose either to accept one, both, or none of the bribes. We leave this question for future research, because this treatment would have increased the complexity of the experiment without directly helping us in answering the central questions of the current paper.

The third additional treatment involves a higher wage for the referee (treatment HighWage). This treatment is similar to treatment KeepWinnerReject, except that referees receive a higher show-up fee ($20 instead of $5). This treatment allows us to test whether some sort of inequality preferences (see, e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) can explain our results. In treatment KeepWinner (and in

---

[4] Treatment KeepBothReject and treatment CoinFlip were conducted in October 2014. We thank the editor and two anonymous referees for suggesting we run these additional treatments.

all other treatments), the referee starts with less money than the workers ($5 vs. $10). Therefore, accepting the higher bribe will *decrease* inequity by making the referee's income more similar to both the winner's income and the loser's income. By contrast, in treatment HighWage, the referee starts with more money than the workers ($20 vs. $10). Therefore, accepting the higher bribe will now *increase* inequity. Thus, inequity aversion would predict referees would be less likely to let the higher bribe win in treatment HighWage than in treatment KeepWinner. In this treatment, we also included the possibility of rejecting bribes in order to give referees the possibility of decreasing inequity by choosing a winner without keeping either of the two bribes.

Our fourth additional treatment, Treatment NoTask, is identical to treatment KeepBoth, except that workers no longer compete on a task. Removing the task does not affect equilibrium predictions. In this case, choosing the higher bribe does not require the referee to distort judgment, and hence we can test whether moral costs of distortion are important, and whether in the absence of such distortion, gift exchange can account for the results. This treatment is more closely related to the existing bribery games in the literature. As discussed above, these studies do not capture the distortionary effect of bribes that our design introduces.

The fifth additional treatment involves a variation of treatment KeepWinner in which workers compete on a different task (treatment Objective). As we discuss in detail below, the main treatments of this paper use a subjective task. In treatment Objective, we replace this task with a more objective one. When evaluating a subjective task, distorting judgment could be easier because referees may convince

themselves that the worker with the higher bribe is also the better performer. If the task is more objective, convincing oneself that the worker with the higher bribe is the better performer could be more difficult, and as a result the moral costs associated with distortion of judgment may be higher. Thus, similar to treatment NoTask, treatment Objective tests the importance of moral costs, but it does so by increasing rather than decreasing them.

The last treatment, treatment CoinFlip, is a variation of treatment KeepBoth in which the referee has to decide between selecting the winning worker (as in the other treatments) or letting the winner be determined at random. In this treatment, choosing to select the winner is costly, which allows us to investigate whether referees are willing to incur a monetary cost to be reciprocal or choose the better performer. If greed is important, referees should choose to flip the coin, and will hence not be reciprocal or select the better joke.

In treatments KeepWinnerReject, HighWage, and Objective, the payoff-maximizing strategies and equilibrium under the assumptions of selfish profit-maximizing behavior are the same as for treatment KeepWinner. In the first case, no payoff-maximizing referee will ever reject a bribe; in the second case, the referee's income level is irrelevant; and in the third case, the nature of the task does not affect equilibrium predictions. Along the same lines, the equilibrium in treatments KeepBothReject and NoTask is the same as for treatment KeepBoth. In treatment CoinFlip, referees' payoff-maximizing strategy is to determine the winner at random; workers should best respond by bribing zero.

1.3 Experimental Design

1.3.1 Task

In all treatments except NoTask and Objective, we chose a task that involves creativity and for which the evaluation is not fully objective but depends partly on the referee's subjective taste. In particular, we asked workers to write a joke either about economists (round 1) or psychologists (round 2). All instructions can be found in Appendix B.

We chose to use a subjective task because in many real-life situations in which bribery is relevant, decision makers cannot exclusively rely on objective criteria when deciding how to allocate resources. In the Blagojevich example, the selection of the most qualified candidate for the Senate seat was partially based on the governor's subjective judgment. Procurement auctions are another example of these situations, because the decision to award a procurement contract to a certain supplier is based on both objective parameters (e.g., price, completion time), which can easily be observed, and partially subjective ones (e.g., esthetics), which are left to the auctioneer's discretion (Burguet and Che, 2004).

The task of judging jokes incorporates both the subjective and the objective component. In terms of the subjective component, humor is at least partially a matter of taste, so that for relatively similar jokes, different referees may have different opinions about which joke is the better one. However, when jokes differ enough in quality, one of them can also be objectively regarded as the better joke. We will see

below that referees were capable of selecting the (objectively) better joke in such cases. For an overview of some of the jokes written by participants, see Appendix C.

By contrast, in treatment Objective, workers were asked to work on a variation of the Stroop Task (Stroop, 1935). Each participant was shown a sequence of color words (e.g., blue, red, yellow) one after the other and was asked to identify the ink color of each word. We chose to use a congruent version of the task, meaning the color word and its ink color were compatible (e.g., blue was always written in blue letters). Participants were informed that their final score was equal to the number of words successfully identified, and were asked to complete as many words as possible.

Upon completion of the task, their final score was graphically represented on a score sheet in which every successfully identified word was represented by a dot (see Appendix B4 for the instructions and a sample score sheet). This procedure meant referees could count the dots to objectively determine which worker performed better, but had some moral wiggle room in that they needed to expend some effort to objectively evaluate workers' performance. We introduced the score sheet both to add some moral wiggle room and to give referees a non-trivial task to perform.

1.3.2 Procedure

We conducted the experiment in the laboratory of the Rady School of Management at the University of California San Diego with a total of 363 participants. Participants were recruited using standard recruitment procedures at the laboratory via an online experimental registration system. All UCSD students are able to register for this system to participate in laboratory experiments.

For each session, we recruited 10 participants to the laboratory. Every participant from the pool was notified about the sessions and was eligible to participate. Each session consisted of exactly six participants, and therefore any time more than six showed up, we randomly selected six and dismissed the extra participants after paying them a $5 show-up fee.

Each session lasted approximately 50 minutes. Upon being selected to participate, participants were randomly assigned to a computer station and were asked to follow the instructions on the screen. Participants were anonymously matched in groups of three, and each of them was either assigned to the role of workers (called participant A and B in the experiment) or the referee. We then moved the referees to separate rooms (one room for each referee), where they received the remainder of the instructions. Workers continued reading their instructions in the main lab. Neither workers nor referees knew which of the other participants were matched with them.

We then informed participants (except those in the NoTask treatment) about the task and the referee's role in determining the winner. In treatment NoTask, we informed participants about the referee's role in determining the winner but did not ask them to complete any task. In all cases, neither the workers nor the referees were yet informed about the workers' opportunity to send money to the referee.

On their desks, workers had an envelope with their $10 show-up fee, in $1 bills. Each referee had an envelope with a $5 show-up fee in all treatments except treatment HighWage, in which the referee received an envelope with a $20 show-up

fee. The information about the other participants' initial show-up fees was made common knowledge.

After all workers read their instructions and completed some attention questions, they learned the topic of the jokes for the first round ("economists") and had 10 minutes to type a joke (in the NoTask treatment, workers were told to wait 10 minutes; in the Objective treatment they had 5 minutes to work on the Stroop task). The experimenters then printed and returned each joke (or score sheet in the Objective treatment) to the workers. Workers received only their own joke or score sheet, and were not informed about the jokes or scores of the other workers in the experiment. While the experimenters were printing the jokes, we asked workers to state their expected likelihood of having a better joke than their opponent ("What do you believe is the probability that you will have a better joke than your opponent?").

The workers then received a second set of instructions on the screen, which notified them of the opportunity to send money to the referee. In particular, workers were asked to put the printed copy of their joke (or score sheet in treatment Objective) in a large envelope labeled with their participation ID, and were given the opportunity to add up to $5 of their show-up fee to the envelope. Meanwhile, the referees also received a second set of instructions telling them about the possibility of workers sending them money.

After all workers had prepared their envelopes, an experimenter collected them, recorded the monetary content of each, and gave the envelopes to the referees. Upon receiving the envelopes, each referee had five minutes to rate on a scale of 0 to

10 the quality of the workers' jokes (except in treatment NoTask and treatment Objective), and to place a winner card and a loser card in the winner and loser's envelope, respectively. After five minutes, the workers returned the envelopes to the experimenter, who then recorded the referees' decisions.

For treatment CoinFlip, referees who were willing to pay $1 to determine the winner themselves were asked to pay the experimenter after they determined the winner using the same procedure as in the other treatments. Referees who wanted the winner to be determined randomly were told to notify the experimenter at the end of the five minutes and to ask the experimenter to flip the coin for them, at no extra cost.

In treatments KeepWinner and Objective, the referee could keep only the winner's monetary transfer and had to return the loser's money by putting it back in the envelope. In treatments KeepWinnerReject and HighWage, the referee had to return any money received by the loser, but were asked to decide whether to keep the winner's money or return both bribes. In treatments KeepBoth, NoTask, and CoinFlip, the referee kept all the money sent by both workers, whereas in treatment KeepBothReject, the referee also had the option to return both the winner and the loser's money. Table 1 summarizes the experimental treatments. Note that we have 60 participants (20 groups) in the two main treatments, KeepWinner and KeepBoth, and in treatment HighWage. For treatment Objective, we have 63 participants (21

groups).[5] For the four remaining additional treatments, we have 30 participants (10 groups) in each treatment. The experiment consisted of two rounds with the same matching of participants. To prevent referees from reciprocating the largest bribe in round 1 for strategic reasons, no feedback was provided between rounds. Workers started the second round while the referees were still evaluating their first round. The procedure for round 2 was identical to that of round 1, apart from the topic of the joke. After the second round, both workers and referees were asked to complete a survey of basic demographic information. The referees were then paid and left the experiment, and workers received back the envelopes for rounds 1 and 2. Each envelope contained either a winner or a loser card indicating the referees' decision. For treatments KeepWinner, KeepWinnerReject, HighWage, and Objective, the envelope with the loser card also contained any money sent to the referee by the worker who lost. For treatments KeepWinnerReject, KeepBothReject, and HighWage, both envelopes could also contain money returned by the referee if the referee decided to reject both bribes. Workers were then paid $10 for each winner card they had and left the experiment.

[5] In treatment Objective, we had to discard one group because the referee did not follow the instructions and rejected both bribes even though he was not allowed to do so. Therefore, we ran one additional session in order to have at least 20 groups.

Table 1.1 The Bribery Game Experimental Treatments

	Which bribe does the referee keep?	Task	Participants	Ref. show-up fee	Who determines the winner
KeepWinner	Only winner's	Jokes	60	$5	Referee
KeepBoth	Both	Jokes	60	$5	Referee
KeepWinnerReject	Chooses whether to keep winner's	Jokes	30	$5	Referee
KeepBothReject	Chooses whether to keep both	Jokes	30	$5	Referee
HighWage	Chooses whether to keep winner's	Jokes	60	$20	Referee
NoTask	Both	No	30	$5	Referee
Objective	Only winner's	Objective	63	$5	Referee
CoinFlip	Both	Jokes	30	$5	Referee or coin flip

1.3.3 Joke Quality

After the experiment was completed, we organized additional sessions in which participants from the same participant pool who had not previously participated in the experiment evaluated the quality of several pairs of jokes. The jokes were evaluated by a total of 792 raters who, for each pair of jokes, had to evaluate the quality of each joke (on a scale from 0 to 10) and had to determine which joke is funnier. Raters were shown the same pairs of jokes the referees evaluated during the experiment, without being informed about the bribes sent by the workers. This procedure provides us with a more objective measure of joke quality, which is not biased by the presence of bribery. Each rater evaluated up to six pairs of jokes, chosen at random by an electronic randomizer among all the possible pairs of jokes. Each independent rater was presented with up to six pairs of jokes selected at random. Each

pair of jokes was evaluated by an average of 22 independent raters. The full instructions are in Appendix B3.

1.4 Results

Table 2 presents some descriptive statistics on our sample. As the table shows, the treatments are balanced with respect to demographics. Joke quality and confidence levels are also not statistically different between treatments and rounds (Bonferroni or Holm-Bonferroni correction for multiple hypothesis testing).

In the remainder of this section, we will use both parametric and non-parametric tests to test for differences between treatments. Whenever we analyze worker behavior, we use one worker as one independent observation; whenever we analyze referee behavior, we use one referee as one independent observation. For non-parametric tests involving data from both rounds, we therefore take the average over both rounds as the unit of observation. In the remainder of this section, we first analyze worker and referee behavior in the two main treatments, KeepWinner and KeepBoth. We then discuss the additional treatments to investigate the robustness of our results and address potential alternative explanations.

1.4.1 Do Workers Bribe?

Figure 1 shows the distribution of bribes in the KeepWinner treatment for both rounds. The first thing to note is that workers did bribe: 41% of bribes were at the maximum $5 and a further 33% of bribes were positive. In 26% of the cases, workers elected not to send a bribe. Overall, the average bribe was $2.80.

Table 1.2 Descriptive Statistics

	Overall	Keep Winner	Keep Both	KW Rej	KB Rej	High Wage	No Task	Objective	Coin Flip
Joke Quality (Round 1)	3.60 (1.18	3.57 (1.10)	3.50 (1.17)	3.47 (1.28)	3.81 (1.4	3.78 (1.11)			3.45 (1.22)
Joke Quality (Round 2)	3.58 (1.26	3.73 (1.32)	3.27 (1.43)	3.39 (1.45)	3.89 (1.1	3.51 (.98)			3.92 (1.13)
Objective Score (Round 1)	176 (15)							176 (15)	
Objective Score (Round 2)	179 (17)							179 (17)	
Worker (Round 1)	.51 (.26)	.48 (.28)	.53 (.27)	.49 (.26)	.55 (.24)	.41 (.24)		.62 (.19)	.53 (.31)
Worker (Round 2)	.49 (.25)	.44 (.28)	.56 (.26)	.41 (.30)	.53 (.28)	.39 (.15)		.59 (.22)	.53 (.27)
Psychology	.15 (.29)	.13 (.33)	.12 (.32)	.13 (.35)	.03 (.18)	.18 (.39)	.17 (.38)	.14 (.35)	.03 (.18)
Economics	.25 (.43)	.20 (.40)	.18 (.39)	.23 (.43)	.43 (.50)	.25 (.44)	.37 (.49)	.16 (.37)	.33 (.48)
Other Social	.07 (.26)	.07 (.25)	.07 (.25)	.03 (.18)	.07 (.26)	.08 (.28)	.10 (.31)	.10 (.29)	.07 (.25)
Biology/Chemist	.26 (.44)	.35 (.48)	.25 (.44)	.43 (.50)	.23 (.43)	.15 (.36)	.23 (.43)	.25 (.44)	.27 (.45)
Engineering/Scie	.20 (.40)	.20 (.40)	.22 (.42)	.17 (.38)	.17 (.38)	.22 (.42)	.07 (.25)	.24 (.43)	.30 (.47)
Humanities	.08 (.27)	.05 (.22)	.15 (.36)	.00 (.00)	.07 (.25)	.10 (.30)	.07 (.25)	.11 (.32)	.00 (.00)
Undeclared	.01 (.07)	.00 (.00)	.02 (.13)	.00 (.00)	.00 (.00)	.02 (.13)	.00 (.00	.00 (.00)	.00 (.00)
Asian Ethnicity	.71 (.46)	.63 (.49)	.82 (.39)	.83 (.38)	.70 (.47)	.72 (.45)	.60 (.50)	.71 (.46)	.60 (.50)
Female	.55 (.50)	.55 (.50)	.60 (.49)	.50 (.51)	.47 (.51)	.57 (.50)	.57 (.50)	.57 (.50)	.43 (.50)
Nonnative	.16 (.37)	.15 (.36)	.18 (.39)	.13 (.34)	.23 (.43)	.15 (.36)	.20 (.40)	.13 (.33)	.13 (.34)
Age	20.7 (1.93	21.1 (2.62)	20.5 (1.46)	20.6 (1.65)	20.3 (2.1	20.6 (1.81)	21.0 (1.30)	20.7 (1.89	20.7 (2.02)
Observations	363	60	60	30	30	60	30	63	30

Notes: Descriptive statistics. Joke quality is the average rating of the joke by the independent raters. Objective performance is the score on the objective task for treatment Objective. Confidence is the worker's confidence in having a better joke or performance than the other worker. The remaining variables are dummies for the respective majors, Asian participants, females, and nonnative speakers, and a continuous variable for age, respectively. Among the nonnative speakers, 12% are Chinese native speakers, 2% are Spanish native speakers, and the remainder report different languages.

As Figure 1 also shows, the workers in treatment KeepBoth bribed less than the workers in treatment KeepWinner, in which no bribe was sent in 66% of cases. Overall the average bribe in this treatment was \$0.90. The difference in the

distribution of bribes between the KeepWinner and the KeepBoth treatments is significant (p<.0001, Mann-Whitney).

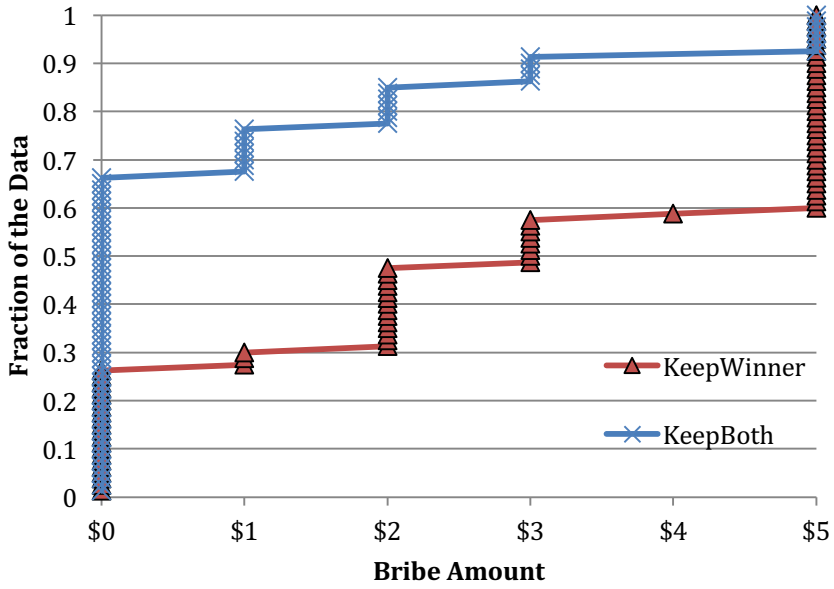

Figure 1.1 CDF of Bribes for KeepWinner and KeepBoth

Table 3 reports the distribution of bribes per treatment and the average bribes per round for all treatments. On average, in the KeepWinner treatment, bribes did not change between rounds: the average bribe was $2.83 in the first round and $2.78 in the second round (the difference is not statistically significant). In treatment KeepBoth, the average bribe was $1.13 in round one and $0.68 in round two and this difference is marginally significant (p=.056, Wilcoxon signed-rank test).[6] Additionally, workers

[6] However, this effect is only driven by a few observations. Thirty of the 40 workers bribed exactly the same in the second round, two people bribed more, and the remaining eight people bribed less.

who bribed more in round 1 were also likely to bribe more in round 2 (r=.52, p=.005 in KeepWinner; r=.60, p<.001 in KeepBoth).

Table 1.3 Bribes across Treatments

	Overall	Keep Winner	Keep Both	KW Reject	KB Reject	High Wage	No Task	Objective	Coin
Fraction of Bribes: all Rounds									
Bribe=0	40%	26%	66%	13%	58%	53%	25%	36%	63%
Bribe=1	6%	4%	10%	3%	18%	5%	5%	6%	15%
Bribe=2	13%	18%	9%	18%	10%	9%	15%	12%	10%
Bribe=3	9%	10%	6%	13%	5%	4%	18%	10%	5%
Bribe=4	2%	1%	0	3%	3%	1%	15%	1%	0%
Bribe=5	30%	41%	9%	53%	8%	29%	23%	36%	8%
Average Bribe									
Round	2. 12 (2.11)	2.83 (2.07)	1.13 (1.76)	3.35 (1.92)	1.25 (1.65)	2.08 (2.34)	2.95 (1.79)	2.36 (2.15)	1.10 (1.68)
Round	1.86 (2.08)	2.78 (2.13)	.68 (1.31)	3.60 (1.79)	.75 (1.41)	1.58 (2.07)	2.25 (1.97)	2.48 (2.22)	.65 (1.23)
Both	1.99 (2.09)	2.80 (2.09)	.90 (1.56)	3.48 (1.84)	1.00 (1.45)	1.83 (2.20)	2.60 (1.89)	2.42 (2.17)	.88 (1.47)
N: per	202	40	40	20	20	40	20	42	20
N: both	404	80	80	40	40	80	40	84	40
Average Difference in Bribes (excluding equal bribes)									
Round	3.14 (1.47)	3.07 (1.59)	2.92 (1.56	2.14 (.69)	2.83 (1.72)	4.21 (1.31)	2.63 (1.30)	3.40 (1.24)	2.86 (1.68)
Round	2.90 (1.51)	3.00 (1.36)	2.30 (1.64)	3.00 (1.07)	2.60 (1.82)	3.42 (1.73)	3.00 (1.41)	3.29 (1.54)	1.86 (1.47)
Both	3.03 (1.49)	3.03 (1.45)	2.64 (1.59)	2.60 (.99)	2.73 (1.68)	3.85 (1.54)	2.82 (1.33)	3.34 (1.37)	2.36 (1.60)

Notes: The table gives the relative frequency of bribes of different sizes in the upper panel. The middle panel displays average bribe size (over all workers) separately as well as jointly for each treatment and round. The lower panel displays the average difference in bribes (over all workers) separately as well as jointly for each treatment and round. Average bribes are computed using all bribes, including zeros. Average differences in bribes are computed by subtracting the highest bribe from the lowest bribe in a given pair of bribes and are based only on observations in which the two bribes were not identical. The numbers in brackets are standard deviations.

Determinants of Bribes—Next, we investigate whether worker-level characteristics are predictive of bribe size. For example, workers who wrote inferior jokes (or performed worse on the objective task) might have sent higher bribes. Similarly, some of the

demographic variables reported in Table 2 might be predictive of how much a worker decides to bribe.

To check which variables are predictive of bribe size, we regressed bribe size on performance on the task, the confidence question, and all of the demographic variables reported in Table 2, where we use biology/chemistry majors as the reference group. We pooled the data from all treatments to have the largest possible sample size. The regression results are reported in Table A1 in the Appendix.

Overall, we find that the coefficient for joke quality (or the performance on the objective task) is not significantly different from zero (for joke quality: β=0.03, p=0.76), suggesting that overall the quality of workers' performance did not affect their bribing behavior. Further, the coefficient for workers' beliefs about having a better joke than the opponent is also not significant (β=-0.02, p=0.93). These coefficients remain insignificant even if we only include either actual quality or the beliefs in the regression.

Our analysis further reveals that non-native speakers (β=0.84, p=.004), older participants (β=0.11, p=.062), and men (β=0.45, p=.064) send higher bribes, whereas social science majors (not including economists) send smaller bribes (β=-0.77, p=.084). For more details and for an additional analysis of the determinants of quality, see Appendix A1.

Differences in Bribes—To analyze referees' behavior and make treatment comparisons, referees in the two treatments must face similar combinations of bribes. That is, the distribution of the difference in bribes between the two workers who

competed against each other must not differ across treatments. The lower panel of Table 3 shows the average difference in bribes by treatment. We include only cases in which the bribes are not identical, because these are the observations that allow us to disentangle greed from reciprocity. We find that the average difference in bribes is similar in treatment KeepWinner ($3.03) and treatment KeepBoth ($2.64, p=.811 for round 1, p=.191 for round 2, p=.482 for both rounds combined) and so is the distribution of bribes (p=1, Kolmogorov-Smirnov for both rounds combined). Thus, we confirm that referees indeed faced similar financial tradeoffs in both treatments, which allows us to make treatment comparisons.

Further, as compared to referees in treatment KeepWinner, referees in treatment KeepBoth were more likely to receive two identical bribes, less likely to face two positive bribes, and less likely to receive two large bribes. We address the former concern in our regression analysis by examining the cases in which bribes differ separately from the cases where bribes are identical. We discuss the latter two differences in Appendix A2 and show that neither of them affects our main results. In particular, the results suggest referees behaved similarly irrespective of the absolute size or the number of bribes received.

1.4.2 Does Bribery Distort the Referee's Judgment?

Joke Quality—To investigate whether bribery results in a distortion of the referee's judgment, we use the evaluation provided by the independent raters as an unbiased measure of joke quality. We will focus on two measures of quality: the difference in average rating between the two jokes in a pair and the fraction of raters

that, for a given pair, chose the same joke as the better one (i.e., the degree of agreement across raters). The two measures are highly correlated ($r=.92$, $p<.001$ for all joke treatments combined).

As with bribes, in our analysis, we look at quality *differences* between the jokes written by the two workers in a given pair. The distribution of differences in quality does not differ between KeepWinner and Keepboth ($p=.739$, Kolmogorov-Smirnov).

With the non-parametric tests, we investigate whether the worker with the better joke in the pair won the prize. Because joke quality is subjective, it is not enough for one joke to have a slightly higher quality on either of the two quality measures. Instead, we need to know whether one joke is *significantly* better than the other. For this purpose, we consider all joke pairs for which at least 65.1% of the independent raters agreed on the winner. With this threshold, the fraction of independent raters who selected a given joke over the other is significantly different from chance (i.e., 50%) at the 10% level ($z=1.28$, $p=.1$, test of proportions for our minimum of 18 independent raters). By this threshold, 63% of joke pairs have a significantly better joke.

In the remaining pairs, the quality of the two jokes was too similar to be statistically distinguishable. In such cases, picking one joke over the other did not constitute a big distortion. Whenever we refer to better-quality jokes in subsequent non-parametric tests, we will only use jokes that are sufficiently different by this criterion. Appendix A3 presents more details on this threshold and also shows that for

a threshold of 69.4%, which corresponds to jokes being significantly different at the 5% level, the main results are similar.

For the Objective treatment, we use participants' actual scores as the performance measure. Similar to the other treatments, we omit the 37% least distinguishable performance pairs whenever we refer to better-quality performers. In practice, this approach means we only look at pairs in which the difference in performance was at least 11 words.

The KeepWinner Treatment—Did bribing result in a distortion of the referees' judgment? In 86% of the cases, referees in the KeepWinner treatment chose the worker who offered the higher bribe as the winner. This number is significantly larger than chance (p=.001, Wilcoxon). By contrast, as Figure 2 shows, the better joke (as judged by the independent raters) won only 57% of the time, which is not significantly different from chance (p=.564, Wilcoxon). Thus, these results suggest that bribery distorted referees' judgment, because they chose the worker who paid them more, not the one who wrote the funnier joke.

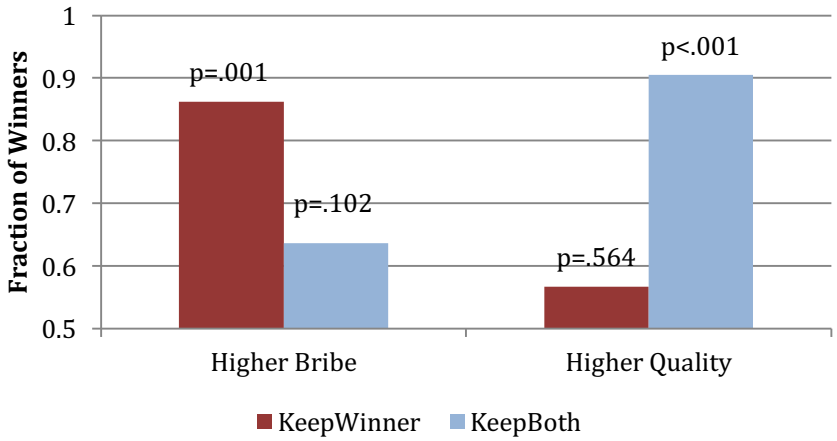

Figure 1.2 Win Chance for Higher Bribe or Better Rating

We further investigate the effect of bribes and quality using OLS. In the regression, we examine how differences in joke quality and bribes between the two workers affect the referee's decision. For a given worker, the regression tells us whether increasing her bribe or quality relative to the other worker increases her probability of winning. The more that referees care about quality relative to bribes, the more beneficial having a better joke should be.

On a more methodological level, because our independent variables are differences between the two workers within a given pair, the observations for the two workers are always the exact inverse of each other. Hence, for the regression, we randomly select one worker per round. We use the same random sample throughout the analysis. In Appendix A6, we show that the reported results do not depend on the particular random sample selected for the regression. Randomly selecting a worker also implies that selected workers on average win approximately 50% of the time; as a result, we do not report the constant in the regression table.

Further, to facilitate comparisons between quality and bribe coefficients, we standardize all independent variables, such that the coefficients represent the effect of a one-standard-deviation increase in the independent variable. For quality, we estimate separate coefficients for cases in which the two bribes are identical and cases in which they differ. The latter coefficient is of particular interest because it allows us to examine the effect of quality when referees could also be influenced by bribes. The former coefficient instead allows us to see whether quality is important when referees have no incentive to distort their judgment. Finally, we compute the p-values reported in the regression tables using a wild bootstrap procedure (Cameron, Gelbach, and Miller, 2008).[7]

Column (1) of Table 4 presents the results. The coefficient for bribes is large, positive, and statistically significant. Indeed, a one-standard-deviation increase in a given worker's bribe (relative to the other worker) increases her likelihood of winning the prize by 31 percentage points. By contrast, the coefficient for quality when bribes differ is small and not statistically significant. Thus, the regression results confirm that bribes, not quality, influenced referees in treatment KeepWinner.

The regression results also show that when bribes are identical, increasing the quality of a given worker's joke (relative to the other worker) does significantly increase her likelihood of winning. This finding shows that despite the subjective

---

7 We thank an anonymous referee for suggesting this method to us. Cameron, Gelbach, and Miller (2008) show that for small cluster sizes and/or a small number of clusters, this approach leads to more accurate (and more conservative) p-values than alternative techniques. In Appendix A5, we also redo all regressions of Table 4, using several alternative techniques, including clustered standard errors and non-parametric bootstraps.

nature of the task, referees were indeed capable of identifying the higher-quality joke in the absence of distortionary incentives.

Finally, the results are similar if we perform this regression separately for each round; see Appendix A7. In Appendix A4, we additionally estimate similar regressions that incorporate only those pairs of workers who had different bribes or wrote sufficiently different jokes. Further, we present the results of alternative specifications that include separate coefficients for bribes depending on whether joke quality was similar or significantly different. The results of the additional analyses are in line with the results reported in Table 4.

The KeepBoth Treatment—Figure 2 and column (2) of Table 4 give an overview of the referees' behavior in the KeepBoth treatment. In 64% of the cases, referees chose the worker who offered the higher bribe as a winner. This number is not significantly larger than chance ($p=.103$, Wilcoxon). By contrast, the better joke, as judged by our independent raters, won 90% of the time. This proportion is significantly larger than chance ($p<.001$, Wilcoxon). In other words, when the referees' payoff did not depend on the choice of winner, bribery did not distort judgment, and referees chose the worker who wrote the funnier joke.

Table 1.4 OLS Regressions for Referees in KeepWinner and KeepBoth

Dependent Variable:	Winner (1=Yes)		
	(1)	(2)	(3)
Bribe Difference	.308***	.086	.274***
	(.000)	(.140)	(.000)
Quality Difference (bribes	.014	262**	.015
	(.762)	(.010)	(.762)
Quality Difference (bribes	.336***	.275**	.255**
	(.020)	(.022)	(.020)
$D_{KeepBoth}$			.008
			(.980)
Bribe Difference X $D_{KeepBoth}$			-.173**
			(.032)
Quality Difference X $D_{KeepBoth}$ (bribes differ)			.222**
			(.014)
Quality Difference X $D_{KeepBoth}$ (bribes identical)			.150
			(.156)
Treatment	KeepWinn	KeepBoth	KeepWinn KeepBoth
Selected Workers	Random	Random	Random
Observations	40	40	80
Clusters	20	20	40

Notes: OLS estimates (p-values). The dependent variable is a dummy that specifies whether the selected worker was selected as the winner. Quality Difference is the difference between the quality of the joke (i.e., the average rating by the independent raters) of the selected worker and the quality of the joke of the other worker in the pair. Bribe Difference is the difference between the bribe sent by the selected worker and the bribe sent by the other worker in the pair. $D_{KeepBoth}$ is a dummy that is equal to one for treatment KeepBoth, and zero otherwise. For column (3), the bribe variable and both quality variables are standardized using the respective variable's combined standard deviation over all included treatments. P-values are calculated using wild bootstraps. For each regression, we randomly select one worker per referee in each round.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

The regression results are similar. Column (2) of Table 4 shows that increasing the quality of a given worker's joke (relative to the other worker) significantly increased her likelihood of winning, whereas increasing the worker's bribe did not pay off. In addition, the effect of joke quality on the likelihood of winning was similar both when bribes were identical and when bribes differed. This finding confirms that referees chose the worker who wrote the funnier joke in this treatment.

Comparison between KeepWinner and KeepBoth—Figure 2 shows that having a higher bribe was more effective in the KeepWinner treatment (86% vs. 64%, p=.048; Mann-Whitney), whereas having a better joke was more effective in the KeepBoth treatment (90% vs. 57%, p=.004; Mann-Whitney). The latter effect is driven exclusively by cases for which bribes were different. When bribes were equal, referees *in both treatments* (80% for KeepWinner, 78% for KeepBoth) picked the better joke (p=.871; Mann-Whitney). When bribes were *unequal*, referees in KeepBoth selected the better joke 100% of the time, compared with only 45% in KeepWinner (p=.003, Mann-Whitney). When we look specifically at cases in which the better bribe corresponds to the inferior joke, referees in KeepWinner selected the better bribe 82% of the time, whereas referees in KeepBoth selected the better joke (and lower bribe) in all cases (p=.007, Mann-Whitney, see Appendix A9).

These findings are confirmed by the regression of column (3) in Table 4, where we included data from both treatments and interacted the quality and bribe variables with a dummy for treatment KeepBoth. The interaction terms confirm that when bribes are unequal, the effect of quality is significantly larger in treatment KeepBoth, whereas the effect of bribes is smaller. Further, in cases in which bribes were equal, the importance of quality was approximately the same in both treatments.[8]

---

[8] The reason that the three non-interacted coefficients in column 3 of Table 4 are not exactly identical to the coefficients in column 1 is standardization. For column 1, we standardized all coefficients with respect to the standard deviation of the explanatory variables in KeepWinner, whereas for column 3 we used the standard deviation for the combined data of KeepWinner and KeepBoth.

Although differences in the bribes received were similar across treatments, referees in KeepWinner were more likely to receive two relatively large bribes and less likely to receive only one bribe. In Appendix A2, we therefore additionally investigate whether a similar difference in bribes becomes less (or more) important as the size of the bribes increases. We also investigate whether referees respond differently to receiving one versus two bribes. The results suggest that referees behaved very similarly irrespective of the absolute size or the number of bribes received.

Another interesting piece of information from our data comes from referees' evaluations of the quality of the two jokes. In particular, we asked referees in all the joke treatments to rate the quality of both jokes on a scale from 0 to 10. This measure was not incentivized, so participants had no incentive to lie. Whereas in treatment KeepBoth, referees' evaluation of the jokes is highly correlated with the independent raters' quality measure ($r=.513$, $p<.001$, OLS), in treatment KeepWinner, this correlation is smaller and not significant ($r=.147$, $p=.194$, OLS). This treatment difference in the accuracy of referees' evaluations of quality is significant ($p=.034$, OLS), which suggests the referees in treatment KeepWinner may have tried to rationalize their choice ex post and that the bribes distorted their quality evaluations.

Overall, the results show that referees awarded the prize to the worker with the higher bribe in treatment KeepWinner, but selected the one with the better joke in treatment KeepBoth. This finding is in line with the greed explanation of bribery. When referees are motivated by greed (treatment KeepWinner), they distort their

judgment. However, when only reciprocity could lead referees to select the higher bribe (treatment KeepBoth), they instead select the better joke. This observation suggests greed is more important than moral costs, which is in turn more important than reciprocity.

1.4.3 Additional Treatments

In this section, we present the results for the six additional treatments. To facilitate comparisons across treatments, we pool the data from all treatments into a joint regression in Table 5. We use KeepWinner as the reference treatment and interact the bribe and quality difference variables with treatment dummies for all other treatments. This approach allows us to verify whether bribes and quality played a larger or smaller role than in treatment KeepWinner. The corresponding figures are presented in Appendix A9.

The Reject Treatments—The results of KeepWinnerReject and KeepBothReject suggest that allowing referees to reject bribes does not change their behavior. In the KeepWinnerReject treatment, referees chose not to keep the winner's bribe only in 10% of the cases. As a consequence, the fraction of winners with the higher bribe (100% vs 86%, p=.176, Mann-Whitney) or with the funnier joke (56.7% vs. 57.1%; p=.978, MW) are similar to those in treatment KeepWinner.

In treatment KeepBothReject, referees also chose to reject both bribes in 10% of the cases. As a result, the results are similar to those of treatment KeepBoth: 86% of the referees awarded the prize to the better joke (vs. 90% in KeepBoth; p=.626, Mann-Whitney). Referees were less likely to award the prize to the worker with the higher

bribe than in treatment KeepBoth (36% vs. 64%; p=.043, Mann-Whitney), though this effect disappears when controlling for quality (as in Table 5).

The regression results are presented in Table 5. For treatment KeepWinnerReject, the only difference relative to KeepWinner is that having a larger bribe than the opponent increases the likelihood of winning significantly more, which is due to the fact that referees let the higher bribe win 100% of the time in this treatment. For treatment KeepBothReject, the coefficients for bribes and quality are similar in size to those of treatment KeepBoth, and significantly different from those of treatment KeepWinner. Overall, the results suggest that allowing referees to reject bribes did not affect their behavior.

Further, allowing referees to reject bribes did not affect worker behavior. There are no significant differences in average bribes for either of the two treatments (KeepWinnerReject = 3.48 vs. KeepWinner 2.80, p=.188, Mann-Whitney; KeepBothReject = 1.00 vs. KeepBoth = 0.90, p=.343, Mann-Whitney).

The HighWage Treatment—In KeepWinner, inequity aversion predicts that referees should accept the higher bribe, whereas in HighWage, it predicts the opposite. However, the higher bribe won 88% of the time, compared with 86% in KeepWinner, and the better-rated joke won 44% of the time, compared with 57% in KeepWinner. Neither difference is significant in either non-parametric tests or in Table 5. Additionally, the referee chose not to keep the winner's bribe in only 7.5% of the cases. These results suggest referees in this treatment behaved similarly to referees in treatment KeepWinner, contrary to the prediction of inequity aversion.

The only difference between the HighWage and the KeepWinner treatment relates to workers' behavior: their average bribe was significantly lower in HighWage than in KeepWinner ($1.80 versus $2.80, p=.020; Mann-Whitney).

The NoTask Treatment—Why are higher bribes ineffective in the KeepBoth treatment? One hypothesis is that the moral costs of distorting judgment are stronger than reciprocity, and hence referees will ignore the bribes. In treatment NoTask, we removed all moral costs of distortion by no longer asking referees to judge performance on a task. Hence, we expect reciprocity to become more important.

This is indeed what we find. Whereas the higher bribe won in 64% of the cases in treatment KeepBoth, in the NoTask treatment, this fraction significantly increases to 94% (p=.044, Mann Whitney). Moreover, 94% is also significantly larger than chance (p=.011, Wilcoxon), and not significantly different from KeepWinner (p=.614, Mann Whitney). The regression analysis of Table 5 confirms these results.

Looking at workers' behavior (Table 3), we find the average bribe in the NoTask treatment ($2.60) was higher than in KeepBoth (p<.0001; Mann-Whitney) and similar to KeepWinner (p=.697; Mann-Whitney). Additionally, in this treatment, the fraction of workers who did not send any bribe is lower than in KeepBoth (25% vs. 66%).

The comparison between the KeepBoth and the NoTask treatment provides further evidence that distorting judgment by rewarding the higher bribe presented referees with a moral cost, which previous bribery games did not capture. Referees were happy to award the prize to the worker who sent them more money when

rewarding them did not require a distortion of their judgment, but not when it did. Thus, in the absence of moral costs, reciprocity guides referees' behavior. By contrast, in other treatments, the moral costs of distorting judgment seem stronger than the norm of reciprocity.

The Objective Treatment—Having a better performance was more effective in the Objective treatment than in the KeepWinner treatment (83% vs. 57%, p=.023; Mann-Whitney), whereas it was equally effective as in the KeepBoth treatment (83% vs. 90%, p=.312; Mann-Whitney). Having a higher bribe was neither less effective than in KeepWinner (76% vs. 86%, p=.443; Mann-Whitney) nor more effective than in KeepBoth (76% vs. 64%, p=.299; Mann-Whitney).

Table 1.5 OLS Regressions for Referees in Additional Treatments

Dependent Variable: Winner (1=Yes)	(1)
Bribe Difference	.295***
	(.006)
Bribe Difference X $D_{KeepBoth}$	-.186**
	(.032)
Bribe Difference X $D_{KeepWinnerReject}$	.186***
	(.004)
Bribe Difference X $D_{KeepBothReject}$	-.197
	(.126)
Bribe Difference X $D_{HighWage}$	-.083
	(.178)
Bribe Difference X D_{NoTask}	.059
	(.596)
Bribe Difference X $D_{Objective}$	-.224**
	(.028)
Bribe Difference X $D_{CoinFlip}$	-.344
	(.272)
Quality Difference (equal bribes only)	.232**
	(.020)
Quality Difference (Bribes Equal) X $D_{KeepBoth}$	.138
	(.156)
Quality Difference (Bribes Equal) X $D_{KeepWinnerReject}$	-.244
	(.320)
Quality Difference (Bribes Equal) X $D_{KeepBothReject}$	-.150
	(.474)
Quality Difference (Bribes Equal) X $D_{HighWage}$	-.207
	(.168)
Quality Difference (Bribes Equal) X $D_{Objective}$	-.081
	(.692)
Quality Difference (Bribes Equal) X $D_{CoinFlip}$	-.400***
	(.004)
Quality Difference (Different Bribes)	.014
	(.762)
Quality Difference (Different Bribes) X $D_{KeepBoth}$	.210**
	(.014)
Quality Difference (Different Bribes) X $D_{KeepWinnerReject}$	-.044
	(.416)
Quality Difference (Different Bribes) X $D_{KeepBothReject}$	.452**
	(.012)
Quality Difference (Different Bribes) X $D_{HighWage}$	.023
	(.842)
Quality Difference (Different Bribes) X $D_{Objective}$	.295**
	(.020)
Quality Difference (Different Bribes) X $D_{CoinFlip}$	-.218
	(.354)
Selected Workers	Random
Treatment Dummies	Yes
Observations	242
Clusters	121

Notes: OLS estimates (p-values). P-values are computed using wild bootstraps. The '$D_{Treatment}$' variables are dummy variables for the respective treatments; KeepWinner serves as the reference treatment. The bribe variable and both quality variables are standardized using the respective variable's combined standard deviation over all treatments. We randomly select one worker per referee in each round. For other variable definitions, see the notes to Table 4.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

The fact that both bribes and performance appear to be important may be the result of a positive correlation between differences in bribe size and differences in performance (r=.31; p=.082, OLS, one-sided.). The regression in Table 5 corrects for this correlation. The results show that having a higher bribe than the opponent affected the probability of winning significantly less than in the KeepWinner treatment. The bribe coefficient is similar to the bribe coefficient estimated for the KeepBoth treatment. Similarly, when bribes differed, having a better performance mattered significantly more than in the KeepWinner treatment; the coefficient is similar to the coefficient of the KeepBoth treatment. Overall, the results suggest that moral costs are higher in the Objective treatment, resulting in a larger emphasis on quality.

The CoinFlip Treatment—In 70% of the cases, referees elected to flip the coin rather than determine the winner themselves, and in 100% of the cases, referees decided to flip the coin in at least one round. As a result, neither the higher bribe (50%; p=1, Wilcoxon) nor the better joke (36%; p=.180, Wilcoxon) won significantly more (or less) often than chance.

Table 5 confirms these results. Given that the majority of referees pick the winner at random, quality and bribe differences do not predict the likelihood of winning. The interaction terms for the CoinFlip treatment tend to be both large and negative, though they are not always estimated precisely enough to be significant. Moreover, if we re-estimate the regression separately for this treatment, none of the quality and bribe coefficients are significant in any of the three specifications, as expected. Thus, the finding that most referees were unwilling to pay $1 to select the

winner, even when workers sent them a bribe, further highlights the importance of greed in driving distortion.

1.5 An Experiment in the Market in Shillong, India

The results of our experiment show that the mechanism by which bribes distort the referees' decisions is greed and not reciprocity. Here, we complement the laboratory evidence with evidence from an experiment in a different, more natural setting. Whereas the lab experiments allow us to sharply disentangle between the different mechanisms of bribery in a clean setting, the extra-laboratory experiment (Charness et al., 2013) allows us to investigate whether our results generalize to a population and environment that are more regularly exposed to bribery than UC San Diego students.

1.5.1 Experimental Design

We conducted the experiment at the market in the city of Shillong, in the state of Meghalaya in northeast India. Bribery and corruption are prevalent in India (Transparency International, 2014), and Meghalaya is thought to be among the most corrupt states in India (Transparency International India, 2008).

Participants (N=120) in the experiment were visitors in the market who were approached, at random, by research assistants. All participants were asked to taste two different pineapples, each purchased from a different vendor, and tell us which of the two they thought tasted better. As with the joke task in the lab experiment, we chose

this task because selecting the tastier pineapple is partially based on the decision maker's subjective judgment.

In contrast to the laboratory experiment in San Diego, the bribers' side was exogenously determined by the experimenters. Before the beginning of the experiment, we approached two sellers, A and B, in the market and invited them to participate in the study. They were both selling pineapples and their stands were not close to each other. We explained to the sellers that we would purchase some of their pineapples and ask shoppers in the market to taste them. We told the sellers that shoppers would be asked to taste their pineapple as well as a pineapple from another seller in the market, and indicate which of the two was tastier. We told the sellers that every time a shopper recommended their pineapple, we would purchase an additional pineapple from their stand at a price of 60 rupees (approximately $1 at the time).

Each seller agreed to pay some money to each shopper who chose his pineapple (except those in the control treatment, see below). In particular, Seller A agreed to pay 10 Rupees and Seller B agreed to pay 20 Rupees. Both sellers also agreed that in half of the cases, they would pay these amounts even if the shopper did not choose their pineapple.

Before starting the experiment, we selected two pineapples from seller A (pineapples A1 and A2), and two from seller B (pineapples B1 and B2). For this purpose, four experimenters tasted several pineapples. We then chose four pineapples such that all four experimenters thought pineapple A1 was tastier than pineapple B1,

and pineapple A2 was tastier than pineapple B2. We then cut each selected pineapple into small pieces that we placed in separate bowls.

Determining the combination of quality and bribes in this way ensures that all participants received the same combination of quality and bribes. Because we always matched the low bribe with the tastier pineapple, a trade-off between quality and bribes always existed, increasing the power of our study.

Two research assistants, a male and a female, conducted the experiment. We instructed the research assistants to approach shoppers at the market to ask them to taste the pineapples from the two bowls and indicate which one was tastier. The first 60 participants tasted pineapple A1 and pineapple B1; the other 60 participants tasted pineapple A2 and pineapple B2. The procedure for the first and the second group was the same.

We conducted three treatments, with 40 participants in each (20 for each set of pineapples). The first treatment was a control with no bribes, whereas the other two were analogous to treatments KeepWinner and KeepBoth. Participants were never informed that they were taking part in an experiment.

In the Control treatment, we asked the research assistants to follow the following script. After approaching the participant, we asked them to tell the participant the following (translated to the local language--Khasi): “Thank you for agreeing to help us. We will pay you 10 Rupees for your time. We would like to ask you to tell us which of these two pineapples is tastier. It is important for us because we

will buy an extra pineapple from the seller who sold us the one you will tell us is tastier. Please taste both and tell us which one is tastier."

The research assistants then asked participants to taste both pineapples and indicate which one was tastier. Participants received their payment of 10 Rupees after making their choice. During the experiment, research assistants were instructed to switch the hands in which they were holding the bowls after each participant, and always start the tasting with the bowl on the left hand. In this way, we counterbalanced any order effect.

Treatment KeepWinner was similar to the control treatment, but instead of paying participants 10 Rupees for tasting the pineapples we told them seller A had offered 10 Rupees to those who chose his pineapple, and seller B had offered 20 Rupees to those who chose his pineapple. Participants were also told they could only keep the money offered by the seller of the pineapple that they indicated as tastier. The following additional wording was added to the script before we asked participants to taste the pineapples: "The seller of this pineapple [the RA holding the bowls raised the bowl containing pineapple A1 or A2] offered you 10 Rupees if you will choose his pineapple, and the seller of this pineapple [now the bowl containing pineapple B1 or B2 was raised] offered 20 Rupees to you if you will choose his pineapple. As a result, you will be paid 10 Rupees if you choose this one and 20 Rupees if you choose this one [again, the respective bowls were raised]." Participants then tasted both pineapples, chose one, and were paid according to their choice.

Treatment KeepBoth was similar to treatment KeepWinner, but participants were told that regardless of their choice, they would be paid both the 10 Rupees offered by seller A and the 20 Rupees offered by seller B. Specifically, the protocol was as follows: "The seller of this pineapple [the RA holding the bowls raised the bowl containing pineapple A1 or A2] offered you 10 Rupees and the seller of this pineapple [the bowl containing pineapple B1 or B2 was raised] offered 20 Rupees. As a result, you will be paid 30 Rupees regardless of your choice." Participants then tasted both pineapples, chose one, and received their payment.

1.5.2 Results

The results are presented in Figure 3. In the control treatment, 77.5% of the participants indicated pineapple A was tastier. This fraction is significantly larger than predicted by chance (i.e., 50%, $p<.001$, test of proportions), which suggests that pineapple A was indeed tastier than pineapple B. Thus, in this treatment, most participants agreed with the experimenters about which of the two pineapples was tastier.

In treatment KeepWinner, shoppers chose pineapple A only 35% of the time. The difference between the fraction of participants choosing pineapple A in KeepWinner and in the control treatment is significant (test of proportions, $z=3.83$, $p<.001$).

In treatment KeepBoth, the fraction of participants choosing pineapple A was 67.5%, which is significantly higher than the fraction observed in the KeepWinner

treatment (test of proportion, z=2.91, p=.003), and does not differ from the control treatment (test of proportion, z=1.00, p=.317).

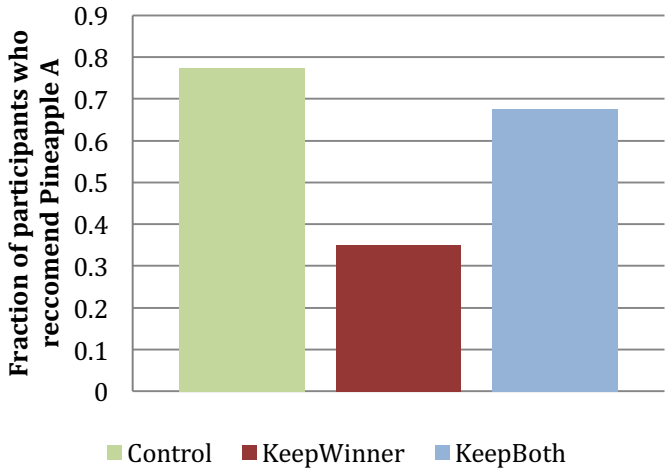

Figure 1.3 Fraction of Participants who Choose the Tastier Pineapple

We further explore these results using OLS regressions in which we estimate treatment effects on the probability of choosing pineapple A. Since observations are not clustered, we use robust standard errors to compute the p-values. The regression (Table 6) confirms that participants in KeepWinner were significantly less likely to choose the tastier pineapple than participants in the control treatment. Participants in KeepBoth are not significantly less likely to choose the tastier pineapple than participants in the control treatment. The difference between the KeepWinner and the KeepBoth coefficients is also significant (F(1,117)=9.22, p=.003). In column (2), we interact the particular pineapple that was tasted by the subjects with treatment

dummies. We find that the treatment effect is similar regardless of the particular pineapple that was tasted.

Table 1.6 OLS Regressions for the India Experiment

Dependent Variable:	Pineapple A Wins (1)	(2)
Constant	.775	.750
	(.000)	(.000)
$D_{KeepWinner}$	-.425***	-.400***
	(.000)	(.008)
$D_{KeepBoth}$	-.100	-.050
	(.322)	(.730)
Pineapple A2/B2		.050
		(.712)
Pineapple A2/B2 X $D_{KeepWinner}$		-.050
		(.808)
Pineapple A2/B2 X $D_{KeepBoth}$		-.100
		(.624)
Observations	120	120

Notes: OLS estimates (p-values). The dependent variable is a dummy that specifies whether the referee selected the best-tasting pineapple (A) as the winner. $D_{KeepWinner}$ and $D_{KeepBoth}$ are dummy variables that are equal to one for KeepWinner and KeepBoth respectively, and zero otherwise. Pineapple A2/B2 is a dummy variable that is equal to one if the Pineapples tasted were Pineapple A2 and B2, and equal to zero if the pineapples tasted were A1 and B1. The control treatment serves as the reference treatment. P-values are calculated using robust standard errors.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

1.6 General Discussion

Bribery is widespread and has an important impact on how decisions are made in politics, business, sports, education, and many other domains, with large economic consequences. Despite some economic arguments that bribes are not necessarily bad for society but are simply used to "grease the wheels" of bureaucracy (e.g., Leff, 1964; Huntington, 1968), even in these cases, when bureaucrats can endogenously choose

the level of corruption, bribes clearly have a negative effect on economic efficiency (Banerjee, 1997).

The purpose of the current paper is to investigate the relative importance of greed and social preferences in motivating bribery. We find that when incentives are contingent on choices, individuals accept and reward bribes: in our experiments, referees systematically rewarded the higher bribe when they could only keep the winner's bribe. However, when bribes were not contingent on delivering a certain outcome, individuals did not distort judgment. This finding supports the greed explanation of bribery. The extra-laboratory experiment we conducted in the market in India confirms the results we observed in San Diego, outside of the lab and with a population that is more accustomed to bribery.

Our experiments show the norm of reciprocity seems to be weaker than the moral costs of distorting judgment, which are weaker than profit maximization. Our ability to rank these different forces comes from the experimental bribery game that we introduce, which is able to capture the moral costs associated with distortion of judgment that is generated when bribes, rather than performance, are rewarded. We find that distortion plays an important role in explaining whether referees reciprocate the higher bribe. When the decision of which worker wins a prize does not involve a distortion of judgment (as in treatment NoTask), we find that reciprocity is an important motivator of behavior. Further, we find that increasing the moral costs of distortion by providing a more objective task (as in treatment Objective) makes referee less likely to go along with the higher bribe and more likely to choose based on

quality. These results thus show that capturing distortion in bribery experiments is important because moral costs of distortion have an important influence on the behavior of participants.

Our investigation of the motives that induce decision makers to accept and reward bribes provides insights into how to better target anti-corruption interventions. One implication is that anti-corruption policies designed to reduce the effectiveness of reciprocity may not be effective. If greed drives behavior, as our results suggest, policy interventions that enforce monetary sanctions may be more effective in preventing corruption.

Policy interventions that focus on increasing the moral costs of distortion may provide an alternative way to reduce bribery. In a related paper (Gneezy, Saccardo, Serra-Garcia, and Van Veldhuizen, 2015), we examine the role of self-deception in distorting judgment, by varying when evaluators are informed about their incentives to recommend one of two options either before or after their initial private judgment. When the information regarding the incentives is provided before (as in treatment KeepWinner here), we find a significant bias in judgment in the direction of the incentive. However, when the information is provided after they privately evaluate the options, the bias in judgment is significantly reduced. In other words, limiting self-deception may increase the moral costs of distortion, which limits the effectiveness of bribery.

Future research could build on our game and findings in at least two other important ways. First, in our experiment, as in the example of Rod Blagojevich, the

workers who lost because of bribery suffered the negative externality of distorted justice. This negative externality did not reduce the overall wealth of the participants. In many real-world cases of bribery, however, the negative externality could be much larger and reduce the overall earnings. Our game could be extended by incorporating a negative externality (e.g., Falk and Szech, 2013) into the design or by making bribery welfare decreasing.

Second, a future design could study how the chance of being audited and penalized for accepting bribes affects the decisions in such games. Our investigation of bribery focused on the case in which the decision of who wins the prize is subjective, in which implementing monitoring and punishment mechanisms is often hard, because of the subjective nature of the choice. Monitoring and punishment could instead be investigated in contexts in which the decision of who wins the prize is objective. Future research could focus on the case in which decision makers have to make objective decisions, and use this design to study the interplay between the probability of being caught and the size of the penalty, and how this interplay affects the decision to offer or accept a bribe.

Chapter 1, in full, has been submitted for publication of the material as it may appear in the Review of Economic Studies, 2015. Gneezy, Uri, Silvia Saccardo, Roel van Veldhuizen, “Bribery: Greed versus Reciprocity.” The dissertation author was the co-primary investigator and author of this paper.

Appendix A. Additional Analyses

A1. The Determinants of Bribing Behavior

Table 1.A1 Determinants of Bribe Amount and Joke Quality

Dependent Variable:	Bribe Amount (1)	Joke Quality (2)	Obj. Score (3)
Joke Quality	0.032		
	(0.764)		
Objective Score	0.036		
	(0.102)		
Confidence	-0.020	.648**	31.88**
	(0.930)	(.022)	(.028)
Female	-0.445*	-.084	9.73**
	(0.064)	(.566)	(.034)
Nonnative speaker	0.843***	-.546**	2.99
	(0.004)	(.020)	(.818)
Not of Asian Ethnicity	0.318	.347**	-6.18
	(0.158)	(.046)	(.270)
Age	0.111*	-.047	-1.75
	(0.062)	(.276)	(.138)
Economics Major	-0.206	-.087	7.08
	(0.502)	(.632)	(.246)
Psychology Major	0.221	.207	-8.33
	(0.566)	(.468)	(.218)
Engineering/Science Major	-0.052	-.138	6.35
	(0.904)	(.496)	(.288)
Other Social Science Major	-0.769*	-.108	6.91
	(0.084)	(.764)	(.312)
Humanities Major	-0.195	.121	-8.07
	(0.664)	(.608)	(.192)
Treatment Dummies	Yes	Yes	Yes
Treatment	All Treatments	All Joke Treatments	Objective
Observations	484	280	84
Clusters	242	140	42

Notes: OLS estimates (p-values). Bribe amount is the bribe amount sent by the worker. Joke quality is the average rating of the joke by the independent judges. Objective score is the worker's score on the objective task. Confidence is the worker's confidence in having a better joke than the other worker. The remaining variables are dummies for females, nonnative speakers, and non-Asians, a continuous variable for age, and dummies for different majors, respectively (the omitted groups are biology/chemistry majors and undeclared majors). P-values (in brackets) are calculated using wild bootstraps.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

Table 1.A1 presents the results of the analysis on the determinants of bribing behavior. The results of column (1) are discussed in the results section. Columns (2) and (3) show that a correlation exists between joke quality (or performance in the objective task) and workers' belief that they will win the prize. The regressions further show that native speakers and subjects that did not have an Asian ethnicity wrote better-quality jokes, and that women performed better in the objective task.

A2. Differences in Bribes

A feature of our design is that although treatment is randomized and workers are randomly paired within sessions, bribes are not determined at random. To analyze referees' behavior and make treatment comparisons, the referees must face similar combinations of bribes across treatments. That is, the distribution of the difference between the bribes the referees receive must be similar across both treatments. In the result sections, we showed that, despite the average bribe being significantly higher in KeepWinner, the difference between bribes was indeed similar in treatment KeepWinner and treatment KeepBoth.

Nevertheless, the fact that bribes are not randomized in our experiment leads to three additional treatment differences between KeepWinner and KeepBoth, which may in turn affect referee behavior. First, referees in treatment KeepWinner are less likely to face two identical bribes. We already addressed this concern in the results section, by estimating a separate coefficient for quality when bribes are equal and when bribes differ in our regressions. Since the latter coefficient is only identified by the observations where bribes are different, any treatment differences we find in the size of this coefficient cannot be explained by a smaller frequency of equal bribes in treatment KeepWinner.

Second, referees in treatment KeepWinner are more likely to face two bribes rather than one. Third, referees in treatment KeepWinner are more likely to have to choose between two large bribes (e.g., \$5 and \$3) than referees in treatment KeepBoth. We will discuss the latter two differences below and provide evidence suggesting that the respective differences cannot explain our results.

One Bribe versus Two Bribes. Referees were more likely to receive two bribes (e.g., \$5 versus \$2) in treatment KeepWinner than in treatment KeepBoth. Indeed, of all the times they received bribes of different sizes, referees in KeepBoth received only one bribe (e.g., 3\$ vs. \$0) 82% of the time, versus only 41% for the referees in the KeepWinner treatment. To correct for this difference, we re-estimate regressions of Table 4 by including separate coefficients for the two-bribes and one-bribe case for both quality and bribes. This approach allows us to compare the importance of bribes and quality between cases in which referees received one or two bribes, respectively.

Table 1.A2 OLS Regressions for Referees One or Two Bribes Positive

Dependent Variable:	Winner (1=Yes)	
	(1)	(2)
Bribe difference	.330**	.131*
[One bribe positive]	(.022)	(.082)
Quality difference (bribes differ)	.066	.236**
[One bribe positive]	(.416)	(.036)
Bribe difference	.280***	.128
[Both bribes positive]	(.002)	(.140)
Quality difference (bribes differ)	-.056	.646
[Both bribes positive]	(.630)	(.162)
Quality difference (bribes identical)	.342**	.273**
	(.022)	(.022)
Treatment	KeepWinner	KeepBoth
Selected Workers	Random	Random
Observations	40	40
Clusters	20	20

Notes: OLS estimates (p-values). The dependent variable is a dummy that specifies whether the selected worker was selected as the winner. For other variable definitions, see the notes to Table 4. Separate coefficients are included for cases in which one bribe is positive and in which both bribes are positive. P-values (in brackets) are calculated using wild bootstraps. We randomly select one worker per referee in each round.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

Table 1.A2 presents the results. For the KeepWinner treatment, the coefficients for bribes are significant and nearly identical in both cases. Similarly, the coefficient for quality is small and not significant in either case as well. This finding suggests that within treatment, KeepWinner referees do not react differently to bribery in the one-bribe versus two-bribes case.

Treatment KeepBoth does not have enough observations with two different bribes for us to estimate the coefficients for the two-bribes case with any precision. For the one-bribe case, the coefficient for bribes is slightly larger than the coefficient in Table 4 and is significant at the 10% level. At the same time, the (standardized) coefficient for quality is still nearly twice as large as the (standardized) coefficient for bribes, suggesting that quality played a larger role in the referee's decision making. Further, the comparison with the KeepWinner treatment shows the coefficient for bribe difference in KeepBoth is 60% smaller than the coefficient for KeepWinner, whereas the coefficient for quality is substantially larger. These results are similar to the results of Table 4. Thus, the treatment difference between KeepWinner and KeepBoth does not appear to have been the result of referees being more likely to receive only one bribe in the KeepBoth treatment.

Absolute Bribe Size—A related difference between KeepWinner and KeepBoth that arises from the fact that bribes in our experiment are not determined at random is that referees in KeepWinner are more likely to face two relatively large bribes. Referees might respond differently to receiving two large bribes than to receiving two smaller

ones, even when the difference in bribes is the same. Explaining our treatment effect would require that referees care *more* about a difference of $5 vs. $4 than $2 vs. $1, as could be the case if referees are more willing to reward a bribe when the absolute value of the bribe is high (note, however, that the converse argument could also be made).

To check whether this was indeed the case, we redid the regressions of Table 4 while controlling for the sum of the two bribes. We also interact the sum variable with the bribe-difference variable. The estimate for the interaction effect tells us whether bribes had a larger (or smaller) role when the sum of the two bribes was larger.

Table 1.A3 shows the results. The interaction effect is not significant for either treatment. If anything, the coefficient estimate for the KeepBoth treatment (p=.234) suggests referees care *less* about bribes as the sum of the two bribes increases, which would increase the size of the treatment difference. Thus, overall, we find no evidence that differences in absolute bribe size can explain our treatment effect.

Table 1.A3 OLS Regressions for Referees with Absolute Bribe Size

Dependent Variable:	Winner (1=Yes)	
	(1)	(2)
Bribe difference	.327***	.110
	(.000)	(.278)
Quality difference (bribes differ)	-.012	.291**
	(.904)	(.012)
Quality difference (bribes identical)	.365**	.274**
	(.016)	(.022)
Sum of the two bribes	.066	.083
	(.376)	(.362)
Bribe Difference X Sum of the two	.031	-.130
	(.802)	(.234)
Treatment	KeepWinner	KeepBoth
Selected Workers	Random	Random
Observations	40	40
Clusters	20	20

Notes: OLS estimates (p-values). The dependent variable is a dummy that specifies whether the selected worker was selected as the winner. The sum of the two bribes is the sum of the bribes of both workers in the pair. For other variable definitions, see the notes to Table 4. P-values (in brackets) are calculated using wild bootstraps. We randomly select one worker per referee in each round.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

A3. The Quality Threshold

All non-parametric tests in the main text that relate to quality use only those pairs of jokes in which at least 65.1% of independent raters agreed on the winner. With this threshold, the fraction of independent raters who selected a given joke over the other is significantly different from chance (i.e., 50%) at the 10% level by a test of proportions.

In this section, we redo the main analysis using a more stringent threshold of 69.4%, which entails that the two jokes are significantly different at the 5% level (test of proportions, Z=1.65, n=18, p=.050). The table below summarizes the non-parametric results. The results are very similar for both thresholds. The biggest difference is that the comparison between the KeepWinner and KeepBoth treatment is significant at the 1% level for the 65.1% threshold and at the 5% level for the more stringent threshold, which is largely due to a loss of observations. Similarly, the difference between KeepWinner and KeepBothReject is no longer significant, because of a lack of observations, although the result is still very similar and not significantly different from KeepBoth (p=.611, Mann-Whitney).

Table 1.A4 Non-Parametric Tests for Alternative Threshold

		Better Quality Wins (%)		Difference vs. KeepWinner	
Treatment	Threshold:	65.1%	69.4%	65.1%	69.4%
KeepWinner		56.7%	60.0%		
		(.564)	(.405)		
KeepBoth		90.5%	88.9%	33.80%	28.90%
		(.001)	(.002)	(.004)	(.030)
KeepWinnerReject		57.1%	58.3%	0.40%	-1.70%
		(.655)	(.655)	(.978)	(.852)
KeepBothReject		90.5%	81.8%	33.80%	21.80%
		(.005)	(.014)	(.026)	(.119)
HighWage		44.0%	40.9%	-12.70%	-19.10%
		(.593)	(.564)	(.439)	(.331)
Objective		83.3%	86.4%	26.60%	26.40%
		(.001)	(.001)	(.023)	(.023)
CoinFlip		36.4%	40.0%	-20.30%	-20.00%
		(.180)	(.414)	(.167)	(.263)

Notes: Percentages (P-values). The first two columns display the percentage of times the best joke won for each treatment, for a given threshold. The last two columns display the difference between the percentage of winners between the respective treatment and treatment KeepWinner. In both cases, two different thresholds are used. For the 65.1% threshold, the two jokes in a pair differ significantly at the 10% level by a test of proportions; for the 69.4% threshold, the two jokes differ significantly from each other at the 5% level. P-values are computed using Wilcoxon tests and a Mann-Whitney tests for columns 1 and 2 and columns 3 and 4 respectively.

A4. Alternative Regression Specifications and Demographic Controls

For our main regression (Table 4 and Table 5), we selected one worker per pair at random for each given round and then investigated how differences in quality and bribes between the selected worker and the opponent affect winning. Table A5 provides the results of two alternative specifications, which we first estimate separately for the two main treatments.

In regression (1) and (2), we focus on the workers who submitted a larger bribe than their opponent in a given round. We then investigate whether for those workers, increasing the quality of the joke relative to the opponent further increased their likelihood of winning. In treatment KeepWinner, the higher bribe already wins 86% of

the time. Hence, increasing the quality of the worker's joke (relative to the opponent) does not further increase the likelihood of winning. By contrast, in treatment KeepBoth (column 2), the higher bribe only wins 64% of the time, and increasing the relative quality of the worker's joke is highly beneficial. These results are in line with the results in the main text: quality matters in treatment KeepBoth, but not in treatment KeepWinner.

Table 1.A5 Alternative OLS Regressions for KeepWinner and KeepBoth

Dependent Variable:	Winner (1=Yes)			
	(1)	(2)	(3)	(4)
Bribe Difference			.270**	.014
			(.016)	(.536)
Quality Difference (bribes differ)	.060	.274**		
	(.306)	(.012)		
Treatment	KeepWinner	KeepBoth	KeepWinner	KeepBoth
Selected Workers	Higher Bribe	Higher Bribe	Higher Quality	Higher Quality
Observations	29	22	30	21
Clusters	16	12	18	17

Notes: OLS estimates (p-values). The dependent variable is a dummy that specifies whether the selected worker was selected as the winner. For other variable definitions, see the notes to Table 4. P-values are calculated using wild bootstraps. For specification (1) and (2), we select only workers with a higher bribe than their opponent in the given round. For specification (3) and (4), we select only workers with a better-quality joke than their opponent in the given round. We consider a joke to be of better quality when the agreement of the independent raters is at least 65.1%.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

In regression (3) and (4), we instead focus on the workers who wrote the better joke in a given round, and investigate whether larger differences in bribes between those worker and their opponents make the workers even more likely to win the prize. The results show that for workers who already have the best joke in the pair, bribes have a strong positive effect on the likelihood of winning in treatment KeepWinner, but not in treatment KeepBoth. Thus, similar to the regressions presented in the main text, bribes matter in treatment KeepWinner but not in treatment KeepBoth.

In Table 1.A6 we estimate the regressions reported in Table A5 jointly for all treatments. We report four regressions. In columns (1) and (2), we use KeepWinner as the reference treatment and interact the bribe and quality-difference variables with treatment dummies for all other treatments. This approach allows us to verify whether bribes and quality played a larger or smaller role than in treatment KeepWinner. Columns (3) and (4) are similar to columns (1) and (2) but also include a series of

demographic control variables. In particular, we add controls for all the demographic variables reported in Table 2 and for the round.[1]

The results largely replicate the main results from Table 5. Relative to KeepWinner, bribes play a less important role in KeepBoth and KeepBothReject, whereas quality plays a larger role in these treatments. The only difference lies in treatment Objective, where the effect of quality is still larger than in treatment KeepWinner, but the effect of bribes is no longer significantly smaller, though the coefficient estimate is still fairly large. In addition, including demographic controls does not substantially affect the coefficient estimates. In particular, p-values are very similar across both specifications.

[1] We could not control for demographics in the regressions presented in the main text. When workers are randomly selected, selected workers have a 50% chance of winning, irrespective of the referee's demographics, and hence all demographic variable coefficients are zero in expectation.

Table 1.A6 Alternative OLS Regressions for All Treatments

Dependent Variable: Winner (1=Yes)	(1)	(2)	(3)	(4)
Bribe Difference		.255**		.245**
		(.030)		(.032)
Bribe Difference X $D_{KeepBoth}$		-.240**		-.194**
		(.012)		(.028)
Bribe Difference X $D_{KeepWinnerReject}$		.210**		.227**
		(.014)		(.034)
Bribe Difference X $D_{KeepBothReject}$		-.305***		-.319**
		(.006)		(.012)
Bribe Difference X $D_{HighWage}$		.012		.035
		(.864)		(.714)
Bribe Difference X $D_{Objective}$		-.162		-.179
		(.122)		(.110)
Bribe Difference X $D_{CoinFlip}$		.066		.108
		(.770)		(.726)
Quality Difference (Different Bribes)	.057		.064	
	(.266)		(.232)	
Quality Difference (Different Bribes) X $D_{KeepBoth}$	.173**		.194***	
	(.030)		(.006)	
Quality Difference (Different Bribes) X $D_{KeepWinnerReject}$	-.057		-.040	
	(.254)		(.418)	
Quality Difference (Different Bribes) X $D_{KeepBothReject}$	.364**		.365***	
	(.012)		(.008)	
Quality Difference (Different Bribes) X $D_{HighWage}$	-.019		-.048	
	(.816)		(.648)	
Quality Difference (Different Bribes) X $D_{Objective}$	.217*		.206*	
	(.050)		(.050)	
Quality Difference (Different Bribes) X $D_{CoinFlip}$	-.321		-.345	
	(.146)		(.132)	
Selected Workers	Higher Bribe	Higher Quality	Higher Bribe	Higher Quality
Treatment Dummies	Yes	Yes	Yes	Yes
Demographic Controls	No	No	Yes	Yes
Observations	163	145	163	145
Clusters	97	98	97	98

Notes: OLS estimates (p-values). P-values are computed using wild bootstraps. The '$D_{Treatment}$' variables are dummy variables for the respective treatments; KeepWinner serves as the reference treatment. Control variables include the round and all demographic variables of Table 2, where we use bio/chemistry majors and undeclared majors as the reference group. Bribe and quality variables are standardized using the combined standard deviation over all included treatments. For other variable definitions and explanation of how we selected the workers in each regression, see the notes to Table 4.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

Identical Quality—In the regressions reported in the main text, we estimate separate coefficients for quality when bribes are equal and when bribes differ. Here, we present a similar analysis in which we also estimate separate coefficients for bribes when quality is equal (using the 65.1% threshold) and when quality differs. The latter coefficient allows us to focus on cases in which quality differs, which is where a larger treatment difference should be expected. We do not use this approach in the main text because the threshold between 'similar' jokes and different jokes is not as clear-cut as the threshold between identical and non-identical bribes.

Table 1.A7 reports the results. For treatment KeepWinner, bribes have a larger effect when quality is equal and therefore cannot influence judgment. Similarly, in treatment KeepBoth, the coefficient for bribes is larger when quality is equal, though it is still not significant. The interaction terms in column (3) shows that the treatment difference in the importance of bribes overall seems to be largely driven by cases in which quality differs, as expected. These results indeed suggest that bribes play a larger role when quality is equal in both treatments.

Table 1.A7 OLS Regressions for referees with separate bribe coefficients

Dependent Variable:	Winner (1=Yes)		
	(1)	(2)	(3)
Bribe Difference (quality identical)	.453**	.172	.348**
	(.020)	(.184)	(.020)
Bribe Difference (quality differs)	.267**	.033	.254**
	(.020)	(.458)	(.020)
Quality Difference (bribes differ)	.009	.270***	.010
	(.838)	(.008)	(.838)
Quality Difference (bribes identical)	.339**	.276**	.257**
	(.020)	(.026)	(.020)
$D_{KeepBoth}$			.024
			(.882)
Bribe Difference (quality identical) X			-.145
			(.314)
Bribe Difference (quality differs) X			-.216**
			(.012)
Quality Difference (bribes differ) X			.234***
			(.004)
Quality Difference (bribes identical) X			.151
			(.168)
Treatment	KeepWinner	KeepBoth	KeepWinner KeepBoth
Selected Workers	Random	Random	Random
Observations	40	40	80
Clusters	20	20	40

Notes: OLS estimates (p-values). The dependent variable is a dummy that specifies whether the selected worker was selected as the winner. For other variable definitions, see the notes to Table 4. P-values (in brackets) are calculated using wild bootstraps. We consider jokes to be of identical quality when fewer than 65.1% of independent raters agree on the winner. We randomly select one worker per referee in each round.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

A5. Robustness Check: Alternatives to the Wild Bootstrap

In the main regression analysis, we computed p-values using wild bootstraps, as suggested by Cameron et al. (2008). In this section, as a robustness check, we provide the results of several alternative techniques, which we use to compute the p-values of the KeepWinner and KeepBoth treatment regressions reported in Table 4.

We use four different techniques. The first is the wild bootstrap procedure, which we use in the main text. Cameron et al. (2008) show that this approach (which they refer to as the "wild bootstrap-t" procedure) leads to more consistently accurate (and more conservative) rejection rates than alternative approaches, and therefore recommend its use in case of a small number of clusters. For more details on how the technique works, see Cameron et al. (2008).

For column (2), we recalculate our p-values using clustered standard errors, a standard approach in experimental economics. For column (3), we use a non-

parametric bootstrap, which is, to our knowledge, the most widely used bootstrap method in experimental economics. Finally, we also redo the main regressions using probit (with clustered standard errors).

The top half of Table A8 reports the results for treatment KeepWinner; the lower half presents the results for treatment KeepBoth. Each column presents the results of a different estimation technique. Column (1) reports the results of Table 4. The other columns show that alternative estimation techniques result in very similar p-values. In particular, both bribes are significant at the 1% level across specifications, and quality (for different bribes) is never significant in any specification. The only difference is that quality (for equal bribes) is significant at the 5% level in column 1, but significant at the 1% level in all other specifications.[2]

For KeepBoth, the results are also very similar across specifications. Bribes are not significant, and quality (for different bribes) is significant at the 1% level in all specifications. As with KeepWinner, the main difference is that the quality variable for equal bribes is significant at the 5% level in the wild bootstrap, but at the 1% level in the other specifications (except probit).

[2] The marginal-effect estimates for probit are typically larger than the coefficient estimates of OLS, which is the result of the marginal effect being estimated at the sample average, where the probability of winning is approximately .5.

Table 1.A8 Separate OLS Regressions with Alternative P-value Estimates for Main Treatments

Dependent Variable: Winner	(1)	(2)	(3)	(4)
(KeepWinner)				
Bribe Difference	.308***	.308***	.308***	.488***
	(.000)	(.000)	(.000)	(.000)
Quality Difference (bribes	.014	.014	.014	.054
	(.762)	(.794)	(.830)	(.619)
Quality Difference (bribes	.336**	.336***	.336***	.526***
	(.020)	(.000)	(.002)	(.002)
(KeepBoth)				
Bribe Difference	.086	.086	.086	.154
	(.140)	(.161)	(.222)	(.205)
Quality Difference (bribes	.262**	.262***	.262***	.613***
	(.010)	(.001)	(.004)	(.005)
Quality Difference (bribes	.275**	.275***	.275***	.420**
	(.022)	(.000)	(.001)	(.012)
Selected Workers	Random	Random	Random	Random
Technique	OLS	OLS	OLS	Probit (Marg.
P-Values	Wild BS	Clustered SE	Non-Par	Clustered
Observations	40	40	40	40
Clusters	20	20	20	20

Notes: Regression estimates (p-values). Each column presents the results of two regressions, which for the top and bottom panel are analogous to regressions 1 and 2 in Table 4, respectively. For variable definitions and other details, see Table 4 and its notes. Columns (1)-(3) use OLS, column (4) uses probit. P-values are computed either by wild bootstrap (column (1)), clustered standard errors (column (2) and (4)) or a non-parametric bootstrap (column (3)).

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

Table 1.A9 presents the regression with interaction effects (column (3) in Table 4). We do not estimate the probit model here, because interaction terms are difficult to interpret with marginal effects. Similar to Table A8, the p-values are quite robust with respect to the estimation technique we use.

Overall, the wild bootstrap technique and alternatives yield similar p-values in both Table 1.A8 and 1.A9. If anything, the wild bootstrap tends to be the most conservative technique, which is in line with Cameron et al. (2008).

Table 1.A9 OLS Regressions with Alternative P-value Estimates For Main Treatments

Dependent Variable: Winner	(1)	(2)	(3)
Bribe Difference	.274***	.274***	.274***
	(.000)	(.000)	(.000)
QualityDif*bribesDif	.015	.015	.015
	(.762)	(.791)	(.839)
QualityDif*bribes identical)	.255**	.255***	.255***
	(.020)	(.000)	(.005)
$D_{KeepBoth}$	.008	.008	.008
	(.980)	(.942)	(.947)
Bribe Difference X $D_{KeepBoth}$	-.173**	-.173**	-.173*
	(.032)	(.035)	(.058)
Quality Difference X $D_{KeepBoth}$	.222**	.222***	.222**
(bribes differ)	(.014)	(.009)	(.049)
Quality Difference X $D_{KeepBoth}$	.150	.150	.150
(bribes identical)	(.156)	(.142)	(.320)
Treatment	KeepWinn KeepBoth	KeepWinner KeepBoth	KeepWinn KeepBoth
Selected Workers	Random	Random	Random
Technique	OLS	OLS	OLS
P-Values	Wild BS	Clustered SE	Non-Par
Observations	40	40	40
Clusters	20	20	20

Notes: OLS estimates (p-values). Each column presents the results of a single regression, which is analogous to regression 3 in Table 4. For variable definitions and other details, see Table 4 and its notes. P-values are computed either by wild bootstrap (column (1)), clustered standard errors (column (2)), or a non-parametric bootstrap (column (3)).

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

A6. Robustness Check: Alternative Random Samples of Referees

In our main regression analyses, we randomly selected one worker per referee in a given round. We use the same random sample for all regressions. Here, we present evidence that shows the results we present do not depend on the particular sample of workers we randomly selected. For this purpose, we re-estimate of the regressions of Table 4 1,000 times for 1,000 random samples of workers. In every random sample, each referee and round combination is represented exactly once; the only thing that differs across samples is whether worker A or worker B is included in the analysis (for a given pair and round).

Table A10 below presents the average resulting coefficient estimates as well as the standard deviation of the coefficient estimates (in square brackets). Overall, coefficient estimates do not differ much across the different samples. The only exception is the dummy for treatment KeepBoth in column (3); this exception is due to the random sample sometimes containing more winners for treatment KeepBoth than for KeepWinner and vice versa. Overall, the results are very robust with respect to the particular random sample selected for these regressions.

Table 1.A10 OLS Regressions with Alternative Random Samples

Probability (winning)	(1)	(2)	(3)
Bribe Difference	0.304	0.092	0.275
	[0.011]	[0.011]	[0.010]
Quality Difference (bribes differ)	0.014	0.253	0.014
	[0.011]	[0.019]	[0.012]
Quality Difference (bribes	0.341	0.274	0.251
	[0.028]	[0.020]	[0.015]
$D_{KeepBoth}$			-0.004
			[0.095]
Bribe Difference X $D_{KeepBoth}$			-0.170
			[0.017]
Quality Difference X $D_{KeepBoth}$ (bribes differ)			0.222
			[0.021]
Quality Difference X $D_{KeepBoth}$ (bribes identical)			0.144
			[0.028]
Treatment	KeepWinner	KeepBoth	KeepWinne KeepBoth
Observations	40	40	80
Clusters	20	20	40

Notes: OLS estimates [standard deviations]. For each column, we re-estimate the regression of, respectively, columns (1), (2), and (3) of Table 4 1,000 times. For each replication, we randomly select a new sample of one worker per referee per round. The presented estimates are the average coefficient estimates (over all replications). The standard deviations are the standard deviations of the coefficient estimates. For variable definitions, see Table 4 and its notes.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

A7. Comparison between rounds

The analysis of referees' behavior aggregates observations for rounds 1 and 2. However, it seems possible that some referees changed their behavior between rounds. To allow for this, Table A11 re-estimates regressions (1) and (2) from Table 4 separately for each round. This allows us to investigate whether the impact of quality and bribes was different in round 1 and round 2. This analysis uses only one observation per cluster, and hence we no longer compute standard errors using wild bootstraps, but use standard robust standard errors to compute p-values instead.

Table 1.A11 OLS Regressions for KeepWinner and KeepBoth Separately for each Round

Probability	(1)	(2)	(3)	(4)
Bribe Difference	.292***	.295***	.184**	-.049
	(.000)	(.001)	(.025)	(.556)
Quality	-.112	.113	.270***	.289**
(bribes differ)	(.118)	(.246)	(.000)	(.014)
Quality	.342***	.372***	.280**	.293***
(bribes identical)	(.003)	(.002)	(.018)	(.001)
Treatment	KeepWinner	KeepWinn	KeepBoth	KeepBoth
Selected Workers	Random	Random	Random	Random
Standard Errors	Robust	Robust	Robust	Robust
Round	1	2	1	2
Observations	20	20	20	20
Clusters	20	20	20	20

Notes: OLS estimates (p-values). The dependent variable is a dummy that specifies whether the selected worker was selected as the winner. For other variable definitions, see Table 4 and its notes. P-values (in brackets) are calculated using robust standard errors. For all specifications, we randomly select one worker per referee in each round.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

Columns (1) and (2) show that in the KeepWinner treatment, bribes matter equally in both rounds. Further, we find that quality has a similar impact in both rounds when bribes are identical, and is not significant in either round when bribes differ. For treatment KeepBoth, columns (3) and (4) show the effect of quality is more important than the effect of bribes in both rounds. Both quality coefficients are very similar across rounds and are significant in both cases. The only difference from the analysis presented in the main text is that the coefficient for bribes is zero in round 2 and positive and significant in round 1. This finding suggests referees may have been

more receptive toward bribes in round 1 than in round 2, though the effect of quality stays the same through both rounds.[3]

Another potential explanation for why behavior might differ across rounds is that referees may have alternated between workers across rounds. This explanation may in particular have been the case when bribes and/or joke quality were equal. In the remainder of this section, we check whether alternating played an important role in either of the two main treatments.

First, we show that alternating cannot by itself explain our results. Indeed, if all referees were alternating, neither quality nor bribes would be significant in any treatment. To illustrate this, we ran simulations in which we used workers' actual behavior from the experiment but replaced referees' choices with a random winner in round 1 and then had them choose the other worker in round 2.

The results of this simulation are displayed in Table A12 and show that, as expected, the average estimated coefficient over all simulations is essentially zero for all variables and both treatments. Further, the quality and bribe variables are only significant in approximately 5% of all simulations in both treatments, again as expected by chance. Third, in only .1% of simulation samples (i.e., 2 out of 2,000 across both treatments) were the same coefficients significant (at the 5% level) and had the same sign as in the actual regression estimates. There were no simulations in which the same coefficients were significant as the actual estimates for both treatments at the same time. Thus, the simulations strongly suggest our results cannot have been generated by alternating alone.

---

[3] One potential reason for this result is that, by chance, a greater number of pairs of jokes in round 1 were very similar in quality. If referees found that deciding between two very similar jokes was difficult, they may have selected the better bribe instead. In line with this reasoning, we show in appendix A4 that referees appear to have been more likely to care about bribes when jokes were of similar quality.

Table 1.A12 Simulations for Alternating

	Actual Estimate	Simulation Estimate	Sim. Est. Significant
KeepWinner			
Bribe Difference	.308***	-.001	5.7%
Quality Difference (bribes differ)	.014	.003	5.0%
Quality Difference (bribes identical)	.336**	-.001	5.6%
KeepBoth			
Bribe Difference	.086	.001	4.9%
Quality Difference (bribes differ)	.262**	.000	5.1%
Quality Difference (bribes identical)	.275**	.001	5.2%

Notes: The first column (Actual Estimates) gives the actual coefficient estimates from column (1) (upper panel) and column (2) (lower panel) of Table 4. The second column (Simulation Estimate) presents the corresponding average coefficient estimates from the simulations with alternating. Column 3 (Sim. Est. Significant) presents the percentage of times in which the respective estimated coefficient was significant at the 5% level (in either direction) in a simulation. Here, we used clustered standard errors to calculate significance, because using bootstraps would have been too computationally intensive.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level (Column 1 only)

Though referees alternating across rounds cannot fully explain our results, alternating behavior might have still played a role in some cases. For example, referees might have alternated in cases in which bribes and/or quality were very similar. To check whether alternating played a role, we control for it directly by adding a dummy for round 1 winners to our regression and interacting it with a round 2 dummy. If alternating plays a role, round 1 winners should be less likely to win in round 2, and hence the coefficient for this interaction effect should be negative and significant.

Table A13 below presents the results in columns (2) and (4). Columns (1) and (3) report the relevant regressions from Table 4 (columns (1) and (2) respectively) for ease of comparison. In KeepWinner, the coefficient for round 1 winners is *positive* and significant at the 10% level. A positive coefficient is inconsistent with alternating, but note the effect is small and could be spurious, for example, if the round 1 winner had the better bribe in both round 1 and round 2. Importantly, the comparison between column (1) and (2) shows that the coefficients for bribes and quality remain mostly unaffected by allowing for alternating.

In KeepBoth (column (4)), however, alternating does seem to have played a role: round 1 winners are 53.1 percentage points less likely to win in round 2 than round 1 losers (who have a 75% chance of winning in round 2). The coefficient for bribes also increases slightly and is now significant, though it is still substantially smaller than the quality coefficients and the bribe coefficient for KeepWinner. Allowing for alternating does not affect the coefficients for quality.

Why does alternating play a role in KeepBoth but not in KeepWinner? One possible explanation is that in the absence of profit-maximization motives, referees typically select the better joke. However, in some cases, the two jokes were of very similar quality, and referees might alternate in such cases. Thus, alternating may have emerged when neither greed nor joke quality could determine referee behavior. At the same time, it is important to emphasize that controlling for alternating does not significantly change the coefficient estimates for quality, which still plays a more important role in treatment KeepBoth. Similarly, controlling for alternating does not affect the coefficient estimates for KeepWinner, where bribes are more important than quality.

Table 1.A13 OLS Regressions Allowing for Alternating

Probability (winning)	(1)	(2)	(3)	(4)
Bribe Difference	.308***	.272***	.086	.143**
	(.000)	(.000)	(.140)	(.020)
Quality	.014	.006	262**	.264***
(bribes differ)	(.762)	(.894)	(.010)	(.008)
Quality	.336**	.326**	.275**	.342**
(bribes identical)	(.020)	(.014)	(.022)	(.014)
Round 2		-.029		.252*
		(.852)		(.072)
Round1Winner		.257*		-.531**
*Round 2		(.094)		(.010)
Treatment	KeepWinner	KeepWinn	KeepBoth	KeepBoth
Selected Workers	Random	Random	Random	Random
Observations	40	40	40	40
Clusters	20	20	20	20

Notes: OLS estimates (p-values). The dependent variable is a dummy that specifies whether the selected worker was selected as the winner. Round 2 is a dummy for round 2 observations, and round 1 winner is a dummy for workers who won in round 1. For other variable definitions, see Table 4 and its notes. P-values (in brackets) are calculated using wild bootstraps. We randomly select one worker per referee in each round.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

A8. Equilibrium for KeepBoth

This section gives an overview of equilibrium bribing behavior for workers in treatment KeepBoth, under the assumption that the referee allocates the prize based on the bribes. Since referees get to keep the bribes of both workers in treatment KeepBoth, however, they are financially indifferent between the two workers. Therefore, we allow for the possibility that the referee may instead use an allocation rule that rewards the *lower* bribe with some probability. Specifically, the referee will reward the higher bribe with probability δ, where δ can range from 0 (referees always select the lower bribe) to 1 (referees always select the higher bribe), and allocates the prize randomly (with equal probability) if both bribes are equal.

Given this allocation rule, the expected value of bribing a given amount *b* equals:

$$E\Pi(b_i = b) = P(b > b_j)\delta p + P(b = b_j)\frac{1}{2}p + P(b < b_j)(1-\delta)p - b$$

Here, b_j is the bribe of the competing worker, and $p = 10$ is the prize obtained by the winning worker. In equilibrium, each worker i chooses a strategy $\sigma_i = \{\sigma_b\}^i = (\sigma_0, \dots, \sigma_5)$ that specifies the probability that worker i plays any given bribe, such that the above expression is maximized given the strategy of the competing worker, σ_j. Since the best response functions depend on the referee's allocation rule, in equilibrium worker behavior depends on δ.

We focus on pure strategy equilibria of the game, if they exist, and otherwise specify the symmetric mixed equilibrium. Deriving the equilibrium for a given δ can be done in three steps. The first step is to eliminate strictly dominated bribes. If more than one bribe amount remains, the second step is to then check for pure strategy equilibria. If necessary, the final step is to compute the symmetric mixed equilibrium (ME). For the ME, no profitable deviations should be possible. This means that all bribe amounts that are assigned strictly positive probabilities in equilibrium need to have equal expected payoffs, and other bribe amounts need to have a lower (or equal) expected payoff. Hence, deriving the ME entails solving the system of equations $E\Pi(b = 0) = E\Pi(b = 1) =..$ subject to the constraint that all σ_b are non-negative and sum to one (for both workers).

For $\delta \leq 0.5$, referees on average let the worse bribe win. As a result, the unique equilibrium is for both workers to bribe zero (i.e., $\sigma_0 = 1$ for both workers), since any larger bribe is both costly and results in a weakly smaller likelihood of winning.

Figure A1 below plots the equilibria for increasing values of δ, on the interval $\delta \in [0.5,1]$, for the parameters used in the experiment. For values close to $\delta = 0.5$, choosing a bribe of 1 or more would mean a sure loss of the bribe (=1 or more) versus an expected gain of $((\delta - 0.5) * 10 < 1)$. Hence for $\delta \in [0.5,0.6)$, both workers bribing zero is also the unique equilibrium.

For $\delta = 0.6$, there is an additional pure strategy equilibrium where both workers bribe 1, as well as an equilibrium where one worker bribes 1 and the other bribes 0.

For $\delta \in (0.6,1)$ there are only mixed equilibria. For $\delta \in (0.6,0.7]$,the equilibrium is for both workers to randomize between 0, 1 and 2, where 0 and 2 are chosen with equal probability. Intuitively, there can be no pure strategy equilibria, since the best responses to 0, 1 and 2 are 1, 2 and 0 respectively. As δ increases over the interval, 2 is becoming an ever more attractive bribe, since the advantage relative to a bribe of 1 is increasing, and the disadvantage with respect to a bribe of 0 is decreasing. Therefore, in equilibrium, the frequency of bribes of 0 and 2 must increase.

For $\delta \in [0.7,0.8]$ the equilibrium is for workers to randomize between 0, 1, 2, 3 and 4, where 0, 2 and 4 are chosen with equal probability, as are 1 and 3. Bribes of 3 and 4 are added to the equilibrium since they are no longer strictly dominated by a bribe of zero, otherwise the intuition is similar to the above.

For $\delta \in [0.8,0.9]$, equilibrium randomization occurs between 0,1,2,3, and 5, again with 0 and 2 chosen equally often, and similarly for 1 and 3, with 5 chosen more often than any of the other bribes. For $\delta \in [0.9,1)$ workers randomize between 0,1 and 5, with 5 chosen most often. Intuitively, as the highest bribe starts winning with high regularity, bribing 5 becomes more attractive, which is amplified by bribes of 6 or higher not being possible in the experiment. Given that many workers bribe 5, it is no longer attractive to bribe 4 or other intermediate amounts, since these bribes will always lose to a bribe of 5 but still incur the certain cost of sending the bribe. However, bribes of zero are still attractive, since they are costless and have a positive probability of wining even against bribes of 5. Finally, bribes of 1 are also not very costly and have the added advantage of beating the 0 bribes with high likelihood.

In addition, for $\delta \in \{0.7,0.8,0.9\}$, any linear combination between both relevant aforementioned equilibria is also a ME. Finally, for $\delta = 1$, the equilibrium is a pure strategy equilibrium where both workers bribe the maximum. Note also that the fraction of referees who picked the highest bribe in treatment KeepBoth in the experiment was .6364. The mixed equilibrium corresponding to this fraction is for workers to bribe 0 or 2 with probability .267 and bribe 1 with probability .467.

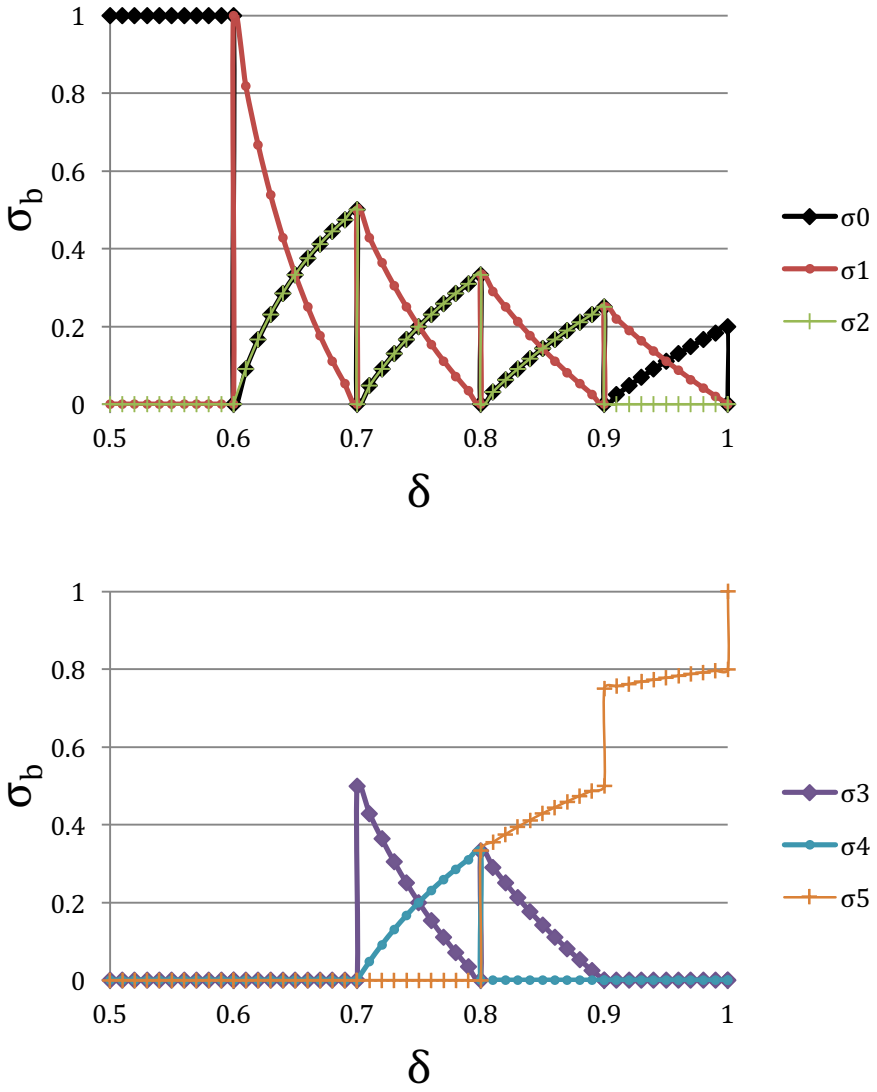

Figure 1.A1 Equilibria For Treatment Keepboth
Notes: The figure displays the probability σ_b that a bribe b is chosen in equilibrium as a function of the referee's allocation rule δ. Here, the referee's allocation rule specifies the fraction of times the referee chooses the highest bribe as the winner.

A9. Additional Figures

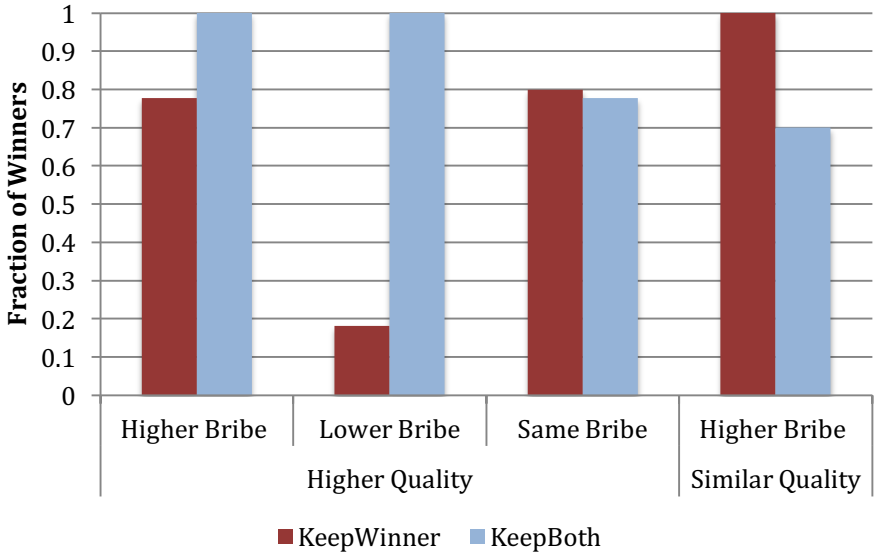

Figure 1.A2 Win chance for KeepWinner and KeepBoth

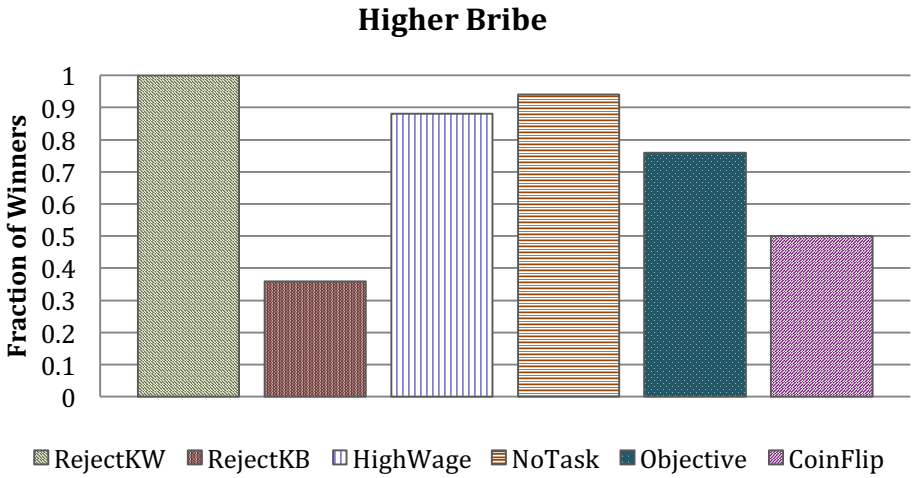

Figure 1.A3 Win chance when having the higher bribe for the additional treatments

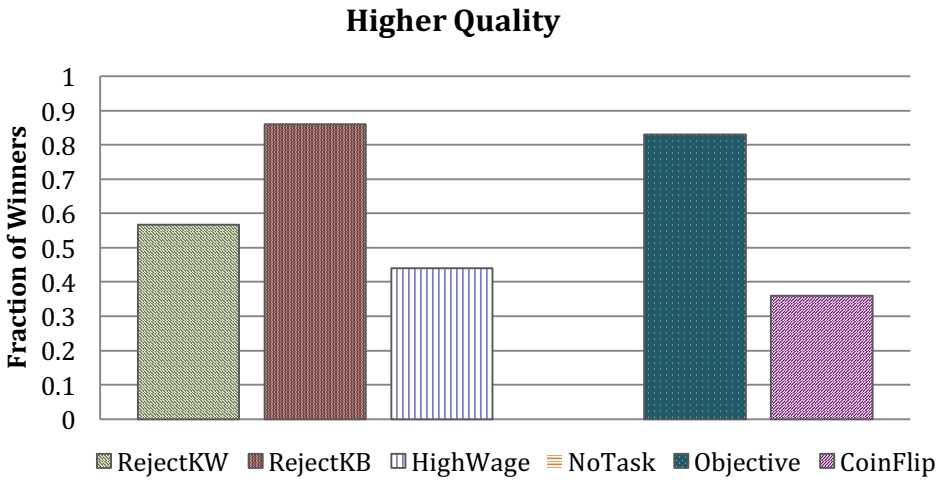

Figure 1.A4 Win chance when having the best joke for the additional treatments

Appendix B. Instructions

B1. Worker Instructions

Welcome to today's experiment.[4] Please read the instructions carefully. If you have any questions, please raise your hand and one of the experimenters will come to your desk to answer your question.[5]--next screen--

---

[4] These are the instructions for the KeepWinner treatment. The instructions for treatment Objective are presented below. The instructions for the other treatments are similar to KeepWinner and available upon request.

[5] A horizontal line indicates that participants went to the next screen.

You have been assigned to the role of Participant A. For the remainder of the experiment you will be matched with two other participants: Participant B and a Referee. The Referee will now be moved to a different room.

--next screen--

On your desk you can find an envelope with 10 dollars. This is your show-up fee for taking part in this experiment. Both you and Participant B have received a $10 show-up fee whereas the Referee has received a $5 show-up fee. Please do not remove the money from the envelope until you are instructed to.

Both you and Participant B will be asked to work on a task for two rounds. The task will be explained below. After each round the Referee will decide whether you or Participant B performed the task better. The Participant that performed better (as decided by the Referee) will receive an additional $10 prize on top of the show-up fee. The other Participant will receive nothing.

You will be matched to the same Referee and Participant B in both rounds. None of you will ever know the identity of the other two participants.
Do you have any questions before we explain the task to you?

--next screen--

Your task:
Your task is to come up with a joke about a certain topic, which will be announced after the instructions. In total, you will have 10 minutes to come up with a joke. The joke can be short or long, a simple one liner or a full anecdote. The experimenter will let you know when you have 5 minutes as well as 1 minute left for the round.

Check-up questions

How much will you earn (in dollars) in a given round if you are the winner?
Who is going to evaluate your task?
True or false: the Referee and Participant B will be the same participants in both rounds of the experiment.

You are now ready to start the experiment. Please raise your hand when you are ready to start the task. Do not proceed to the next page. The experimenter will instruct you to start when the other participants have finished reading the instructions.

Please write a joke about economists. You have 10 minutes to complete the task.

What do you believe is the probability that you wrote better jokes than Participant B?

Please wait while we are printing your joke. After you have received your joke, please put it into the large envelope with the number so it can be handed over to your Referee. You also have the option to add money for the Referee and put it in the envelope with the joke. For this purpose, you can take up to $5 out of the smaller envelope with your show-up fee and put it into the larger envelope together with the joke you wrote.

Participant B also has the option to add up to $5 to the envelope he/she sends to the Referee.

The Referee will be given both your envelope with the joke and the money and Participants B's. He/she will then be asked to read the jokes and decide which one wins. If the Referee chooses your joke, then you will get an additional $10 and the Referee will keep the money you sent him/her. Participant B will get the money he/she sent to the Referee back. If the Referee chooses Participants B's joke, then Participant B will get an additional $10 and the Referee will keep the money he/she sent to him. In this case you will get back the money that you sent to the Referee.

-- next screen--

Please raise your hand when the envelope for the Referee is ready. The experimenter will bring it to the Referee in the next room. After the Referee has determined the winner, the envelope will be collected by the experimenter. The envelope will be returned to you after the Referee has finished grading the second round of jokes.[6]

-- next screen--

B2. Referee Instructions

Welcome to today's experiment. Please read the instructions carefully. If you have any questions, please raise your hand and one of the experimenters will come to your desk to answer your question. --next screen—You have been assigned to the role of the Referee. For the remainder of the experiment you will be matched with two other participants: Participant A and Participant B.
Please raise your hand. The experimenter will escort you to a different room.

--next screen--

---

[6] Instructions for round 2 started from "please write a joke about economists" onwards and were identical to the instructions for round 1, except that workers were instructed to write a joke about psychologists instead.

On your desk you will find a small envelope with $5. This is your show-up fee for taking part in this experiment; Participant A and Participant B have received a $10 show-up fee for the experiment.

Today your task is to rate the quality of a joke written by Participant A and a joke written by Participant B. You will be matched to the same Participant A and Participant B in both rounds. None of you will ever know the identity of the other two participants.

Both Participants have 10 minutes to write a joke. After Participants A and B have finished their jokes, they will print them and put them in an envelope which will be brought to you by an experimenter.

You will then have 5 minutes to read both jokes and determine who of the two did the best job, i.e. determine the winner. The winner will receive a prize of $10, whereas the loser will receive nothing. Please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope.

You will also be asked to rate the quality of both the winner's and the loser's joke on a scale from 0 to 10 (on the evaluation form).

The envelopes will then be collected by the experimenter and you will be asked to grade a second round of jokes, written by Participants A and B while you were grading.

The envelopes for both rounds will be returned to Participants A and B at the end of the second round.

The topic for the first round will be 'economists', the topic of the second round will be announced to you after you finish grading the first round.

Please wait while Participants A and B finish writing their jokes. If you have any questions in the meantime, please ask the experimenter.

-- next page --

In a few moments you will receive two envelopes containing the jokes written by Participants A and B. To grade their jokes, please indicate you rating for both Participants on the evaluation form on a scale from 0 to 10. Participants A and B also have the opportunity to add money to their envelope. You have the option to keep the money sent to you by either Participant A or Participant B. If you keep the money of a Participant, he or she will automatically be the winner. The loser's money will then be returned.

After determining the winner, please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope. After five minutes, an experimenter will collect the envelopes. The envelopes will be returned to Participants A and B at the end of the second round.

Please remain patient while we are printing the jokes.[7]

B3. Independent Raters Instructions

Welcome!

In this experiment you will be shown six pairs of jokes. Jokes in each pair will either be about economists or about psychologists. Participants in a previous experiment wrote the jokes in 10 minutes. For each pair of jokes, you will be asked to rate the quality of both jokes and to indicate which one is better.

Please rate the quality of the following jokes about economists (psychologists). Make sure to read both jokes before rating.
What is the quality of this joke? (0-10)
Which one is the best joke? (Joke A, Joke B)

B4. Worker Instructions Treatment Objective

Welcome to today's experiment. Please read the instructions carefully. If you have any questions, please raise your hand and one of the experimenters will come to your desk to answer your question.

-- next section –

You have been assigned to the role of Participant A. For the remainder of the experiment you will be matched with two other participants: Participant A and a Referee.

The Referee will now be moved to a different room.

--next section--

On your desk you can find an envelope with 10 dollars. This is your show-up fee for taking part in this experiment. Both you and Participant B have received a $10 show-up fee whereas the referee has received a $5 show-up fee. Please do not remove the money from the envelope until you are instructed to.

Both you and Participant B will be asked to work on a task for two rounds. The task will be explained below. After each round the Referee will decide whether

---

[7] Instructions for round 2 contained the topic of the second round. Otherwise, they were identical to the last page of the instructions for round 1 (from "In a few moments" onwards).

you or Participant B performed the task better. Your goal is to complete as many words as possible in 5 minutes.

The Participant that performed better (as decided by the Referee) will receive an additional $10 prize on top of the show-up fee. The other Participant will receive nothing. You will be matched to the same Referee and Participant B in both rounds. None of you will ever know the identity of the other two participants.

Do you have any questions before we explain the task to you?

--next section--

During each round of the experiment you will be shown a sequence of words. These words will be printed in different colors: yellow, blue, purple, orange, or red. Your task is to indicate the color of each word. Only the colors named correctly will counts towards your total. This task will last for a total of 5 minutes.

You can indicate the color of your choice using the keyboard. The relevant keys are y (for yellow), r (red), p (purple), o (orange) and b (blue). The key-color combinations will also be visible at the bottom of the screen throughout the task. Be aware: if you press any key other than the one corresponding to the correct color, this will not be counted as a correct response. This also holds for keys that do not refer to any color. On the next page you will have the opportunity to practice the task with a sequence of 10 words.

--next section--

After you finish the task, your score will be printed on a score sheet that will be handed over to your referee. Your score sheet will be similar to the example below. Every color you successfully indicated will be represented by a dot on the score sheet.

--next section--

Please answer the following questions before proceeding to the next page.

Question 1:How much will you earn (in dollars) in a given round if you are the winner?

Question 2:Who is going to evaluate your task? Participant A, Participant B or the Referee?

Question 3:True or false: the Referee and Participant will be the same participants in both rounds of the experiment.

--next section--

You are now ready to start the experiment.Please raise your hand when you are ready to start the task.

Do not proceed to the next page. The experimenter will instruct you to start when the other participants have finished reading the instructions.

--next section--

What do you believe is the probability that you have a better score than Participant ?

--next section--

Please wait while we are printing your score sheet.

After you have received your score sheet, please put it into the large envelope with the number so it can be handed over to your Referee.

You also have the option to add money for the Referee and put it in the envelope with the score sheet. For this purpose, you can take up to $5 out of the smaller envelope with your show-up fee and put it into the larger envelope together with your score sheet.

Participant B also has the option to add up to $5 to the envelope he/she sends to the Referee.

The Referee will be given both your envelope with the score sheet and the money and Participant B's envelope. He/she will then be asked to determine which Participant wins. If the Referee decides that you win, then you will get an additional $10 and the Referee will keep the money you sent him/her. Participant B will get back the money he/she sent to the Referee.

If the Referee decides that Participant B wins, then Participant B will get an additional $10 and the Referee will keep the money he/she sent to him. In this case you will get back the money that you sent out to the Referee.

Please raise your hand when the envelope for the Referee is ready. The experimenter will bring it to the Referee in the next room.

--next section--

You are now ready to start round 2. This round will be similar to round 1: you will again have to indicate the color of a sequence of words and the task will again be graded by the Referee. Please remember that you will be matched to the same Referee

and the same Participant as before. You will again have the option to send money to the referee after you finish your task.[8]

[8] Instructions for round 2 were the same as round 1, starting from "what do you believe is the probability …".

B4. Referee Instructions Treatment Objective

Welcome to the experiment. On your desk you will find a small envelope with $5. This is your show-up fee for taking part in this experiment; Participant A and Participant B have received a $10 show-up fee for the experiment.

Today your task is to determine the score of Participant A and Participant B on a task. You will be matched to the same Participant A and Participant B in both rounds. None of you will ever know the identity of the other two participants.

Participant A and B's task is to determine the color of a series of words displayed on their computer screen. The participants will be shown a sequence of words one at the time and they will have to indicate the colors of the words. Their goal is to complete as many words as possible in 5 minutes. A screenshot of the task has been provided to you on a separate sheet.

Both Participants have 5 minutes for the task. After 5 minutes, their scores will be printed on a score sheet and each one of them will get his or her own printout. The printout score sheet will be similar to the sample score sheet provided to you as an example. Each color successfully determined by the participants will be represented by a single dot on the score sheet. Each participant will then put his/her score sheet in an envelope that will be brought to you by the experimenter.

You will then have 5 minutes to determine the winner. The winner will receive a prize of $10, whereas the loser will receive nothing. Please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope.

The envelopes will then be returned to Participants A and B and you will be asked to grade a second round of score sheets representing the number of colors successfully indicated by Participants A and B while you were grading.

Please wait while Participants A and B complete the first round. If you have any questions in the meantime, please ask the experimenter.

--next page--

In a few moments you will receive two envelopes containing the score sheets of Participants A and B.

Participants A and B also have the opportunity to add money to their envelope. You have the option to keep the money sent to you by either Participant A or Participant B. If you keep the money of a Participant, he or she will automatically be the winner. The loser's money will then be returned.

After determining the winner, please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope. After five minutes, the experimenter will collect the envelopes and return them to Participants A and B in the other room.

Please remain patient while we are printing the score sheets.

--next page--

Please wait while Participants A and B are finishing the second round. After Participants A and B have finished the second round, the procedure will be similar to round 1.

You will again receive two envelopes containing the score sheets of Participants A and B.

Participants A and B also again have the opportunity to add money to their envelope. You have the option to keep the money sent to you by either Participant A or Participant B. If you keep the money of a Participant, he or she will automatically be the winner. The loser's money will then be returned.

After determining the winner, please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope. After five minutes, an experimenter will collect the envelopes and return them to Participants A and B in the other room.

Please remain patient while Participants A and B are finishing the second round.

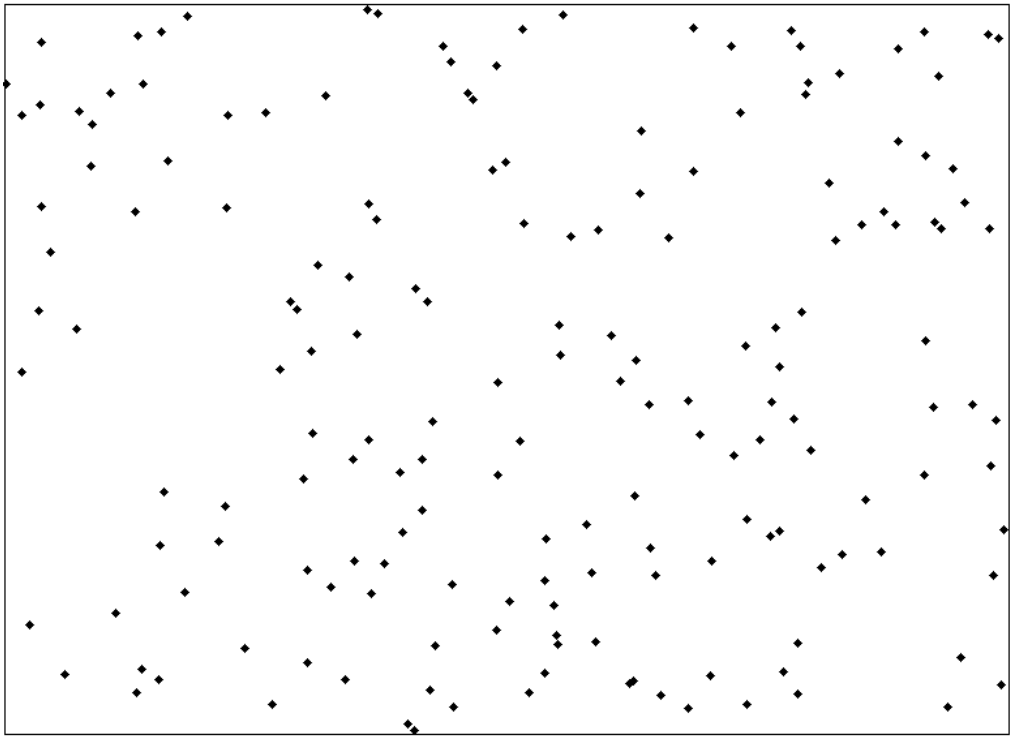

Figure 1.A5 Sample Score Sheet for Treatment Objective

Appendix C. Examples of Jokes

In this section, we present 9 examples of jokes written by participants in the experiment. The following jokes are the three best, the three worst and the three median jokes (as determined by the independent raters). All other jokes are available upon request.

C1. Good Jokes

A psychologist was conducting a group therapy session about addictions and obsessions, in which four mothers and their children were participating. Each of the mothers was asked by the psychologist to share their obsession as well as their kid's names. The first mother said, "I am obsessed with eating, and my daughter's name is Candy." The second mother said, "I am obsessed with money, and my daughter's name is Penny." The third mother said, "I am an alcoholic, my daughter's name is Brandy." The fourth mother got up, took her son by the hand, and whispered in his ear, "Come on, Dick, let's go home." (Average Rating 8/10)

A man is at the library and is trying to find an open seat to study. He finds an open spot next to an attractive young woman and asks if he can sit there. She responds rather loudly, "NO, I DON'T WANT TO SPEND THE NIGHT WITH YOU!" Everyone in the library turns to stare at the man. Embarrassed by the attention, the man goes on and finds another spot to study. Later, the same young woman goes up to the man and tells him, "I'm a psychologist and I study social behavior. I know I made you feel embarrassed, right?" The man looks up and responds rather loudly, "$200 DOLLARS JUST FOR ONE NIGHT?! THAT'S TOO MUCH!" Everyone in the library turns to stare at the young woman this time. The man then proceeds to tell her in a subdued voice, "I'm a lawyer. I know how to make people feel guilty." (Average rating 7.63/10)

Why did the psychologist get kicked out school? The professor caught him committing Freud (Average rating 6.6/10)

C1. Median Jokes

A psychologist, an economist, and a physicist were asked for their professional input on ways to improve execution by guillotine. The physicist said "To make the execution less painful, the blade should be heavier because then the blade will travel more quickly and kill the victim sooner". The economist said "The blade shouldn't be cleaned in between executions, because then you can save the cost of cleaning

supplies. They're going to die anyway, so sterilizing the blade isn't an ethical concern". The psychologist said "How do we know how much pain the person is in to make it less painful? I think we need more trials, but that's not possible because people only have one life to live. We should use cats! They have nine lives!" Everyone else decided to use the psychologist for a trial because they all owned cats. (Average Rating 3.5/10)

Economists. What my friends think I do: sit back and stare at money. What my parents think I do: earn money. What my colleagues think I do: scam money and help with money laundering. What the academics think I do: create awesome financial theories and win the Nobel Prize. What the public think I do: nothing. What I really do: look at lines and graphs all day long (Average Rating 3.5/10)

Three psychologists are looking up at the stars. The first, a Freudian, sees the Big Dipper in close proximity to Orion's belt and understands instantly the sexual frustration nestled there. The second, a social psychologist, scoffs, and asks the first what Cassiopeia's Little Dipper means, then. He sees the sky in aggregate, a multitude of decision-making stars cohering to a wider social contract. The third is silent. "Hey, Silence of the Lambs," the Freudian psychologist calls out. "Who's right about the stars?" Number Three, an abnormal psychologist, is rather convinced that the stars are, in fact, a 1970s construct remnant from the Star Wars campaign, part of a government conspiracy, and also happen to be transmitting this very conversation to (secretly) Red Russia. Then a goat comes along and speaks. None pay it any heed. (Average Rating 3.5/10)

C1. Bad Jokes

One economist one day went to the shooping centre to buy a keyboard, the price labbled on the hat was $59.99. While the keyboard is using solar as its battery, he start to computer the profit he can get from the keyboard. Since the solar keyboard is much expensive than the normal one, he think that he can use it 3 years, and if he uses the normal keyboard the battery is ... As he thinking, here is a college student came to the store, he bought the keyboard without thinking, and the solar keyboard is out of stock! (Average Rating 1.1/10)

One day the economics was walking beside the beach and began to wonder what the white coloring was up ahead. / Once the economics had reached it then it suddenly had the realization that it was the face of mist. Economics decided to walk into it, and the mist decided to walk through economics as well....and what became of them after that?...That was when economists began their journey! (Average Rating 1.1/10)

A man asked a Psychologist, "Sir, I had a dream that I was swinging on a tree swing like the one in my old house when I younger. The trees were sturdy Oaks that were many years old, and I remember their branches being slightly gnarled. I remember the sun peeking through the leaves and my mother called my name, but as I was running to reach her, the ground opened up and swallowed me. What can this

mean?" The Psychologist looked at the man with a furrowed brow, leaned his head back and stared at the ceiling. "Well, the branches mean you are very sexually repressed as they blocked your view of the sun clearly, and that the woman of your dreams is your mom. Clearly you have an Oedipus complex, as the ground breaking up is a sign of your father stopping you from gaining access to your mother. All in all, you love your mom and need to kill your dad." The man blinked a few times, then stood up. "You made that up didn't you?" he asked the Psychologist. With a hearty sigh the Psychologist sat up straight and looked

References

Abbink, Klaus. 2004. “Staff Rotation as an Anti-corruption Policy: An Experimental Study.” *European Journal of Political Economy* 20 (4): 887–906.

Abbink, Klaus. 2005. “Fair Salaries and the Moral Costs of Corruption.” In *Advances in Cognitive Economics*, ed. Boicho N. Kokinov, 2–7. Sofia: NBU Press.

Abbink, Klaus, Utteeyo Dasgupta, Lata Gangadharan, and Tarun Jain. 2014. “Letting the Briber Go Free: An Experiment on Mitigating Harassment Bribes.” *Journal of Public Economics* 111: 17–28.

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner. 2002. “An Experimental Bribery Game.” *Journal of Law, Economics, and Organization* 18 (2): 428–454.

Abbink, Klaus, and Danila Serra. 2012. “Anticorruption Policies: Lessons from the Lab.” In *New Advances in Experimental Research on Corruption (Research in Experimental Economics, Volume 15)*, ed. Danila Serra and Leonard Wantchekon, 77–115. Emerald Group Publishing Ltd.

Akerlof, George A. 1982. “Labor Contracts as Partial Gift Exchange.” *The Quarterly Journal of Economics* 97 (4): 543.

Armantier, Olivier, and Amadou Boly. 2013. “Comparing Corruption in the Laboratory and in the Field in Burkina Faso and in Canada.” *The Economic Journal* 123 (573)): 1168–1187.

Banerjee, Abhijit V. 1997. “A Theory of Misgovernance.” *The Quarterly Journal of Economics* 112 (4): 1289–1332.

Barr, Abigail, and Danila Serra. 2010. “Corruption and Culture: An Experimental Analysis.” *Journal of Public Economics* 94 (11-12): 862–869.

Bazerman, Max H., George Loewenstein, and Don A. Moore. 2002. "Why Good Accountants Do Bad Audits." *Harvard Business Review* 80, no. 11: 96-103.

Becker, Gary S., and George J. Stigler. 1974. “Law Enforcement, Malfeasance, and Compensation of Enforcers.” *The Journal of Legal Studies* 3 (1): 1–18.

Belot, Michèle, and Marina Schröder. 2013. “Sloppy Work, Lies and Theft: A Novel Experimental Design to Study Counterproductive Behaviour.” *Journal of Economic Behavior & Organization* 93: 233–238.

Bertrand, M., S. Djankov, R. Hanna, and S. Mullainathan. 2007. "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption." *The Quarterly Journal of Economics* 122 (4): 1639–1676.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90 (1): 166–193.

Burguet, Roberto, and Yeon-Koo Che. 2004. "Competitive Procurement with Corruption." *RAND Journal of Economics* 35 (1): 50–68.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3): 414–427.

Cameron, Lisa, Ananish Chaudhuri, Nisvan Erkal, and Lata Gangadharan. 2009. "Propensities to Engage in and Punish Corrupt Behavior: Experimental Evidence from Australia, India, Indonesia and Singapore." *Journal of Public Economics* 93 (7-8): 843–851.

Cappelen, Alexander W., Erik Ø. Sørensen, and Bertil Tungodden. 2013. "When Do We Lie?" *Journal of Economic Behavior & Organization* 93: 258–265.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn. 2013. "Experimental Methods: Extra-Laboratory Experiments-Extending the Reach of Experimental Economics." *Journal of Economic Behavior & Organization* 91: 93–100.

Del Monte, Alfredo, and Erasmo Papagni. 2001. "Public Expenditure, Corruption, and Economic Growth: The Case of Italy." *European Journal of Political Economy* 17 (1): 1–16.

Dreber, Anna, and Magnus Johannesson. 2008. "Gender Differences in Deception." *Economics Letters* 99 (1): 197–199.

Erat, Sanjiv. 2013. "Avoiding Lying: The Case of Delegated Deception." *Journal of Economic Behavior & Organization* 93: 273–278.

Erat, Sanjiv, and Uri Gneezy. 2012. "White Lies." *Management Science* 58 (4): 723–733.

Falk, Armin, and Nora Szech. 2013. "Morals and Markets." *Science* 340 (6133): 707–11.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90 (4): 980–994.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics* 114 (3): 817–868.

Fitzgerald, Patrick. 2008. "Justice Department Briefing on Blagojevich Investigation." *The New York Times*. http://www.nytimes.com/2008/12/09/us/politics/09text-illinois.html?_r=0.

Gneezy, Uri. 2005. "Deception: The Role of Consequences." *American Economic Review* 95 (1): 384–394.

Gneezy, Uri, Marta Serra-Garcia, Silvia Saccardo, and Roel van Veldhuizen. 2015. "Motivated Self-Deception, Identity and Unethical Behavior." *Mimeo*.

Goldsmith, Arthur A. 1999. "Slapping the Grasping Hand: Correlates of Political Corruption in Emerging Markets." *American Journal of Economics and Sociology* 58 (4): 865–883.

Huntington, Samuel P. 1968. *Political Order in Changing Societies*. New Haven: Yale University Press.

INDEM. 2005. "Diagnosing Corruption in Russia: A Sociological Analysis." *Russian Social Science Review* 46 (1): 19–36.

Kaufmann, Daniel. 2005. "Myths and Realities of Governance and Corruption." In *Global Competitiveness Report 2005-06*, 81–98. World Economic Forum.

Leff, Nathaniel H. 1964. "Economic Development Through Bureaucratic Corruption." *American Behavioral Scientist* 8 (3): 8–14.

Lightle, John P. 2013. "Harmful Lie Aversion and Lie Discovery in Noisy Expert Advice Games." *Journal of Economic Behavior & Organization* 93: 347–362.

Malmendier, Ulrike, and Klaus M. Schmidt. 2012. "You Owe Me." *NBER Working Paper No. 18543.*

Mauro, Paulo. 1995. "Corruption and Growth." *Quarterly Journal of Economics* 110 (3): 681–712.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–249.

Olken, Benjamin A., and Rohini Pande. 2012. "Corruption in Developing Countries." *Annual Review of Economics* 4: 479–509.Pareto, Vilfredo. 1896. "Course of Political Economy." In *Sociological Writings*. New York: Praeger, 1966.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83 (5): 1281–1302.

Reinikka, Ritva, and Jakob Svensson. 2004. "Local Capture: Evidence from a Central Government Transfer Program in Uganda." *The Quarterly Journal of Economics* 119: 679–705.

Sequeira, Sandra, and Simeon Djankov. 2014. "Corruption and Firm Behavior: Evidence from African Ports." *Journal of International Economics* 94 (2): 277–94.

Stroop, John R. 1935. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology* 18 (6): 643–662.

Sutter, Matthias. 2009. "Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams." *The Economic Journal* 119 (534): 47–60.

Transparency International. 2011. "Bribe Payers Index 2011." http://bpi.transparency.org/bpi2011/results/.

Transparency International. 2014. "Corruption Perceptions Index 2014." https://www.transparency.org/cpi2014/results.

Transparency International India. 2008. "India Corruption Study - 2008." http://www.transparencyindia.org/resource/survey_study/India%20Corruptino%20Study%202008.pdf.

Van Veldhuizen, Roel. 2013. "The Influence of Wages on Public Officials' Corruptibility: a Laboratory Investigation." *Journal of Economic Psychology* 39: 341–356.

2. Motivated Self-Deception, Identity, And Unethical Behavior

Abstract

We examine the role of self-deception in distorting judgment. In two experiments, we vary when evaluators are informed about incentives to recommend one of two options: before or after their initial private judgment. When the information regarding the incentives is provided before, we find a significant bias in judgment in the direction of the incentive. However, when the information is provided after they privately evaluate the options, the bias in judgment is significantly reduced. We term this behavior "motivated self-deception," arguing that in the before treatment judgment is biased such that evaluators can earn more money without compromising their self-image.

2.1 Introduction

Unethical behavior, such as corruption or dishonesty, is widespread and comes with efficiency and fairness costs. A large fraction of companies report that it is often necessary to pay bribes to win business—from 15%-20% in industrialized countries to 40% in China, Russia, and Mexico (Transparency International, 2011). Instances of fraud, including examples covered in the media by both firms (Enron, Worldcom, VW) and individuals (Bernie Madoff, Diederik Stapel), can have large consequences on efficiency: for example, the bankruptcy of Enron led to the loss of 4,000 jobs and employees were left with worthless savings plans (for a discussion of the efficiency costs of corruption, see Banerjee, 1997, and Svensson, 2005).

For some people who are involved in such behavior, no psychological cost is associated with it. But for others, distorting ethical judgment comes with a cost to self-image (Bem, 1972; Akerlof and Kranton, 2000; Bénabou and Tirole, 2006; Mazar, Amir and Ariely, 2008; Ariely, Bracha and Meier, 2009; Gneezy et al., 2012). All else equal, people who have such psychological costs prefer an outcome that is achieved without unethical choices to one that requires an action they consider unethical. To avoid this cost to self-image, people may choose actions that reduce their material payoffs. This conflict creates a tension between maintaining the self-image as a moral person and the desire to increase material goals. However, this tension may be attenuated if individuals can convince themselves that their behavior is ethical.

Consider the healthcare sector, where overtreatment is estimated to cost $210 billion in wasteful annual spending in the US (IOM 2012), as well as obvious non-

monetary costs to patients. One possible reason for overtreatment is that doctors frequently have financial incentives to recommend certain procedures for which they are directly compensated (Emanuel and Fuchs, 2008; Clemens and Gottlieb, 2014). Take for example the growing number of surgeries in response to back pain, many of which have been shown to be unnecessary and even harmful (Mafi et al., 2013). Another example is the large fraction of doctors who recommend unneeded C-sections for birth delivery when such procedures are financially compensated (see e.g., Gruber, Kim and Mayzlin, 1999; Johnson and Rehavi, 2014).

Some doctors may recommend unnecessary care knowingly in order to earn more money. Others, given that medical judgment is partially subjective, may convince themselves that the treatment they are prescribing is needed, thereby preserving their self-image as ethical professionals. In general, when judgment is subjective, evaluators may form their recommendation in a self-serving manner, to preserve their identity. We call this behavior "motivated self-deception," where the decision maker can convince herself that her behavior is ethical, preserving her identity as a moral person, while choosing the option that increases her personal gain.

The main question we ask in this paper is whether the evaluator knows that her evaluation is biased. That is, do evaluators distort their judgment knowingly, or do they engage in self-deception, convinced that their choice is ethical. In the experiment we report in this paper, an advisor is asked to recommend one of two investment choices to a client. The two options differ in risk and expected return, there is no correct or incorrect choice, and the advisor receives a bonus if he recommends a specific investment option.

The key experimental manipulation contrasts two timelines of decision-making. In the first, the evaluator is told about the incentives to choose one of the options *before* she is presented with the options she needs to consider. In the second, the evaluator is told about the incentives only *after* seeing the two options and being asked to consider which one she would recommend. Importantly, in both cases, the evaluator knows about the incentives before we observe her choice. That is, any judgment about the different options before providing the final recommendation only occurs in the evaluators' mind.

If the evaluator is informed about the incentives before evaluating the options, she might be biased in her evaluation, without even realizing she is. If she first decided about her choice, and only then learns about the incentives, she might still recommend the option for which she is incentivized, but she would not be able to maintain the self-image of ethical choice.

Comparing these two basic manipulations allows us to answer what we called the main question above: delaying the information regarding the incentives results in a significant reduction in the proportion of evaluators who favor the incentivized option. Advisors are less likely to choose the incentivized investment choice. As in the doctors' example above, some people choose the incentivized option in any case, but a large portion does so only when able to convince themselves that they are not cheating.

In an additional treatment, we further show that when the evaluation task is objective such that one investment option strictly first-order stochastically dominates the other, and therefore advisors cannot convince themselves that the incentivized

option is the ethical one, delaying the information about incentives has no effect on recommendations. This offers support to our argument that the difference in recommendations arising from a delay in the information regarding incentives in a subjective task is caused by self-deception.

Put together, our results support the hypothesis that incentives influence judgment to a much greater degree when evaluators can convince themselves they are not distorting judgment, and hence reduce the cost to self-image. Our findings suggest a simple solution to some of the biased outcomes discussed above: separating the evaluation task from the information about incentives such that evaluations are formed before incentives can distort judgment. For those who consider themselves ethical but may fall prey to motivated self-deception, this would prevent or lessen the extent of their unethical behavior.

2.2 Related Literature

Our paper is motivated by the early work of Freud (1933) and Festinger (1957) on cognitive dissonance, and subsequent work in psychology on motivated reasoning (Lord, Ross and Lepper 1979; Kunda, 1990). This work suggests that individuals adjust their cognitions to reduce or eliminate two conflicting desires. In this paper we investigate whether such adjustment can occur in the realm of unethical behavior, where individuals face a conflict between achieving their material self-interest and maintaining a positive self-image as ethical. That is, we investigate whether individuals engage in motivated self-deception to eliminate this tension.

Gur and Sackeim (1979) define self-deception as a situation in which an individual holds two contradictory beliefs without being aware of holding one of them, and such awareness is motivated. Self-deception has been studied in the context of assessments of own ability, to explain phenomena such as overconfidence (e.g., Trivers, 2011). In economics, Benabou and Tirole (2002) introduced a model with a dual-self to study self-deception. Two selves characterize an individual: one self who receives information about the individual's ability, but may suppress it, to induce the other uninformed self to exert higher effort levels.[1] In this model, equilibria exist whereby the individual engages in systematic denial of negative signals regarding his ability. Chance et al. (2011) provide experimental evidence showing that individuals, who were given the answers to a test while performing it, erroneously predicted a higher performance in future tests. This behavior is consistent with self-deception, but could also be explained by biased ex-post rationalization, whereby individuals think about their past performance and rationalize their success as a product of their own ability rather than the availability of the answers.

In terms of experimental procedure, the literature on self-serving biases in negotiation (Babcock et al., 1995; Babcock and Lowenstein, 1997), is closest to our work. Participants in their experiment were given a set of legal documents and were asked to study them before negotiating over a settlement. Differently from our set up, individuals in the before treatments in Babcock et al (1995) are asked to provide their

---

[1] See also the related models by Bodner and Prelec (2003), in which individuals may choose their beliefs self-servingly, and Brocas and Carrillo (2008), who model asymmetric information conflicts using a dual-self model of the brain, based on neuroscience evidence.

written assessment about fair outcomes before knowing their roles in the negotiation, while in the latter treatment, they provide their assessed fair outcome after knowing their roles. When participants learned their role after reading the legal documents, they were more likely to reach an agreement than when they knew their role before reading the instructions. Because individuals' views of what is fair are biased in a self-serving way, bargaining impasses in court settlements decrease when individuals read the case documents before being informed about the party they represent, compared to after.

The before and after manipulations in the current paper are built on this idea, with some important differences. First, our experiment studies individuals' desire to preserve their self-image as moral while behaving unethically, and does not regard their fairness assessment of a situation. Second, because fairness considerations were done before negotiating with other people, the evaluation included strategic element in it. Third, because the goal of the Babcock et al (1995) paper to study self-serving biases and not self-deception, they used written assessment. Providing fairness assessment in writing makes the original decision harder to ignore using self-deception.

Konow (2000) examines self-deception in the context of fairness. His paper provides a theoretical model of self-deception, where individuals consciously trade-off the benefits and costs of self-deception. He also shows experimentally that individuals, who adopt a self-serving fairness criterion when distributing monetary payments between themselves and another person, continue to apply this self-serving bias when distributing monetary payments between two other individuals.

Two concepts related to motivated self-deception are self-serving justifications (or rationalizations) and bounded ethicality. Several experiments in psychology show that providing individuals with scope for justifying unethical behaviors increases the frequency of such behaviors (e.g., Mazar et al., 2008; Shalvi et al., 2011 and 2012). For example, Shalvi et al. (2011) use the die-roll paradigm of Fischbacher and Foellmi-Heusi (2013), whereby an individual is asked to roll a die in private and report the outcome to the experimenter. The higher the reported outcome, the higher is the payoff that the individual receives. Shalvi et al. (2011) find that individuals are more likely to report a large outcome when they are asked to roll the die three times, but report the outcome of the first roll, compared to when they are only asked to roll the die once. Their study suggests that when opportunities to justify dishonesty are readily available (by reporting the outcome of the second or third roll, instead of the first), individuals are more likely to cheat.

Finally, the concept of bounded ethicality, introduced by Chugh, Bazerman and Banaji (2005), refers to the fact that ethical judgment is bounded in ways that unconsciously favor a particular vision of the self. Hence, an individual's desire to view herself as moral and competent prevents her from identifying conflicts of interest that involve the self. Self-deception may be considered one of the ways in which individuals exhibit bounded ethicality. Our paper contributes to this literature by studying how incentives can facilitate the scope for motivated self-deception, and thereby contributes understanding the persistence of unethical behavior.

2.3 The Distorted Advice Experiment

2.3.1 The Setting

In this experiment, we study a sender-receiver game in which the sender ("advisor") is informed about the details of two investment opportunities, A and B, and is asked to send a recommendation to an uninformed receiver (the "client") regarding which of the two to choose. This game differs from standard sender-receiver games in that the sender is asked to make a judgment instead of reporting an objective piece of information, such as the state of nature (Crawford and Sobel, 1982).

The timeline of the experiment was as follows. First, the advisor was presented with information regarding the investment opportunities, A and B. Then she wrote a message recommending an option to a client. The client was a participant in a different experimental session and received no information about A and B. He only received the recommendation of the advisor and was asked to choose between A and B.

The information was presented to the advisors on four separate pages on their computer screen (all instructions are provided in the Appendix). On the first page, the advisor was informed about her role in the experiment and that she would be given a fixed payment of $1 for participation. She was told that her role in the experiment would be to recommend one of two investment options (A and B) to another participant in a different session. She also learned that the other participant received no information about A or B except her recommendation.

On the second page of the instructions, advisors were presented with the details of A and B. The investment opportunities, labeled as product A and B, were described

as having a 50% chance of being of high quality and a 50% chance of being of low quality. The payoff to the client for investment A was a 50-50 lottery between $2 and $4. Investment B was a 50-50 lottery between $1 and $7 dollars. The expected payoff of B ($4) was higher than that of A ($3). However, B had a higher variance. Thus, a tradeoff existed between risk and return across the two lotteries, such that the advisor could justify either choice by arguing (to herself) that risk or return was the more important factor for the recommendation.

In addition to receiving information about the lotteries, the advisor was asked (at the bottom of the screen) to think about her recommendation and continue to the next screen once she was ready to provide it. Once the advisor moved to the third screen, the instructions asked her to raise her hand so that the research assistant could bring her the paper on which she would write her recommendation. Once she received the paper, she was asked to move onto the fourth and final screen, where she would provide her recommendation both on paper and on screen. This procedure allowed us to have the advisor send a message in her own handwriting, making the recommendation more tangible, as well as have a direct electronic record of recommendations.

The experiment had three treatments. In the Control treatment, advisors received no additional payment for recommending A or B. In the Before and After treatments, the advisor was told she would receive an additional commission of $1 if she recommended A. The key difference between the Before and After treatments was *when* the advisor was first informed about the additional payment. In the Before treatment, advisors learned this information on the first screen, *before* learning the

details of the investments. By contrast, in the After treatment, the advisor learned about the commission only on the fourth and final screen, *after* reading about the investments and having already thought about her recommendation, but before making the recommendation. To introduce only one change across treatments, the information on the commission was also presented on the fourth screen in the Before treatment.

If the only factor affecting which product the advisor recommends is the incentive, we should see no difference between the Before and After treatments. Assuming advisors are self-interested, and assuming they expect the client to follow their recommendation, they would recommend A in both treatments. If their self-image cost of distorting judgment is large enough, and the timeline of the experiment does not bias their evaluation of A and B, they would recommend A at the same rate in both treatments as in the Control treatment.

However, if the timing of the information about the incentive affects self-deception, whether the advisor knows about the commission of $1 *before* or *after* reading about the investments may make a big difference. In the Before treatment, self-deception is easier, because the advisor learns about the incentives before seeing the products and may be able to convince herself that risk is undesirable, thereby giving her a reason to recommend A.

By contrast, in the After treatment, self-deception is harder. Here, the advisor has already made a decision about her evaluation of the tradeoff between risk and return before receiving information about incentives. Having initially decided to favor B, changing her recommendation to A may come at a cost to self-image, because she cannot deceive herself. Hence, motivated self-deception predicts that advisors will

recommend A more often in the Before treatment than in the After and Control treatments.

2.2.2 Procedures

We conducted the experiment at the University of California, San Diego. Participants took part in an hour-long experimental session involving other studies. The experiment was run during two weeks and the order of studies in a session was the same within each week.[2] Randomization across the three treatments occurred at the participant level. As mentioned above, instructions were presented on computer screens and participants were asked to submit their recommendation for the client on a separate piece of paper, which only included the message "I recommend you to choose Product (A or B) _____."[3]

We aimed at collecting 100 observations per treatment. Sessions were run for a whole day, and we stopped collecting data at the end of the day in which we achieved 300 observations. In total, 324 participants provided their recommendation as advisors (106 in Before, 110 in After, and 108 in Control). Forty-six percent of participants were female and the average age was 21.

One out of every ten recommendations was randomly selected and given to a client in different sessions. Because the total number of recommendations was not a multiple of 10, we rounded it up and provided 33 recommendations to 33 clients.

---

[2] All other studies in a session were not incentivized and unrelated to our study. They were surveys in the fields of marketing and management, remained always the same and were presented in the same order within a week.

[3] Some participants (34 out of 324) did not follow the instructions as indicated. They did not raise their hand to request the paper for the message. We leave these participants in the sample and thereby report results conservatively. If we exclude these participants from the sample, results are strengthened in the direction of our prediction.

A majority of the clients, 28 (84.8%) out of 33, followed the advisor's recommendation. We found no difference in following depending on the recommendation, A or B (11 (91.7%) out of 12 and 17 (81%) out of 21, respectively; Fisher's exact test, p=.630). Hence, the advisor's recommendation had a high chance of directly affecting the client. In what follows, we focus on the behavior of advisors and examine the treatment effects on advisors' recommendations.

We ran a second experiment in a different domain to examine the robustness of our results to a different setting in which there also is scope for self-deception. In this experiment, a referee is asked to award a prize to one of two workers according to their performance on a subjective real-effort task. Workers are given the opportunity to send money to the referee to influence her judgment. The same qualitative findings are obtained as in the main experiment and hence, for brevity, we report its results in Appendix A.

2.2.3 Results

Figure 1 displays the fraction of advisors recommending investment A in the three treatments. When information about the incentive tied to A is provided *before* reading about A, advisors are significantly more likely to recommend it. They recommend A in 43.4% of the cases in the Before treatment, compared to 27.7% of the cases in the After treatment. The difference is statistically significant (test of proportions, Z-stat=2.481, p=.013).

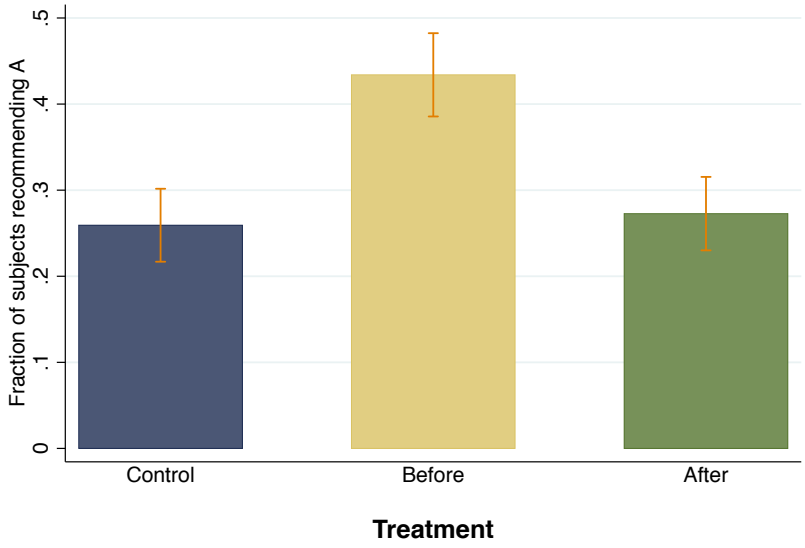

Figure 2.1 Fraction of Advisors Recommending A

Note: The figure presents the fraction of advisors who recommended option A in each of the three treatments respectively. The error bars represent +/- 1 S.E.

The percentage of advisors recommending A is 25.9% in the Control treatment and does not differ significantly from that in the After treatment (Z=0.225, p=.822). It does differ significantly from Before (Z=2.687, p=.007). Hence, we observe that the $1 commission does not significantly distort judgment when announced after the information regarding the two lotteries; relative to control, the change is from 25.9% to 27.7%. However, it leads to a significant bias in recommendations when announced before the information on the lotteries, increasing A recommendations to 43.4%.

Table 1 below confirms these results in a probit regression analysis. Column (1) confirms the average effect of Before relative to After: the likelihood of

recommending A increases by 0.15 in the Before treatment, relative to the After treatment.[4]

We extend our analysis to examine the role of gender. Previous research has shown that females are more risk averse (e.g., Croson and Gneezy, 2009). In column (2), we introduce a control for gender, and in line with previous findings, we find that females are more likely to recommend A, which has a lower variance than B. In other words, female risk aversion appears to be reflected in female recommendations to others.

Next, we examine whether the treatment effects vary by gender. Trivers (2011) suggests that men are more prone to self-deception than women. In the experiment, self-deception occurs only if an advisor would have recommended B in the absence of the incentive. Because a larger fraction of women recommends A in the Control treatment, the difference between Before and After may vary by gender. Columns (3) and (4) report the treatment effects splitting the sample by gender. We observe that whereas the effect of Before is strongly significant for men ($p<.001$), it is not significant for women ($p=.176$). However, the increase in A recommendations among men in the Before treatment, 16.5 percentage points, is not significantly different from

[4] We conducted additional experiments, as will be described in what follows. In parallel, we conducted a Replication Experiment, in which we replicated this experiment to test its robustness to cohort effects in our subject pool. We recruited an additional 311 advisors following the same procedures (104 in the Control treatment, 103 and in the Before treatment and 104 in the After treatment). We obtained even stronger treatment effects than in this experiment. There was a significant difference in A recommendations between the Before (60.2% of the cases) and the After (34.7% of the cases) treatments ($Z=3.826$, $p<.01$). In the Control treatment, A was recommended in 29.8% of the cases, which is not significantly different from the frequency of A recommendations in the After treatment ($Z=0.596$, $p=.551$).

that among women, 12.5 percentage points (p=0.319). Thus, in the context of our experiment there is limited evidence of a gender difference in self-deception.

Table 2.1 Treatment effects on the Likelihood that A is Recommended

	(1)	(2)	(3)	(4)
		P(A is recommended)		
	All	All	Male	Female
Before Treatment	.146***	.159***	.165***	.125
	(.051)	(.049)	(.044)	(.093)
Control Treatment	-.014	-.007	-.009	-.002
	(.061)	(.059)	(.071)	(.095)
Female		.178***		
		(.045)		
Share recommending A in After Treatment	.273	.273	.164	.382
Observations	324	324	174	150

Notes: Columns (1) to (4) report marginal effects from probit regressions on the likelihood that A is recommended. In columns (1) and (2), all advisors are included, whereas column (3) reports results only for male advisors and column (4) only for female advisors. The variables 'Before Treatment' and 'Control Treatment' are dummy variables taking value 1 if the treatment is Before or Control, respectively. The omitted category is the After treatment. Female is a dummy variable that takes value 1 if the participant is a female. Marginal effects are evaluated for a man (column 2) in the After treatment (columns 1 to 4).

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Our results support the prediction that providing incentives to recommend A leads to a stronger bias towards this option when the information about the incentive is revealed *before* the advisor evaluates the two options (A and B) than when it is revealed *after* the options have been privately evaluated. This suggests that motivated self-deception may indeed have been harder in treatment After. When advisors were informed about the incentives before evaluation, motivated self-deception may have allowed them to color their judgment in the direction of the incentives. By contrast, when they were informed about the incentives after the initial evaluation, judgment was less biased.

Our results could also be consistent with two alternative explanations. One alternative explanation is that participants in treatment Before may have avoided evaluation altogether and simply recommended the incentivized option, either because of the incentives per se, or because they perceived the incentives as a signal that the incentivized option was in fact the better product. This would imply that participants in treatment Before require less time to finish the experiment. However, we do not find a significant difference in the time taken to complete the experiment between the Before and After treatments (Mann-Whitney test, p=.170), or relative to Control (Mann-Whitney test, Before vs. Control, p=.215; After vs. Control, p=.829). Second, the smaller bias could also result from preferences for consistency (see Cialdini, 1984) according to which advisors in the After treatment might have a preference to stick to the first judgment they formulated in their minds. We provide further evidence in support of self-deception, and against these alternative explanations, in an additional experiment in which we remove any scope for self-deception.

2.3 Limiting the scope for motivated self-deception

According to our prediction, motivated self-deception would occur only when judgment is subjective. When evaluation occurs on multiple dimensions, such as risk and return, and no option strictly dominates the other, individuals can choose the dimension they consider most relevant. Given that such choices are subjective, there is scope for participants to convince themselves that the dimension that is materially more advantageous to them is the most important. When *strict* dominance in all

dimensions is introduced, there is no scope for motivated self-deception and hence we predict that the timing of the information regarding the incentives will not affect choices.

2.3.1 Strict Dominance Experiment

We ran an additional experiment with strict dominance between investments A and B. The only change relative to the previous experiment was the value of B: a 50-50 lottery between $5 and $7, instead of $1 and $7. Investment A remained unchanged-- a 50-50 lottery between $2 and $4. Thus, in this experiment investment B strictly first-order stochastically dominated A.

As in the Distorted Advice Experiment, there were three treatments. In the Control treatment, there was no additional incentive for recommending A or B. In the Before and After treatments, advisors received a commission of $1 for recommending A. Advisors were informed about the commission either *before* or *after* evaluating A and B.

Introducing first-order stochastic dominance in the experiment removes the scope for motivated self-deception, since advisors cannot any longer convince themselves that A is the better option, as B strictly dominates A in every state of the world. Therefore, we predict no difference between Before and After in this experiment.

Importantly, this prediction of no difference between Before and After in this experiment also allows us to address the two alternative explanations discussed above,

and for which the prediction for the new experiment is the same as the prediction for the first Distorted Advise experiment—a difference between Before and After. First, if the difference between Before and After is driven by participants avoiding evaluation in Before, we would still expect a higher rate of A recommendations in this treatment than in the After treatment. Second, if preferences for consistency would explain the lower rate of A recommendations in After, because individuals stick with the judgment formed before learning about the incentive, we would also still expect a difference in recommendations between Before and After.

The procedures followed in this experiment were the same as in the Distorted Advice Experiment. We recruited 334 participants who provided their recommendation as advisors (113 in Control, 109 in Before, and 112 in After). Fifty-four percent of participants were female and the average age was 21.

A majority of the clients, 25 (73.5%) out of 34, followed the advisor's recommendation. We found no difference in following depending on the recommendation, A or B (8 (80%) out of 10 and 17 (70.8%) out of 24, respectively; Fisher's exact test, p=.692). Hence, as in the other experiment, the advisor's recommendation had a high chance of directly affecting the client.

2.3.2 Results

Figure 2 displays the fraction of advisors recommending A in each treatment. In the Control treatment, where advisors do not receive any incentive for recommending A or B, 15.9% recommend A. When an incentive to recommend A is introduced, the rate of A recommendations increases to 31.2%. Importantly, the

increase in the Before treatment is not significantly different from the increase of 30.4% in the After treatment. The difference between the Before and After treatments is not significant (Z=0.1345, p=.893).

The presence of the incentive significantly increases A recommendations in both Before and After, relative to control (Z=2.685, p=.007 comparing Before and Control; Z=2.567, p=0.010 comparing After and Control).

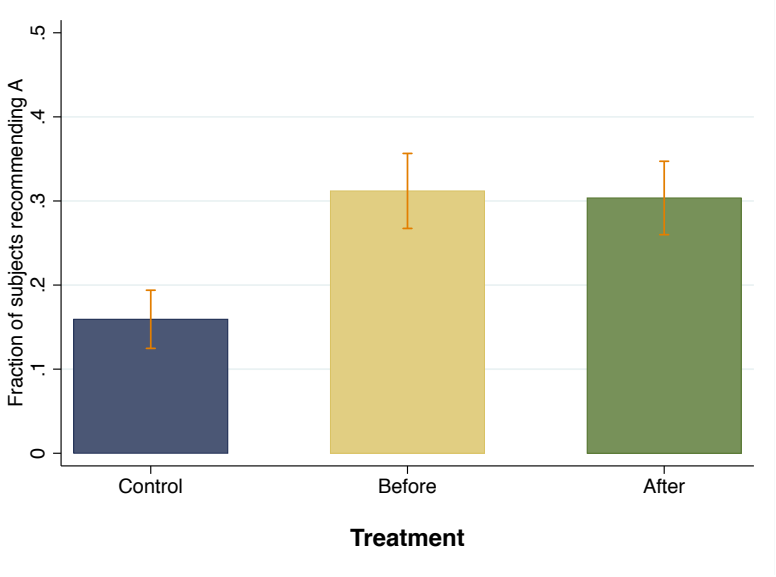

Figure 2.2 Fraction of Advisors Recommending A in the Strict Dominance Experiment

Note: The figure presents the fraction of advisors who recommended option A in each of the three treatments respectively. The error bars represent +/- 1 S.E.

Table 2 confirms the results using a probit regression analysis. Column (1) shows that there is no statistically significant difference between A recommendations in Before and After. Further, the magnitude of the marginal effect is very small, 0.008, in line with the difference in frequencies observed in Figure 2. The rate of A

recommendations in the Control treatment is significantly higher than in the After treatment (p=.022), as is the difference between the coefficients for the Before and Control treatments (p<.01).

Examining the role of gender, we find that there are no significant differences in A recommendations between female and male participants, as shown in column (2). This is in line with the strict dominance of B, which does not yield a risk-return tradeoff that could lead to different recommendations depending on the individual's degree of risk aversion, the explanation for the gender difference observed in the Distorted Advice Experiment. In columns (3) and (4) of Table 2 we examine the effects of the Before and Control treatment by gender. We do not find significant gender differences in the effect of the Before treatment (p=.826), or the Control treatment (p=.897).

Table 2.2 Treatment Effects on the Likelihood that A is Recommended in the Strict Dominance Experiment

	(1)	(2)	(3)	(4)
		P(A is recommended)		
	All	All	Male	Female
Before Treatment	.008	.009	.024	-.003
	(.061)	(.061)	(.089)	(.085)
Control Treatment	-.169**	-.168**	-.179*	-.159
	(.074)	(.074)	(.108)	(.100)
Female		-.016		
		(.053)		
Share recommending A in After Treatment	.304	.304	.309	.298
Observations	334	334	154	180

Notes: Columns (1) to (4) report marginal effects from probit regressions on the likelihood that A is recommended. In columns (1) and (2), all advisors are included, whereas column (3) reports results only for male advisors and column (4) only for female advisors.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

The results of this experiment provide further evidence in support of the presence of motivated self-deception when judgment is subjective. Removing the

scope for motivated self-deception by introducing a strict dominance relationship between the items to be judged removes any difference in recommendations when information about incentives is delayed. This suggests that our original treatment effect was not due to the avoidance of evaluation or a preference for consistency. This result is in line with Falk and Zimmermann (2011), who show that individuals exhibit preferences for consistency only when they formulate their first judgment in writing, not when they do so in their mind, as in our experiments.

2.3.3 The Persistence of Motivated Self-Deception: Weakening Dominance

The results thus far reveal that evidence for motivated self-deception is found when evaluation is performed on multiple dimensions and no option strictly dominates others in all dimensions. Yet, if *strict* dominance is introduced, no evidence of motivated self-deception is found. In this section we repeat the experiment using only weak, instead of strict dominance. In the new experiment we again changed the payoffs associated with investment B. In this Weak Dominance Experiment, B was a 50-50 lottery between $2 and $6. A remained a 50-50 lottery between $2 and $4. Hence, B weakly dominated A. There are two competing hypotheses. On the one hand, weak dominance could limit the scope for motivated self-deception in the same way as strict dominance does, since investment B weakly dominates investment A. On the other hand, previous findings suggest that even a minor reason to favor the incentivized option could be used by individuals to convince themselves that recommending that option is ethical (Kunda, 1990, see also, Konow, 2000). Hence, if

the advisor focused on the "bad" state, or if she compared the difference between her payoff and the expected payoff of the recipient, which was $1, she could find reasons to recommend A.

We ran the Weak Dominance Experiment following the same procedures as in the experiments above. There were 300 advisors (100 in the Control treatment, 101 in the Before treatment and 99 in the After treatment).

The results of the Weak Dominance Experiment reveal that, when B only weakly dominates A, we still find a significant difference between the Before and After treatments. In the Before treatment, advisors recommended A in 53.5% of the cases, while in the After treatment, they recommended A in 25.3% of the cases ($Z=4.081$, $p<.01$). In the Control treatment, participants recommended A in 14% of the cases. This frequency was significantly lower than that in the Before treatment ($Z=5.913$, $p<.01$) and than in the After treatment ($Z=1.999$, $p=0.046$).

The results suggest that motivated self-deception can be persistent, as long as there is weak dominance on some dimension upon which several items are evaluated. However, when strict dominance is introduced, the bias introduced by incentives through motivated self-deception vanishes entirely.

2.4 Conclusion

Understanding why people behave unethically can help structuring policies to reduce such behavior. For example, many physicians believe incentives such as receiving fees for each procedure they perform or gifts from pharmaceutical companies do not influence their judgment. This belief allows them to receive the incentives while maintaining their self-image as unbiased physicians. The evidence suggests the physicians are wrong, and incentives do distort their judgment in many cases (Steinman et al., 2001; Cain, Loewenstein, and Moore, 2005 and 2011; Malmendier and Schmidt, 2012). This biased judgment comes at a cost to the patients who may not receive the best available treatment and/or may pay more for it.

Examples in which ethical choices are biased by incentives are plentiful and have a huge impact on efficiency and fairness. How can policy makers change this practice? One clear way is to outlaw such incentives when possible, and enforce these laws. But in some cases, changing the law (e.g., due to lobbyists) or monitoring behavior (e.g., because judgment could be subjective) can be hard. Even when this type of solution is feasible, enforcing it could be very costly.

In this paper, we propose an additional approach to reducing the effectiveness of incentives in distorting judgment. By having the decision maker first evaluate the options and only then receive information about the incentives, we made them face their biased choices, changing the behavior of a significant fraction of our participants. We argue that this reduction in unethical behavior results from the psychological cost

to the self-image: when faced with the bias, the decision maker cannot easily convince him/herself that the choices are ethical.

Our message is clear. Some people have psychological costs associated with distorting judgment. Creating procedures that reinforce the role of these psychological costs can reduce unethical behavior by ethical-but-biased individuals. For instance, going back to the doctors example, one solution to prevent overtreatment could be to inform doctors about incentives, e.g., details of their patients' insurance, only *after* they have a chance to evaluate which types of medical procedures are needed. Although timing the information about incentives this way will not help reduce unnecessary care by doctors who recommend procedures knowingly to earn more money, it might reduce this unethical behavior by people who consider themselves ethical.

Another example in which ethical decisions may be biased is the recent discussion in academia around the failure to replicate many published findings. Even though instances of data fabrication are part of the problem, another reason for this crisis is researchers who engage in questionable research practices that increase the chance of false positives (e.g. Simmons, Nelson, and Simonsohn, 2011; Gelman, 2013). Examples are making predictions or choosing which analysis to perform only after looking at the data. Such degrees of freedom in the research practices may allow researchers to get the significant results needed to publish their papers but at the same time feel as if they did not break any ethical rule, preserving their self-image.

Chapter 2, in part, is currently being prepared for submission for publication of

the material. Uri Gneezy, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen, “Motivated Self-Deception, Identity, and Unethical Behavior.” The dissertation author was the co-primary investigator and author of this paper.

Appendix A: An additional Experiment

A.1. The Game

The distorted choice game (Gneezy, Saccardo and van Veldhuizen, 2013) involves three players: two workers and a referee. The workers compete against each other in a real-effort task. The third player, the referee, is asked to judge the tasks and select the winner, who gets a prize of p. Each worker i is allowed to send an amount of money ($m_i \in [0, \frac{1}{2}p]$) to the referee, with only integer amounts allowed in the experiment. The referee can only keep the money of the worker who wins the prize.

The referees' payoff-maximizing strategy in this game is to choose as the winner the worker who sends the highest amount of money. Instead, if referees have moral costs associated with lying about who was the best performer (e.g., Gneezy, 2005), they will prefer to award the prize to the best performer of the real-effort task and will be willing to forgo some monetary benefit by doing so.[1]

However, even ethical referees may bias their judgment of the real-effort task to favor the worker who sent the highest amount of money. Such motivated self-deception could occur, for instance, if referees are able to convince themselves that the worker who sent the highest amount also performed better on the real-effort task, even if she in fact performed worse. If motivated self-deception is successful, referees can thus avoid the self-image cost associated with choosing the worst performance.

To investigate how motivated self-deception affects the referee's judgment, we ran three treatments that share the same payoff structure but differ in their scope for motivated self-deception. In our main Before and After treatments, we asked referees to evaluate workers' performance on a subjective real-effort task that consisted of writing a joke about a pre-specified topic. Though some jokes are clearly better than others, humor is at least partially a matter of taste. As a result, we expected motivated self-deception to be relatively easy in this task.

As in the Distorted Advice Experiment, our main manipulation contrasts two timelines of decision-making. In the Before treatment, the referee received the jokes and the money sent by the workers simultaneously and was then asked to select the winner. Therefore, referees in this treatment had a chance to see the money sent *before* making their judgment about the quality of jokes. As a result, we expect referees to be able to engage in motivated self-deception, convincing themselves that the joke that corresponds to the highest amount sent is also the best joke. Thus, we predict that the

---

[1] Gneezy, Saccardo, and Van Veldhuizen (2013) use this game to study the relative importance of greed and reciprocity in accepting bribes. Their key comparison is between a treatment in which referees can keep only the money sent by the winner and a treatment in which they keep the money from both workers. The main finding is that sending money is significantly more effective in the former than in the latter.

choices of the referee in this treatment will favor the worker who sent the highest amount of money, regardless of the quality of the jokes.

In the After treatment, the referee received the money sent by the workers two minutes *after* receiving the jokes. For the first two minutes, the referee had a chance to evaluate the joke without being influenced by the incentives. Hence, the referee could form an unbiased judgment of the jokes before she received the incentives, and convincing herself that the worker who sent the higher amount of money was also the one with the best joke would have become more difficult. Choosing in favor of the worker who sent the highest amount of money is therefore likely to generate higher self-image costs than doing so in the Before treatment. Thus, we predict that incentives will play a smaller role and the quality of the jokes will play a larger role in this treatment.

We also ran a third treatment, "Objective," as an alternative test of our hypothesis. In this treatment, referees had to judge workers' performance on an objective real-effort task. In particular, workers were asked to identify the colors of a sequence of words (Stroop, 1935). As in the Before treatment, referees in this treatment received the task output and the money sent by the workers simultaneously. However, because workers' performance was objective, engaging in motivated self-deception and appearing ethical to oneself is harder. Therefore, we predict that referees will select the worker with the best performance more often than in the Before treatment.

A.2. Procedures

We conducted the experiment at the University of California San Diego with 273 total participants, 6 in each session.[2,3] Among the participants, 56% were female and the average age was approximately 21.

Upon arrival, we randomly assigned participants to computer terminals and provided them the instructions on computer screens. Participants were anonymously matched in groups of three and were assigned to the role of worker or referee. Each referee was then seated in a separate room and received a $5 show-up fee. Each worker received a $10 show-up fee in $1 bills.

In the Before and After treatments, participants had 10 minutes to type a joke. The topic of the joke was "Economists," and it was communicated immediately before the beginning of the task. After they typed their jokes, workers were asked to report

---

[2] The data of 123 participants (60 in Before, 63 in Objective) are also reported in Gneezy, Saccardo, and Van Veldhuizen (2013). Because we wanted to have 90 observations per treatment for this paper, we also collected 30 additional observations for these treatments as well as 90 new observations for the After treatment. Results remain essentially unchanged if we consider only the first 60 observations in each treatment (results available from the authors).

[3] In one group in treatment Objective, the referee did not follow the instructions and rejected both amounts sent even though this was not part of the instructions. The experimenter only realized this at the end of the session. We decided to discard this observation. To reach the sample size we had originally planned, we ran an additional session.

how confident they were that their joke was better than their competitor's. Each joke was then printed on a sheet of paper. Afterwards, workers were informed via a second set of on-screen instructions that they had an opportunity to send up to $5 of their show-up fee to the referee.

In the Before treatment, workers were informed they could put the money for the referee in a single envelope (labeled with their participant ID) together with the printed copy of their joke. In the After treatment, workers were asked to put the money and the jokes into two separate envelopes. In both cases, workers were also informed that the referee would keep their money only if they won the prize and that it would be returned to them otherwise.

After recording the monetary content of each envelope in private, the experimenter delivered the envelopes with jokes to the referees. In the Before treatment, the envelopes also contained the money sent by workers; in the After treatment, the referee received the envelopes with the money two minutes after the envelopes with the joke.

In the Objective treatment, participants had five minutes to identify the color of as many words as possible using the computer keyboard. We showed participants a sequence of color words on screen (e.g., blue, red, yellow) one after the other and asked them to identify the printed color of each word as quickly as possible. We used a congruent version of the task, meaning that the color word and its printed color were compatible (e.g., blue was always written in blue letters). The number of correctly identified words determined the worker's score for the task. The worker's final score was printed on a score sheet using a scatter plot, where a dot on a random coordinate in the plot represented each correctly identified word. The workers' instructions regarding the money were the same as in the Before treatment. In particular, workers had to put the printed scatter plot and the money in one envelope that would be delivered to the referee.

In all three treatments, the instructions informed the referees that they could only keep the winners' money and had to return the losers' money by putting it back into the loser's envelope. The referees had five minutes to determine the winner, after which all envelopes were returned to the experimenter who then recorded their decisions. In the Before and After treatments, we also asked referees to rate the quality of each joke on a scale from 0 to 10; these ratings were collected at the end of the experiment.

The experiment consisted of two rounds with the same matching of participants. To prevent referees from letting the highest amount of money sent win in round 1 for strategic reasons, no feedback was provided between rounds. Workers started the second round while the referees were evaluating their first round. The procedure for round 2 was identical to that of round 1, apart from the topic of the joke ("Psychologists").

We subsequently recruited additional participants as independent raters. These participants had not previously participated in the experiment and were asked to rate the jokes in exchange for class credit. Each rater was presented with up to six randomly selected pairs of jokes that had "competed" in the experiment, and was asked to rate their quality on a scale of 0 to 10 and determine which was the best joke.

Between 18 and 28 different raters rated each joke. This gives us an unbiased measure of quality for the Before and After treatments.

A.3. Results

In this section, we focus on the analysis of referee behavior below, using one referee as one independent observation. No significant differences in worker behavior exist across treatments, allowing us to focus on referees. In particular, there is no significant difference in the average amount of money sent across treatments (Mann-Whitney, $p>.15$) or in the distributions of amount sent (Kolmogorov-Smirnov, $p>.45$) both when we look at one of the rounds individually or when we combine them. Furthermore, the quality of the jokes was similar in the Before and After treatments (Mann-Whitney $p>.55$; Kolmogorov-Smirnov, $p>.75$).

We use both parametric and non-parametric tests to investigate differences between treatments. For non-parametric tests involving data from both rounds, we take the average over both rounds as the unit of observation.

Joke Quality. For the non-parametric tests discussed below, we examine whether the joke with the highest quality won. For this purpose, we do not include all joke pairs because in some cases, the jokes were simply too close in quality to be reliably distinguishable. Hence we only consider two jokes within a joke pair to be sufficiently different from each other if the fraction of independent raters choosing one joke over the other as winner is different from chance at the 10% level in a test of proportions. For our minimum number of raters per pair (18), this implies taking only those pairs in which at least 65.1% of independent raters picked one of the jokes as the winner (test of proportions, $Z=1.281$, $p=0.100$).[4,5] By this criterion, 66% of pairs over the two joke treatments combined are sufficiently different from each other. Furthermore, to facilitate direct comparisons across treatments, we also use a threshold value for the Objective treatment to exclude the performance levels that were very similar. We picked the threshold value to be 11 points, because this value includes the same fraction of data points included in the subjective treatments.

In the regression analysis that follows after the non-parametric tests, we do not use thresholds and incorporate all observations, including those in which quality was similar across the two workers.

Referee Choices. Figure A.1 displays the fraction of referees choosing the worker who sent the highest amount of money (Amount Sent, left section of Figure

---

[4] For a threshold of 69.4%, which corresponds to jokes being significantly different at the 5% level, the results are similar. To keep the largest number of observations, we chose to focus on the threshold of 65.1% instead.

[5] The agreement of raters is also reflected in the difference in our measure of quality, the average rating provided by the independent raters. The average difference in ratings within pairs of jokes that exceed the 65.1% threshold (1.63) is significantly larger than the average difference in quality in jokes below the threshold (0.75) (Mann-Whitney test, $p<0.001$).

A.1) and the fraction of referees choosing the worker with the highest quality (right section) as winner across the three treatments.

In the Before treatment, in which the incentives and the joke were received simultaneously, 84% of the workers who sent the highest amount of money won the prize, which is significantly greater than chance (Wilcoxon signed-rank (WSR) test, p=.001). By contrast, only 56% of the best jokes won the prize, a fraction that is not significantly different from chance (WSR test, p=0.491). Thus, incentives appear to distort judgment in this treatment.

In the After treatment, the percentage of workers with the highest amount sent who win decreases to 73%, which is still significantly larger than chance (Mann-Whitney (MW) test, p=.003) and not statistically different from the Before treatment (MW test, p=.369). However, the percentage of workers with the best joke who won in this treatment is 81%, which is significantly higher than what we observed in the Before treatment (MW test, p=.027). Further, it is significantly different from chance (WSR test, p<.001). This finding is consistent with our hypothesis that making motivated self-deception more difficult increases the importance of quality. The treatment difference in the importance of quality is strong and economically significant. The best joke winning 81% of the time is equivalent to 62% of referees going for quality; the corresponding percentage for the Before treatment is 12%.

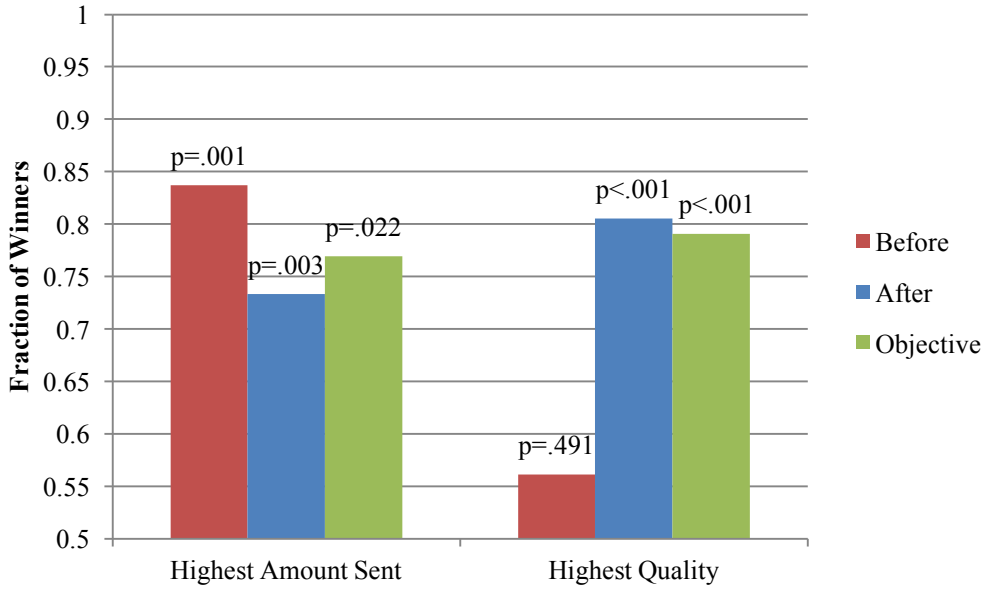

Figure 2.A1 Fraction of Winners Conditional on Highest Amount Sent or Quality
Notes: The p-values are calculated using a Wilcoxon signed rank test that tests if the reported fraction is significantly larger than .5. Workers are classified as having a better rating when at least 65.1% of independent raters agree their joke is better (treatments Before and After) or when their performance on the Stroop task is at least 11 words better (treatment Objective).

In the Objective treatment, 77% of workers who sent the highest amount of money won the prize (MW test, p=.021), which is not significantly different from either the Before (MW test, p=.730) or After (MW test, p=.572) treatments. Further, 79% of the workers with the best score on the task won (MW test, p<.001). This percentage is larger than the one observed in the Before treatment (56%, MW test, p=.035) but similar to the percentage observed in the After treatment (81%, MW test, p=.790). Thus, making self-deception more difficult by using an objective task also increases the importance of quality in determining the winner.

To investigate the effect of incentives and quality simultaneously, we also report the results of probit regression analyses in Table A.1. To facilitate comparisons between coefficients and treatments, we report marginal effects and have standardized all independent variables, so that the coefficients represent the effect of a one-standard-deviation increase in the independent variable. We also allow the importance of quality to differ depending on whether or not the referee received two identical amounts of money. Intuitively, when the amounts sent are identical, referees no longer have a monetary incentive to distort the outcome and can therefore be expected to be more interested in quality.[6]

Column 1 shows that in the Before treatment, relative to a situation with equal amounts of money, increasing the difference in the amount sent by one standard deviation increases the likelihood of winning by 42 percentage points. By contrast, having the best joke does not increase the likelihood of winning when referees receive different amounts of money from the workers. However, the quality of the joke does matter when referees receive two identical amounts. This result shows that when referees no longer have an incentive to distort the outcome, they choose the better joke as the winner, whereas when incentives are in place, their judgment is biased.

Column 2 shows that the observed pattern is different in the After treatment. In contrast to the Before treatment, the quality of the joke matters even when the two amounts sent are different. Conversely, incentives matter less than in the Before treatment. Column 3 reports the results for the Objective treatment. In this treatment, when the amounts sent differ, both quality and incentives matter, with the (normalized) marginal effect for quality being somewhat larger than the coefficient for incentives. Quality also matters when the amounts of money are identical. Additional analyses are provided in section A.4 where we provide several robustness checks.

We also examine referee behavior distinguishing between cases in which referees received two positive amounts of money and one positive amount, respectively. Intuitively, justifying letting the worst performer win when both workers send money might be easier, and as a result, self-image costs might be higher when only one worker sends money. The analysis provides some support for this conjecture, as shown in section A.4.

---

[6] Because the two workers in each pair are the exact inverse observation of one another and therefore not independent observations, we randomly select one worker per pair to include in the analysis. In section A.4 below, we redo the analysis with 1,000 random samples to show that the results reported here are not due to the particular random sample that was selected.

Table 2.A1 Probit Regressions for Referees

Probability (winning)	(1)	(2)	(3)
Quality Difference (amounts sent differ)	-.001	.148**	.369***
	(.103)	(.066)	(.110)
Quality Difference (amounts sent identical)	.401***	.586***	.229**
	(.144)	(.186)	(.116)
Amount Sent Difference	.422***	.206**	.197**
	(.102)	(.081)	(.091)
Treatment	Before	After	Objective
Standard Errors	Clustered	Clustered	Clustered
Observations	60	60	62
Clusters	30	30	31

Notes: Probit estimates (marginal effects). Quality Difference is the difference between the quality of the joke (i.e., the average score among independent raters) of the selected worker and the other worker in the group. Amount Sent Difference is the difference between the amount of money sent by the selected worker and the amount sent by the other worker in the group. In each specification, we randomly select one worker per referee per round. Robust standard errors are clustered at the referee level.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Quality Ratings. As mentioned above, in addition to asking referees to determine the winning worker, we also asked referees in the Before and After treatments to rate the quality of both jokes on a scale from 0 to 10. This measure was not incentivized. Interestingly, the correlation between the referees' ratings and the grades given by independent raters is 0.27 for the Before treatment and 0.54 for the After treatment. An OLS regression with ratings from the referees as a dependent variable and ratings from the independent raters as an independent variable shows this correlation is much stronger for the After treatment (β=1.19, p<.001) than for the Before treatment (β=.46, p=.019); including an interaction term between treatment and independent ratings shows that the difference in coefficients is significant (β=.73, p=.006). Thus, referees in the After treatment gave a less biased judgment of joke quality than referees in the Before treatment. This finding is in line with self-deception being harder in the After treatment as well.

Taken together, these results are in line with motivated self-deception. As in the Distorted Advice Experiment, when the task is subjective, quality plays a larger role in determining a winner when referees perform their judgments before being aware of the incentives. When incentives are provided at the same time as jokes, referees' judgment shifts toward workers who sent the highest amount of money. Conditional on amounts sent being different, quality no longer plays a role.

Additionally, when the task is more objective, receiving the incentives together with the task does not lead to the same bias.

A.4. Robustness checks

We investigate differences in the effect of quality of jokes and the effect of receiving money on referees' choices across treatments, using OLS regressions and interacting these variables with treatment dummies. In this analysis, we use OLS rather than probit to facilitate treatment comparisons. Table A.2 reports the results. Column 1 suggests that the difference in amount sent is a less important determinant of referees' choices in the After treatment than in the Before treatment (p=.11).[7] Conversely, quality difference between jokes plays a larger role in the After treatment than in the Before treatment (p=.072). Column 2 shows that a similar pattern emerges when comparing the Before treatment with the Objective treatment: the difference in amount sent by the two workers is more important in the Before treatment (p=.047), whereas quality plays a larger role in the Objective treatment. Finally, column 3 shows that amount sent and quality have similar effects in the Objective and After treatments.

[7] The significance of this coefficient varies depending on the random sample drawn. In 575 out of 1,000 random samples, the coefficient is significant at the 10% level or lower. In the draw randomly selected for Table A.2. the coefficient is not significant. All other interaction terms are robust and remain significant in at least 900 random samples out of 1,000.

Table 2.A2 OLS Interaction Terms for Referees

Probability (winning)	(1)	(2)	(3)
Quality Difference (amounts sent differ)	-.019	-.019	.219***
	(.058)	(.058)	(.060)
Quality Difference (amounts sent identical)	.244***	.244***	.222**
	(.048)	(.048)	(.090)
Amount Sent Difference	.298***	.298***	.136*
	(.039)	(.039)	(.069)
Quality Diff. (amounts sent differ) * After	.152*		-.084
	(.079)		(.086)
Quality Diff. (amounts sent identical)*	.155*		.174*
	(.085)		(.110)
Amount Sent Difference * After	-.122		.040
	(.076)		(.095)
Quality Diff. (amounts sent differ) *		.239***	
		(.083)	
Quality Diff. (amounts sent identical) *		-.022	
		(.101)	
Amount Sent Difference * Objective		-.162**	
		(.080)	
Treatment	Before &	Before &	Objective &
Standard Errors	Clustered	Clustered	Clustered
Observations	120	122	122
Clusters	60	61	61

Notes: OLS estimates. Quality Difference is the difference between the quality of the joke (i.e., the average score among independent raters) of the selected worker and the other worker in the group. We standardize this variable to have the same mean and standard deviation in the objective task as in the joke task. Amount Sent Difference is the difference between the amount of money the selected worker sent and the one sent by the other worker in the group. In each specification, we randomly select one worker per referee per round. The regressions also include treatment dummies to correct for difference in the overall fraction of winners as the result of randomly selecting workers; their coefficients are always small and not significant. Robust standard errors are clustered at the referee level.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

For the analyses reported in Table 2, we randomly selected one worker for each pair. To ensure that our results are not due to the specific random sample selected for the analysis, we additionally redo the regressions reported in Table A.1 with 1,000 different random samples and report the average results as well as the standard deviation in the estimated marginal effects (Table A.3). The results are very similar to those reported in Table A.1, which reveals that our results are robust to the particular random sample we used.

Table 2.A3 Probit Regressions for Alternative Random Samples

Probability (winning)	(1)	(2)	(3)
Quality Difference (amounts sent	.00	.15	.36
	[.01]	[.01]	[.02]
Quality Difference (amounts sent	.41	.56	.23
	[.03]	[.04]	[.02]
Amount Sent Difference	.43	.20	.20
	[.02]	[.02]	[.01]
Treatment	Before	After	Objective
Observations	60	60	62
Clusters	30	30	31

Notes: Probit estimates (marginal effects). Quality Difference is the difference between the quality of the joke (i.e., the average score among independent raters) of the selected worker and the joke of the other worker in the group. Amount Sent Difference is the difference between the amount of money sent by the selected worker and the one sent by the other worker in the group proposed. In each specification, we re-estimate the regression reported in Table 2 1,000 times with different random samples of one worker per referee per round; the marginal effect is the average of the 1,000 marginal effect estimates, and the number in square brackets is the standard deviation of the 1,000 marginal effect estimates.

So far, our analysis has looked only at differences in the monetary amount sent and quality (in regressions) or the effect of one amount/performance being greater than the other (in non-parametric tests). However, referees might respond differently to quality when they receive two strictly positive monetary amounts compared to situations in which only one person sends a positive amount. For example, a referee might be happy to take $5 over $2 when both workers send money, but not $3 over $0, because the dishonesty of this act is more salient. Although justifying taking the highest amount sent is relatively easy when both workers send a positive amount, because they are both being dishonest, justifying taking the highest amount sent when one of the participants behaves honestly may be harder.[8]

In Table A.4, we estimate separate coefficients for both the amount sent difference and quality difference for the cases when two workers or only one worker sent a positive amount, respectively. In the Before treatment, differences in amount sent are always important and quality only matters when the amounts sent by two workers are identical. In the After treatment, a shift occurs in the relative importance of quality and amount sent. It is especially strong for the cases in which only one worker sent a positive amount of money. In such cases, quality matters and the amount sent to the referee does not. By contrast, when both workers send positive amounts, the effect of quality is not statistically significant and the effect of amount sent is.

Figure 2.A2 illustrates this result graphically. When the referee receives a positive amount from one worker only, the better-quality joke wins 63% of the time in

[8] We did not originally intend to incorporate this analysis in our paper; we only included it after it had been repeatedly suggested to us by seminar participants and others.

the After treatment (WSR test, p=.248), compared to 93% in the Before treatment (WSR test, p=.004). This difference is statistically significant (MW test, p=.049). By contrast, when referees receive positive amounts by both workers, the better joke does not win significantly more (or less) often than chance in either treatment.

For the Objective treatment, the effect of amount sent is similar to the After treatment: it matters only when two workers send positive amounts. However, in this case, the importance of quality does not seem to depend on whether one worker or two workers sent a positive amount.

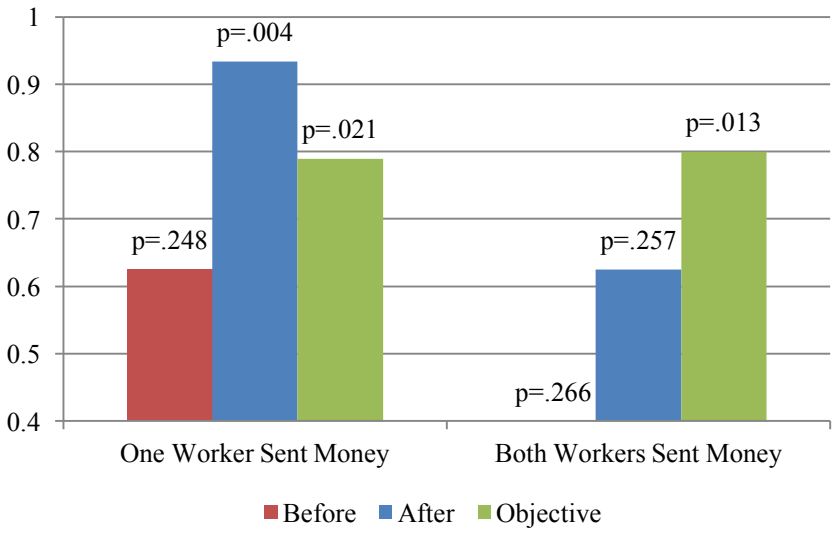

Figure 2.A2 Fraction of Winners conditional on one worker vs. both workers sending money
Notes: The p-values are calculated using a Wilcoxon signed rank test that tests if the reported fraction is significantly larger than .5. Workers are classified as having a better rating when at least 65.1% of independent raters agree their joke is better (Before and After treatments) or when their performance on the Stroop task is at least 11 words better (treatment Objective).

Table 2.A4 Probit Regressions One versus Two Positive Amounts Sent

Probability (winning)	(1)	(2)	(3)
Quality Difference (amounts sent differ)	.059	.392**	.343**
(one worker sent a positive amount)	(.169)	(.188)	(.125)
Quality Difference (amounts sent differ)	-.008	.085	.636**
(both workers sent positive amounts)	(.112)	(.113)	(.177)
Quality Difference (amounts sent	.402***	.584***	.229**
	(.143)	(.187)	(.116)
Amount Sent Difference	.527***	.140	.177
(one worker sent a positive amount)	(.156)	(.113)	(.124)
Amount Sent Difference	.349***	.465***	.513**
(both workers sent a positive	(.136)	(.156)	(.226)
Treatment	Before	After	Objecti
Selected Workers	Random	Random	Rando
Standard Errors	Clustere	Clustered	Cluster
Observations	60	60	62
Clusters	30	30	31

Notes: Probit estimates (marginal effects). Quality Difference is the difference between the average quality of the joke (as judged by the independent raters) of the selected worker and that of the other worker in the group. Amount Sent Difference is the difference between the amount of money sent by the selected worker and the amount of money sent by the other worker in the group. All marginal effects are evaluated at the means for all independent variables. Robust standard errors are clustered at the referee level.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

Appendix B: Instructions

In this section, we present the instructions for the experiments. We first present the instructions for Experiment 1, separately for advisors and clients.

B.1 Advisor – Client Game

The instructions for advisors are presented in Subsection B.1.1. Treatment variations are added in squared brackets, indicating the treatment they apply to. The instructions for clients are presented in Subsection B.1.2.

B.1.1 Instructions for Advisors

Your role

In this study, you have been assigned the role of advisor. You will be asked to recommend one of two products to another participant, the client, who will participate in another session of this study.

How it works

- As advisor you are informed about the details of two products, product A and B
- You recommend one of the products, A or B, to the client
- The client, who does not have any information about A and B, receives your recommendation
- The client chooses product A or B. The client's choice and chance will determine his/her payoffs.

Pairing between advisors and clients

At the end of this study, we will randomly select one advisor out of ten and give his or her recommendation to a client, who will be paid accordingly. As mentioned above, the client will be a participant in another session of this study.

Payments

You will receive $1 for providing your recommendation.

[Before**: If you recommend product A, you will receive an additional $1 as commission**.]

---next page---

Information about the two products

Product A

- Product A is produced by Company A and has two potential qualities, low or high
- If the quality of product A is low, it will yield $2 to the client.
- However if the quality of product A is high, it will yield $4.
- There is an equal chance that the quality will be either low or high.
- Whether the quality of the product is actually high or low will be determined by a coin toss later.

Product B

- Product B is produced by Company B
- As product A, it has two potential qualities, low or high.
- If the quality of product B is low, it will yield $1 to the client.
- However if the quality of product B is high, it will yield $7.
- There is an equal chance that the quality will be either low or high.
- Whether the quality of the product is actually high or low will be determined by a coin toss later.

Hence, after the client chooses between product A or B, we will use the outcome of a coin toss, made by a volunteer, to determine his/her payoff.

Please consider your recommendation
In a few minutes you will be given a decision sheet and you will be asked to complete the sentence:

"I recommend you to choose product (A or B) _____"

This decision sheet will be shown to the client before he or she chooses between product A or B. You will be asked to put it in an envelope and the envelope will be delivered to the client.

Please take a minute to decide which product to recommend. Click the arrow below when you are ready to provide your recommendation.

---next page---

Please raise your hand now

The experimenter will give you your decision sheet where you can write your recommendation to the client.

Once you have your decision sheet, click below to proceed to the next screen.

---next page---

Decision sheet

Please write down your recommendation on the decision sheet

[Before and After: **If you recommend product A, you will receive an additional $1 as a commission.**]

Your recommendation

To make sure all records are kept, please input your recommendation on this screen as well.

"I recommend you to choose product ___

B.1. 2. Instructions for Clients

Your role

Welcome to this study on decision-making. In this experiment you are matched with another participant. Neither your identity nor the identity of the participant you are matched with will be revealed.

In this study, you have been assigned the role of the client. Your task will be to choose one of two products, which will result in some monetary payments to you. The monetary payment you will receive depends on the product you choose.

In a previous session of the study another participant, the advisor, was provided with information about the two products and was asked to recommend a product to you.

In a moment, you will receive the recommendation from the advisor. Please raise your hand.

---------- After receiving the recommendation ----------

Which product do you choose?

Product A
Product B

References

Akerlof, George A., and Rachel E. Kranton. 2000. “Economics and identity.” *Quarterly Journal of Economics* 115 (3): 715-753.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. “Doing good or doing well? Image motivation and monetary incentives in behaving prosocially.” *American Economic Review* 99 (1): 544-55.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased judgments of fairness in bargaining." *American Economic Review* 85 (5): 1337-1343.

Babcock, Linda, and George Loewenstein. 1997. “Explaining bargaining impasse: The role of self-serving biases.” *Journal of Economic Perspectives* 11 (1): 109-26.

Banerjee, Abhijit V. 1997. “A theory of misgovernance.” *Quarterly Journal of Economics* 112 (4): 1289–1332.

Bem, Daryl J. 1972. “Self-perception theory.” *Advances in Experimental Social Psychology* 6: 1-62.

Bénabou, Roland, and Jean Tirole. 2006. “Incentives and prosocial behavior.” *American Economic Review* 96 (5): 1652-78.

Cain, Daylian M., George Loewenstein, and Don A. Moore. 2005. “The dirt on coming clean: perverse effects of disclosing conflicts of interest.” *Journal of Legal Studies* 34 (1): 1-25.

Cain, Daylian M., George Loewenstein, and Don A. Moore. 2011. “When sunlight fails to disinfect: understanding the perverse effects of disclosing conflicts of interest.” *Journal of Consumer Research* 37 (5): 836-57.

Cialdini, Robert. 1984. “Influence, the Psychology of Persuasion.” New York: Harper Collins.

Clemens, Jeffrey, and Joshua D. Gottlieb. 2014. "Do physicians' financial incentives affect medical treatment and patient health?" *American Economic Review* 104 (5): 1320-1349.

Crawford, Vincent P., and Joel Sobel. 1982. “Strategic information transmission.” *Econometrica* 50 (6): 1431-51.

Emanuel, Ezekiel J., and Victor R. Fuchs. 2008. "The perfect storm of overutilization." *JAMA: The Journal of the American Medical Association* 299 (23): 2789-2791.

Falk, Armin, and Florian Zimmermann. 2011. "Preferences for consistency." *IZA Working Paper* 5840.

Freud, Sigmund., 1933. *New Introductory Lectures on Psycho-Analysis*. W.W. Norton & Company. The Standard Edition edition (1990).

Gelman, Andrew. 2013. "Preregistration of studies and mock reports." *Political Analysis* 21: 40-41.

Gneezy, Ayelet, Uri Gneezy, Gerhard Riener, and Leif D. Nelson. 2012. "Pay-what-you-want, identity, and self-signaling in markets." *Proceedings of the National Academy of Sciences* 109 (19): 7236-40.

Gneezy, Uri. 2005. "Deception: the role of consequences." *American Economic Review* 95 (1): 384–394.

Gneezy, Uri, Silvia Saccardo, and Roel van Veldhuizen. 2013. "Bribery: greed versus reciprocity." *Working paper*.

Gruber, Jonathan, John Kim, and Dina Mayzlin. 1999. "Physician fees and procedure intensity: the case of cesarean delivery." *Journal of Health Economics* 18 (4), 473-490. IOM (Institute of Medicine). 2012. *Best care at lower cost: The path to continuously learning health care in America*. Washington, DC: The National Academies Press.

Johnson, Erin M., and M. Marit Rehavi. "Physicians treating physicians: information and incentives in childbirth." *National Bureau of Economic Research Working Paper* No. 19242.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological Bulletin* 108 (3): 480.

Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. "Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence." *Journal of Personality and Social Psychology* 37 (11): 2098.

Mafi, John N., Ellen P. McCarthy, Roger B. Davis, and Bruce E. Landon. 2013. "Worsening trends in the management and treatment of back pain." *JAMA internal medicine* 173 (17): 1573-1581.

Malmendier, Ulrike, and Klaus Schmidt. 2012. "You owe me." *NBER Working Paper*

No. 18543.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The dishonesty of honest people: A theory of self-concept maintenance. " *Journal of Marketing Research* 45 (6): 633-644.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22 (11): 1359-66.

Steinman, Michael A., Michael G. Shlipak, and Stephen J. McPhee. 2001. "Of principles and pens: attitudes and practices of medicine housestaff toward pharmaceutical industry promotions." *American Journal of Medicine* 110 (7): 551-7.

Svensson, Jakob. 2005. "Eight questions about corruption". *Journal of Economic Perspectives* 19 (3): 19-42.

Stroop, J. R. 1935. "Studies of interference in serial verbal reactions." *Journal of Experimental Psychology* 18 (6): 643–662.

Transparancy International. 2011. "Bribe Payers Index 2011." http://bpi.transparency.org/bpi2011/results/.

Trivers, Robert. 2011. *The folly of fools: the logic of deceit and self-deception in human Life*. Basic Books.

Westen, Drew. 1985. *Self and society: narcissism, collectivism, and the development of morals*. Cambridge University Press.

Westen, D. 1998. "The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science." *Psychological Bulletin*, 124, 333–371.

3. A Must Lie Situation - Avoiding Giving Negative Feedback

Abstract

We examine under what conditions people provide accurate feedback to others. We use feedback regarding attractiveness, a trait people care about, and for which objective information is hard to obtain. Our results show that people avoid giving accurate face-to-face feedback to less attractive individuals, even if lying in this context comes at a monetary cost to both the person who gives the feedback and the receiver. A substantial increase of these costs does not increase the accuracy of feedback. However, when feedback is provided anonymously, the aversion to giving negative feedback is reduced.

3.1 Introduction

Feedback is crucial to learning. The transmission of information from an informed agent to a receiver who might benefit from it is studied theoretically in the standard principal-agent model in economics (Crawford and Sobel 1982; Prendergast, 1993; Levitt and Snyder, 1997; Morris, 2011; Olszewski, 2004; Ottaviania and Sørensen, 2005a, 2005b). The question in this literature is how to design incentives contracts such that the agent with the private information will send an honest signal to the principal. We expand the discussion by studying cases in which the agent might suffer a psychological cost from sending a negative signal, and hence avoids it.

In some cases receivers would get little value out of receiving accurate signals regarding their ranking or performance, because they can do little to change it. Consider an episode of "Seinfeld" when Jerry and Elaine are invited to see their friends' baby. One look and they both agree the baby is "the ugliest baby you have ever seen." They of course do not tell this to the proud parents. Jerry's insight later is "And, you know, the thing is, they're never gonna know, no one's ever gonna tell them…it's a must lie situation."

In the ugly baby case, since parents can do little about the baby's appearance, the feedback might not be that valuable. However, in other cases feedback can help people achieve better outcomes. Consider, for example, a person on the job market who keeps applying to jobs that he might be qualified for on paper, but is considered unsuitable based on less tangible character traits displayed during the interview

process. He might talk, act or dress in a way that displays a low work ethic or just does not fit with the company image. Honest feedback about his personal characteristics could help such a person to revise his application strategy and thus to be more successful in the process – either by applying to companies that are a better fit to his personality or by adapting his behavior to the companies he is applying for. Consequently, a lack of feedback could lead to frustration, extended unemployment spells and superfluous investment into further education.

In order to study the provision of feedback, we designed a novel experiment in which participants are asked to give feedback to others on their level of attractiveness. We decided to use attractiveness as a proxy for similar less-tangible traits that could be subject to feedback for four main reasons. First, whereas for some attributes people have a good knowledge of their relative rank (e.g., height), the feedback regarding own attractiveness is noisy and relies on indirect measures such as success in dating. Hence, receiving accurate feedback could be very informative. Second, attractiveness is an attribute most people care about a lot, and thus receiving an informative negative signal could hurt. Third, attractiveness is correlated with economic success (Solnick and Schweitzer, 1999). Fourth, attractiveness can be judged within seconds in an experimental setting, while other traits might only be revealed after an extensive interaction.

In our experiment we asked groups of men and women to rank the members of the opposite sex by attractiveness. We then incentivized participants to correctly judge the rank of another participant of their own sex in private, and compared these assessments to those provided in a treatment in which the attractiveness judgments

were provided to the assessed individual as face-to-face feedback. We find that participants are reluctant to provide honest negative face-to-face feedback to other people even if lying in this context comes at a cost to both the feedback provider and the receiver. Further, we find that a considerable increase in these costs does not change the accuracy of individuals' feedback provision.

One reason for the avoidance to provide negative face-to-face feedback could be a personal cost. The receiver of the information might decide to "shoot the messenger"—blame the (innocent) carrier of bad news. Psychologists, starting with Freud (1999), studied this phenomenon, arguing that people may blame the messenger for the message as a mechanism to fight feelings of powerlessness and a lack of control.

Alternatively, the avoidance of giving negative feedback could rise from trying to shield the receiver from negative information that could hurt. That is, individuals may experience negative utility from providing signals that they anticipate to be painful to the receiver.

To better understand what drives the reluctance to provide honest negative feedback that we find in our experiment, we ran a treatment in which feedback provision is anonymous, i.e., the identity of the feedback provider is not revealed to the receiver. We find that participants provide more honest feedback when their anonymity is guaranteed as compared to when their identity is revealed. This suggests that the reluctance to give face-to-face feedback to less attractive individuals is driven by unwillingness to be identified as the messenger of the bad news.

The paper is organized as follows. The experimental design is presented in the following section. Our main results are presented in Section III. Section IV concludes.

3.2 Experimental Design

3.2.1 The Setting

The experiment consists of four treatments: Judgment, Face-to-Face (F2F) Feedback, High Stakes and Anonymous Feedback. Participants took part in four stages. The first two stages are procedurally the same in all treatments, while the experimental design differs in stages three and four.

In each session of the experiment we recruited 20 participants, 10 men and 10 women. Upon arrival to the laboratory, men were instructed to line up on one side of the room while women formed a line on the opposite side, such that the two groups faced each other. Then participants received the instructions as well as an ID sticker, which they were asked to wear visibly on their chest. Women received ID letters from FA to FJ; men received ID letters from MA to MJ. After putting on the stickers, participants were asked to start reading the instructions for stage one (all the instructions are reported in Appendix D).

Ranking the opposite sex: In the first stage of the experiment participants were asked to rank the members of the opposite group by attractiveness from 1 (the most attractive person) to 10 (the least attractive person), such that each person in the other group received a different number. The ranks given to each participant in a group were

added up and the resulting sums were ordered from the lowest to the highest value. Based on this order, we created an aggregate attractiveness ranking for the group from 1 to 10, such that each participant received a different rank. In the unlikely case of a tie, ranks were determined by the flip of a coin. To incentivize accuracy in the ranking, we promised participants a monetary reward if their rankings matched the aggregate ranking for at least five people. This reward was specified as $10 in the Judgment, F2F Feedback and the Anonymous Feedback treatment, and as $50 in the High Stakes treatment.

Guessing own rank: After everyone had completed the first task (and before knowing whether their own ranking was in line with the group's ranking), we instructed participants to continue with the second stage. In this stage we asked participants to form a circle with their own sex group, so that all other group members were clearly visible to everyone. We reminded them of the aggregate attractiveness ranking provided by participants of the opposite sex group, and asked participants to guess their own rank in the aggregate ranking. That is, we asked participants to guess how the other group had ranked them. We promised participants $10 if their guess matched their actual position in the aggregate ranking. In the High Stakes treatment, the incentive was again increased to $50 instead of $10.

Without knowing whether their guess of their own rank was correct, participants were then asked to turn to the next page of their instructions to continue with stage three. From that point, the experimental design differed across treatments.

Judgment — In the Judgment treatment we asked participants to guess the rank in the aggregate ranking in stage one of another same sex participant in their group.

They were again told that if they guessed correctly we would pay them $10. We told participants that their guess would not be revealed to the participant whose rank they were guessing.

In stage four, we then asked participants to provide a complete attractiveness ranking of their own group including themselves. The incentive structure was identical to that of stage one.

F2F Feedback — In the F2F Feedback treatment we also asked participants to guess the rank in the aggregate ranking (from stage one) of another person in their group. However, unlike the Judgment treatment, participants had to send a message to their counterpart with their assessment. Each participant gave feedback to one participant and received feedback from a different participant afterwards. Senders knew that the receiver would know the rank assessment they gave and their identity.

In particular, participants were asked to write their guess in a message that was delivered to the receiver by the experimenter (see in Appendix D a sample of the message individuals had to complete). The message stated *"My guess about participant (ID)'s position in the aggregate ranking (1-10): __*", and participants had to enter a number from 1 to 10. The sender's ID was pre-written on the message.

To incentivize participants, they were informed that in stage four, after receiving the message, each participant would have the opportunity to update the guess of their own rank provided in stage two. If a participant guessed his/her rank correctly in stage four he/she would receive $10, and so would the person who provided feedback to him/her.

After sending their messages, all participants received a similar message from another participant of the same sex with feedback on their ranking. The ID of the participant who sent the message was visible on the message sheet. Upon receiving their message, participants were asked to continue to stage four, in which they decided whether to update their personal rank prediction from stage two. Participants were notified that a correct, unaltered guess in both stages two and four would only be rewarded once and that the guess provided in stage four overruled the one provided in stage two. If the participant did not guess correctly in stage four, he/she and the person who sent him/her feedback did not receive additional money in this stage.

High Stakes — In the High Stakes treatment, the procedure was identical to that of the F2F Feedback treatment except that the incentives for each stage of the experiment were increased from $10 to $50.

Anonymous Feedback — In the Anonymous Feedback treatment the procedure in stages three and four was the same as in the F2F Feedback treatment, except that the participants did not know the identity of the feedback provider in stage three. Each participant's guess about the other's attractiveness ranking was included in a message that the experimenter delivered to the receiver. However, the sender's ID was not indicated on the message sheet, so that his/her identity was not revealed.

3.2.2 Procedure

We conducted the experiment with students at the University of California, San Diego. A total of 400 participants (50% female) participated in the experiment, with 100 participants in each of the 4 treatments. Our sample consists of 62% Asians,

14% Caucasians, 10% Hispanics, and 9% indicated a different ethnicity. For 6% of the sample we do not have any ethnicity information. The average age is 20.5 years with a standard deviation of 1.9.

Participants were recruited through the online participant database of the university. In order to ensure a total of 20 participants per session, we recruited 30 participants to the lab (half men and half women). We then selected 10 men and 10 women at random for the experiment. All the remaining participants were dismissed after receiving a $5 show up fee. No participant took part in more than one session. We ran the experiment using pen and paper in the spring and fall of 2013 and winter of 2014. On average, each session lasted around 20 minutes. The average payment for the experiment was $9.92 plus a $5 show-up fee. To guarantee confidentiality to all participants, individuals were communicated their total earnings without being told their earning for any given stage of the experiment. None of the participants succeeded in earning money for all stages of the experiment.

3.3 Results

In the following, we first present the results on participants' self-assessment. We then turn to their evaluation of others' attractiveness and the resulting updating behavior for each treatment separately. Throughout the analyses we will use the aggregate ranking computed using participants' evaluations of all opposite sex participants during stage one as a measure of participants' actual rank. To make sure that the aggregate ranking is a meaningful indicator of their relative attractiveness, we

first verify that participants perceived other peoples' attractiveness similarly. We find that the rankings of participants of the opposite sex are highly correlated among group members in the $10 incentive treatments (Cronbach's Alpha for men=0.85 and for women=0.88, N=298).[1,2] We find similar results in the High Stakes treatment (Cronbach's Alpha for men=0.90 and for women=0.92, N=100). Hence, even though taste might differ across individuals, in our experiment we observe a high degree of agreement regarding the relative attractiveness of others.

3.3.1 Self-Assessment

Feedback is particularly useful in situations where one's own perception and the perception of others differ. A necessary prerequisite for our study is thus a discrepancy between individuals' self-assessment and their rank in the aggregate attractiveness ranking. Since the self-evaluation stage of the experiment (stage two) was identical across the Judgment, F2F-Feedback and Anonymous Feedback treatment, we pool all the observations from these treatments for this analysis. For the High Stakes treatment we present the results separately, as participants faced different incentives for this stage.

In order to detect any possible bias in individuals' self-assessment we compare the distribution of guesses of participants' own rank to the distribution we would

---

[1] The number of observations is 298 instead of 300 as one participant indicated numbers instead of ID letters in this stage and a second participant left out more than one ID number in his ranking. Out of all 400 participants in all four treatments 18 left out one ID letter in their ranking and repeated another one instead. In this case we randomly assigned the missing ID number to one of the rank positions with the repeated ID.

[2] Cronbach's Alpha is a measure of internal consistency. It measures the average correlation of the individual rankings we use to create the aggregate ranking. The higher the score, the more reliable is the generated scale.

expect under full information. If participants were perfectly aware of their position in the ranking, we should observe a discrete uniform distribution of guesses, because each rank can only be awarded once per group. However, as can be seen in Figure 1A, the distribution of the self-assessments when individuals are incentivized with $10 is skewed to the left and significantly different from a uniform distribution ($\chi^2(9)=95.20$, $p<0.001$, N=298). Participants assign a mean rank of 3.85 to themselves, instead of the expected 5.5.

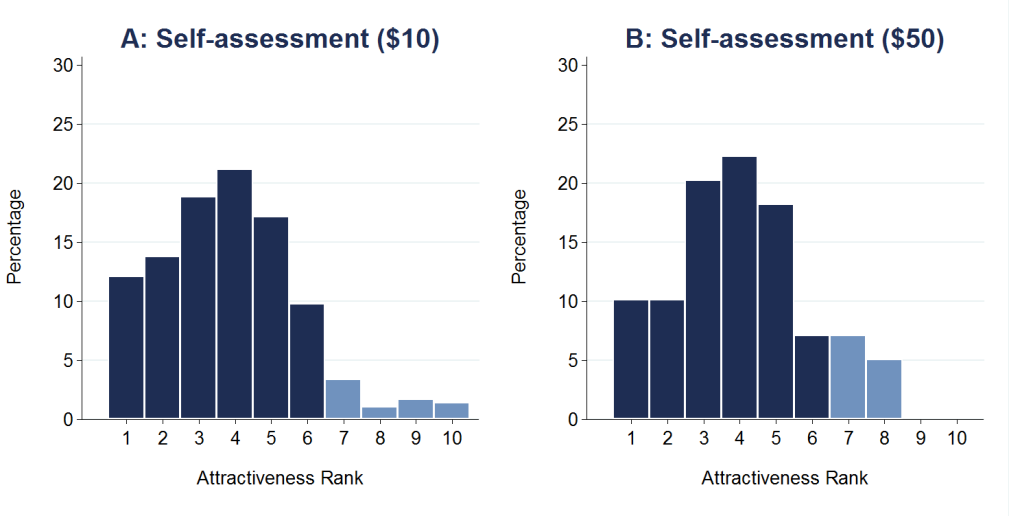

Figure 3.1 Distribution of Participants' Guesses of Own Attractiveness

When investigating the extent at which individuals' perception of their own attractiveness deviates from the aggregate ranking provided by the other participants in stage one (i.e., their actual rank), we find that on average individuals rank themselves as 1.63 ranks more attractive than their actual rank (SD=2.35, N=298). A Wilcoxon signed rank (WS) test confirms that this difference is statistically significant ($z=-9.97$, $p<0.001$). Overall, 63% of the participants guess that they are more

attractive than others perceive them to be. Those individuals on average deviate by 3 ranks from their actual rank (SD=1.73, N=189). Hence, we observe a considerably biased self-assessment in our sample.

This overconfidence is well documented in the psychology and economics literature (see Moore and Healy (2008) for a review, or Balafoutas et al. (2012) and Burks et al. (2013) for applications). We also observe that this overly positive self-evaluation is true for both sexes: there are no differences in the self-assessment distributions of men and women ($\chi^2(9)=12.33$, p=0.195) or in the average deviation of their self-assessment from their actual rank (Mann Whitney (MW) test, z=1.19, p=0.235).

As can be seen in the distribution of self-assessments depicted in Figure 1A, there is a considerable drop in the fraction of participants who rank themselves as 7 or higher. While ranks 1 to 6 are guessed by a fraction of people that is equal or larger than the expected 10% per rank, the fraction of participants who indicate any of the ranks above 6 is significantly lower than the expected 10% per rank. Overall, only 7% of the participants guessed that they were ranked among the less-attractive in the group as opposed to the expected 40% (test for proportions, z=-11.37, p<0.001). We will use 7 as a threshold for the distinction of "attractive" (ranks 1-6) and "less-attractive" ranks (ranks 7-10) throughout our analyses. In Appendix C we present robustness checks in which we (i) split the sample at the median using a threshold of 6, as well as (ii) split the sample at the 70th percentile using a threshold of 8. Splitting the sample at the alternative thresholds does not qualitatively change our main results.

Looking at the extent at which participants' self-assessments deviate from their actual rank, we see that on average the self-assessment of the "less attractive" individuals is more biased (average deviation=3.58, SD=1.90) than that of the "attractive" individuals (average deviation=0.36, SD=1.63). The difference is statistically significant (MW test, $z=11.68$, $p<0.001$).

RESULT 1: Participants are overconfident in their self-evaluation

The overconfidence in self-evaluation comes with a monetary cost to the participants. For a small amount of money, participants might get more benefit from deceiving themselves than from being honest about their attractiveness ranking. According to this explanation, when the cost of an overconfident answer goes up, the net benefit of self-deception will decrease and possibly become negative. This explanation predicts that the level of overconfidence would decrease as the cost of it increases.

To test this hypothesis, we compare the results above with the results of the High Stakes treatment, in which the cost of providing a biased self-evaluation is five times higher. The main result is that participants do not become more precise when the incentives to do so are increased (see Figure 1B).

Comparing the distribution of participants' guesses of their own rank observed in this treatment to the one observed in each of the $10 incentive treatments shows that increasing the incentive does not change the distribution of self-assessments ($\chi^2(9)=10.12$; 8.02; 10.91, $p\geq0.282$). When incentivized with $50 for a correct guess, individuals rank themselves on average 1.46 ranks better than their actual rank

(SD=2.35, N=99). Overall, 73% of the participants are overconfident, rating themselves as more attractive than they are perceived by others by 2.57 ranks on average (SD=1.58, N=72). Only 12% of the participants evaluate themselves as less attractive (ranks 7-10). This percentage significantly differs from the expected 40% (test for proportions, $z=-5.73$, $p<0.001$). When evaluating themselves, attractive individuals deviate from their actual rank by 0.17 on average (SD=1.88, N=59), whereas for the less attractive ones the deviation is higher with a value of 3.38 (SD=1.53, N=40). The difference is statistically significant (MW test, $z=7.10$, $p<0.001$).

RESULT 2: Overconfidence in self-evaluation is insensitive to the size of the incentives

3.3.2 Judgment of others

Judgment — Next, we explore individuals' accuracy when asked to privately judge the attractiveness of another individual of the same sex (stage three). Only if individuals are able to assess the rank of a random same-sex person objectively when incentivized, feedback can be useful to correct the bias in individuals' self-assessment.

We also explore whether the perception of attractiveness of individuals of the same sex is different from how individuals are perceived by the opposite sex. For this investigation, we use the data collected in the last stage of the Judgment treatment, in which individuals were asked to provide a full ranking of the individuals of their own sex. We pool the rankings assigned by the own group with the rankings assigned by

the opposite sex group. Merging all men and women's opinions about each individual provides a Cronbach's alpha of 0.94 for the men's ranking and a Cronbach's alpha of 0.95 for the women's ranking. The high correlation between the 19 rankings given to each individual (9 by own sex, 10 by opposite sex) suggests that men and women have a similar perception of attractiveness.

In an objective judgment distribution, we should observe all ranks from 1 to 10 equally often. The distribution of guesses (see Figure 2) is not significantly different from a discrete uniform distribution ($\chi^2(9)=11.00$, $p=0.275$, $N=100$). The average guessed rank of another same-sex participant in this treatment is 4.9.

RESULT 3: The average guessed rank of another same-sex participant is not biased

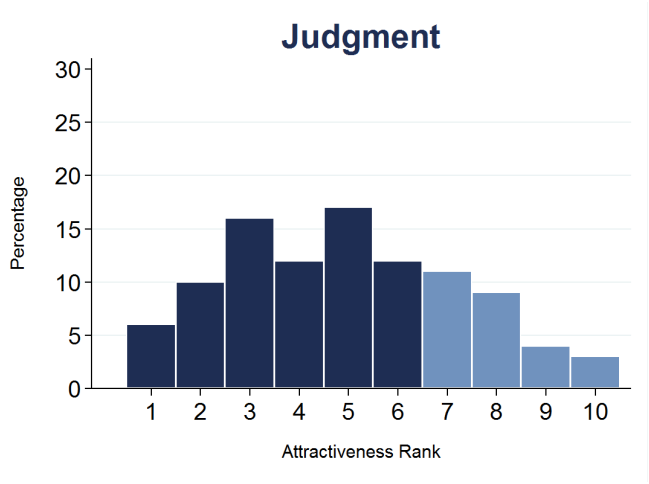

Figure 3.2 Distribution of Participants' Guesses of Others' Attractiveness

Figure 3A displays, for each given rank, the average deviation between a participant's actual rank and her counterpart's guess. Complementary to the figure,

Table 1 presents, for each given actual rank, detailed information on participants' average judgments and the resulting deviations from the actual ranks. A deviation of zero indicates that participants are on average correct in their assessment. We find that on average individuals ranked their counterparts 0.56 ranks better than their actual rank (SD=2.28, N=100). This deviation is significantly smaller than the 1.68 deviation (SD=2.07) when participants assess their own attractiveness (WS test, z=-3.03, p=0.001). These results suggest that while participants tend to considerably overestimate their own attractiveness, they perform better when asked to judge others.

Table 3.1 Average Assessment of Other's Attractiveness per Actual Rank

	Judgment		**F2F Feedback**		**High Stakes**		**Anonymous Feedback**	
Actual	Guess	Deviation	Guess	Deviation	Guess	Deviation	Guess	Deviation
1	2.3	-1.3	2.9	-1.9	2.6	-1.6	2.7	-1.7
2	3.2	-1.2	2.7	-0.7	2.9	-0.9	4.3	-2.3
3	4.6	-1.6	4	-1	4	-1	4	-1
4	4.1	-0.1	3.9	0.1	3.4	0.6	3.8	0.2
5	4.8	0.2	3.6	1.4	4.9	0.1	4.8	0.2
6	4	2	4.8	1.2	4.6	1.4	5	1
7	5.6	1.4	4	3	4	3	5.3	1.7
8	6.8	1.2	5.7	2.3	4.4	3.6	5.3	2.7
9	6.4	2.6	5.5	3.5	5.3	3.7	6.3	2.7
10	7.6	2.4	4.9	5.1	6.1	3.9	6.8	3.2
Overall	4.9	0.56	4.2	1.3	4.2	1.28	4.8	0.68

When we analyze the behavior of participants matched with attractive (ranks 1-6) versus less attractive (rank 7-10) counterparts, we see that this deviation is closer to zero for the former group (average deviation=-0.33, SD=2.18, WS test, z=-1.17,

p=0.243) than for the latter (average deviation=1.90, SD=1.71, WS test, z=4.95, p<0.001). The difference in deviations between the two aforementioned groups is statistically significant (MW test, z=4.94, p<0.001). While participants are on average precise when they evaluate attractive counterparts, they tend to evaluate less attractive counterparts as slightly better looking than they actually are. Nevertheless, when we compare how less attractive individuals assess themselves to how another same sex participant assesses them we find that in the latter case the deviation from the actual rank is significantly smaller (3.35 versus 1.9, MW test, z= -3.58, p<0.001).

Finally, we do not find evidence that the attractiveness of a participant affects the way he/she assesses an attractive/less attractive counterpart (MW tests, z=0.098; 0.987, p≥0.323).

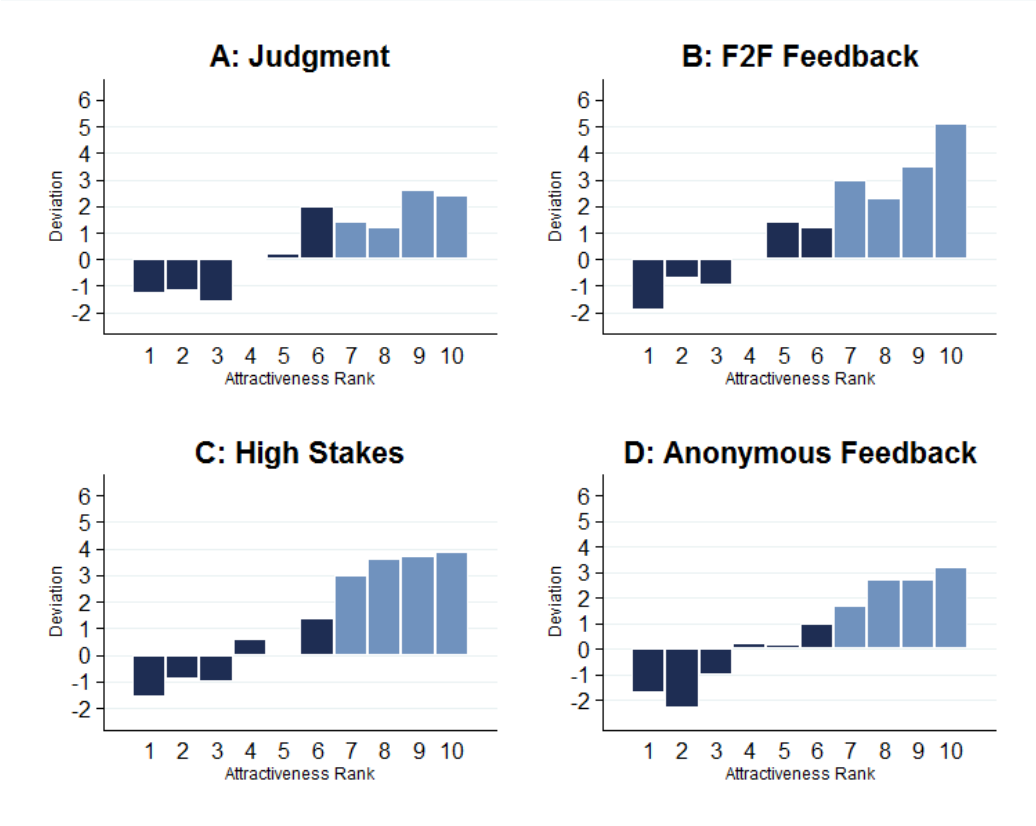

Figure 3.3 Deviation of Average Assessment of Others' Attractiveness Rank from Actual Rank

Note: A deviation of zero indicates that participants are on average correct in their assessment. The x-axis shows the given ranks 1-10 and on the y-axis the according deviation between the assessed rank of the other person and that person's actual rank is displayed. Negative values indicate that individuals attribute on average a rank to their counterpart that is worse than that person's actual rank. Likewise, positive values indicate that on average individuals attribute a rank to their counterpart that is better than that person's actual rank.

3.3.3 Feedback and Updating

F2F Feedback — If feedback is honest, the messages sent to others in the third stage of the experiment should not differ from the evaluations provided in stage three of the Judgment treatment. This is not what we find. In the F2F Feedback treatment participants send their counterparts overly positive messages (mean rank 4.2, N=100; Figure 4A). We find that the distributions of ranks in the Judgment and the F2F-

Feedback treatment are significantly different ($\chi^2(9)=17.89$, $p=0.036$). In line with our "must lie" hypothesis, this result is driven by the participants who are matched with a less attractive receiver. The distribution of ranks provided by these participants significantly differs between treatments ($\chi^2(8)=18.99$, $p=0.015$). By contrast, the distribution of ranks provided by participants matched with attractive individuals does not differ ($\chi^2(8)=5.91$, $p=0.657$). While 27% of the participants in the Judgment treatment evaluated their counterpart as "less attractive," only 7% in the F2F Feedback treatment did so ($\chi^2(1)=14.17$, $p<0.001$). Strikingly, none of the participants of the F2F Feedback treatment gave ranks 9 or 10 as feedback. Thus, individuals appear to avoid evaluating others as less attractive when this information is delivered face to face to the other person.

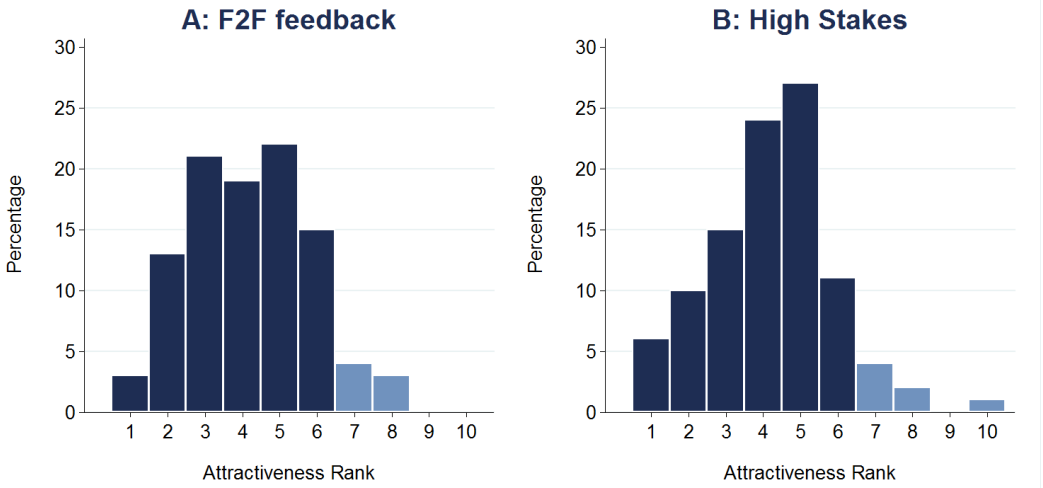

Figure 3.4 Deviation of Average Assessment of Others' Attractiveness Rank from Actual Rank

Figure 3B shows the average deviation between people's actual rank and the message they receive in stage three of the F2F Feedback treatment. We define feedback as dishonest if the deviation of individuals' assessment from the actual rank significantly differs from the deviation observed in the Judgment treatment. When limiting our analysis to participants who were matched with an attractive individual, in the F2F Feedback treatment this deviation is on average -0.15 (SD=1.74), which does not significantly differ from the -0.33 (SD=2.18) observed in the Judgment treatment (MW test, $z=-0.34$, $p=0.737$). This suggests that participants give honest feedback to attractive counterparts. However, when matched with a less attractive participant, the average deviation is significantly larger in the F2F feedback treatment than in the Judgment treatment (3.48 (SD=1.72) vs. 1.90 (SD=1.71) respectively (MW test, $z=-3.68$, $p=<0.001$).

An additional investigation of the precision of feedback is reported in Appendix B, where we look at the number of correct guesses depending on the receiver's attractiveness, allowing for a one-rank deviation from participants' actual rank in both directions. This analysis confirms our results. We find that when matched with less attractive individuals, a significantly lower fraction of participants correctly guess their counterpart's rank in the F2F Feedback treatment than in the Judgment treatment.

Table 3.2 Likelihood of Providing Negative Feedback to the Less Attractive Participants

Probability (Negative feedback)	(1)	(2)	(3)	(4)
F2F Treatment	-.32***	-.31***	-.31***	-.31***
	(.06)	(.06)	(.07)	(.07)
High Stakes Treatment	-.29***	-.29***	-.30***	-.28***
	(.06)	(.06)	(.07)	(.07)
Anonymous Treatment	-.12	-.12	-.12	-.11
	(.08)	(.08)	(.08)	(.09)
Female		-.01	.01	-.01
		(.07)	(.07)	(.08)
Age			-.01	-.02
			(.02)	(.02)
Ethnicity Control	N	N	N	Y
Observations	160	160	152	152

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Note: The table presents marginal effects estimated from Probit Regressions. Negative feedback is defined as feedback that categorizes the receiver as less attractive (ranks 7-10). Probability (Negative Feedback) is the predicted probability of giving negative feedback in each of the three treatments as compared to judging a less attractive individual negatively. Therefore, the baseline treatment is the Judgment treatment. Female is a gender dummy and age is a continuous variable. All marginal effects are evaluated at the change of the dummy variable from 0 to 1.

We further examine how the likelihood that less attractive individuals are ranked as less attractive varies across treatments using probit regression. The results are reported in Table 2. The regression analysis confirms that participants in the F2F Feedback treatment are significantly less likely to provide accurate ranks to the less attractive individuals than participants in the Judgment treatment. In particular, the estimated marginal effects in column 1 show that participants in the F2F Feedback treatment are 32 percentage points less likely to send a message with a negative rank (ranks 7-10) than participants in the Judgment treatment. In column 2 we add a dummy variable to control for gender and show that the result stays significant and almost identical in effect size (31 percentage points). The same is true for the inclusion of the age variable in column 3 and ethnicity dummies in column 4.

In Appendix A, (Table A1) we report an additional analysis on the participants who were matched with attractive individuals, investigating possible treatment differences in the likelihood to provide accurate feedback (rank 1-6) to attractive individuals. We find that participants in the F2F Feedback treatment are 5 percentage points more likely to evaluate their attractive counterpart as attractive than participants in the Judgment treatment. This result, however, is only marginally significant and becomes insignificant when including controls for gender and age. Taken together, these results suggest that while feedback towards attractive individuals tends to be precise, the feedback provided to the less attractive individuals is biased.

RESULT 4: When the sender of feedback is identified (F2F), participants provide positively biased feedback to less attractive participants

Next, we investigate how participants react to the feedback they receive and whether they use it to update their own rank. A class of equilibria exists in this game in which the receivers of F2F Feedback expect the feedback to be biased, and treat it as such. This does not seem to be the case here. We find that 81% of the participants received feedback that was different from their self-evaluation and most of them (67%) updated their self-evaluation in the direction of feedback. Out of the participants who received positive feedback as compared to their self-evaluation, 75% updated to a better rank, suggesting that individuals do not treat feedback as biased. Out of the participants who received feedback that was negative compared to their self-evaluation, 60% updated to a worse rank. The difference in the fraction of people

who update in the direction of the message received after negative and positive feedback is not statistically significant ($\chi^2(1)=1.87$, $p=0.171$). These results are confirmed by a probit regression investigating the likelihood that participants update in the direction of the feedback they receive, controlling for the distance between individuals' self-assessment and the message that they receive (see Appendix A, Table A2).

Next, we investigate the *extent* to which individuals update after receiving feedback. Figure 5 displays the number of ranks by which participants update in stage four as a function of the number of ranks by which the feedback received deviates from participant's initial self-assessment. A negative deviation indicates that participants received positive feedback compared to their self-evaluation, while a positive deviation indicates that individuals received negative feedback compared to their self-evaluation. The OLS regression lines plotted in the figure display updating behavior as a function of this deviation, separately for cases in which the feedback received is positive and cases in which it is negative. As can be seen from the figure, the regression line is less steep when participants receive negative feedback, suggesting that these participants update to a lower extent than participants who receive positive feedback.

We further investigate differences in updating behavior in these two cases in Table 3. Column 1 shows that the larger the difference between participants' self-evaluation and the feedback they receive, the more they update. This is the case both if participants receive positive feedback ($\beta=.54$, $p<0.001$) and if they receive negative

feedback (β=.23, p<0.001). However, the difference between the two coefficients is significant, which confirms that participants update to a greater extent when feedback is positive than when it is negative (F(1, 95)=10.83, p=0.001).

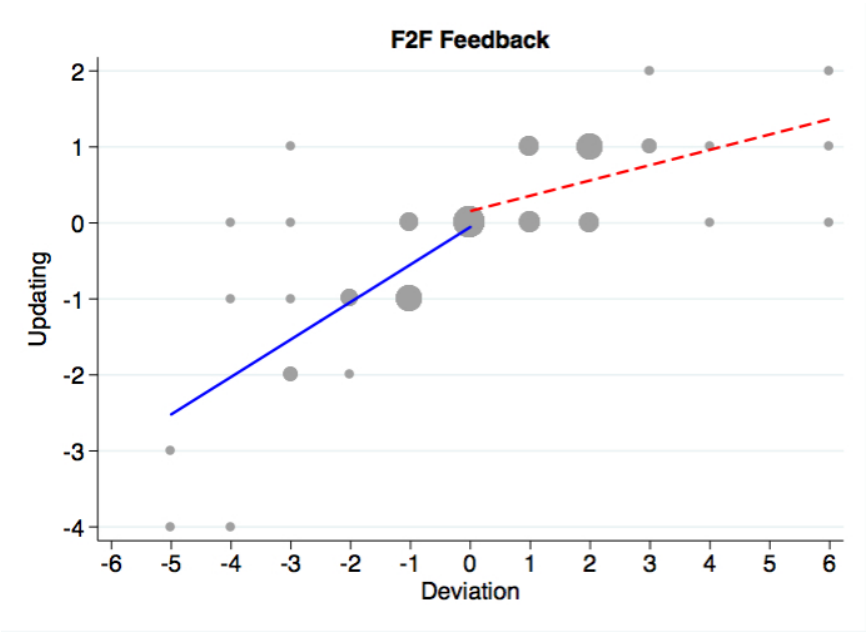

Figure 3.5 Updating Behavior after F2F Feedback

Note: The x-axis shows the deviation between the rank that participants receive as feedback and their initial self-evaluation (from Stage 2). Negative values indicate feedback that is more positive than individuals' self-evaluation while positive values indicate feedback that is negative compared to participants' initial self-evaluation. The y-axis shows the difference between the updated self-assessment in stage four and the self-assessment in stage one. Negative values indicate an update to a better rank while positive values indicate an update to a worse rank. The blue line displays the estimated OLS regression line for when participants receive positive feedback, while the red line displays the estimated OLS regression line for the cases in which participants receive negative feedback.

This result is not fully in line with the findings of Eil and Rao (2011), who found that updating based on negative signals is either strongly discounted or completely ignored by the individuals.

Table 3.3 Degree of Updating by Treatment

Extent of Updating	(1)	(2)	(3)	(4)	(5)	(6)
	F2F Feedback	F2F Feedback	High Stakes	High Stakes	Anonymous	Anonymous
Self-mes received (pos.)	.54***	.53***	.44***	.44***	.41***	.41***
	(.06)	(.06)	(.08)	(.08)	(.08)	(.08)
Self-mes received (neg.)	.23***	.24***	.25***	.25***	.28***	.28***
	(.05)	(.05)	(.07)	(.07)	(.05)	(.05)
Attractiveness		.13		.09		-.04
		(.14)		(.14)		(.15)
				.		
Constant	.08	-.01	-.04	-.11	.03	.05
	(.10)	(.14)	(.11)	(.13)	(.11)	(.14)
Observations	98	98	99	99	99	99

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Notes: The table presents OLS estimates. Extent of updating indicates by how much individuals update in stage four as compared to their original self-evaluation provided in stage two: negative values indicate that individuals updated to a better rank while positive values indicate that individuals updated to a worse rank. Self-message received (pos.) is a variable that indicates the difference between participants' self- evaluation in stage two and the message they receive from others when feedback is positive (as compared to subjects' self-evaluation), and is coded as 0 otherwise. Self-message received (neg.) is a variable that indicates the difference between participants' self-evaluation and the message they receive from others when feedback is negative, and is coded as 0 otherwise. Positive (negative) feedback means that the feedback indicates a better (worse) rank than the self-evaluation from Stage two. Attractiveness is a dummy coded as 1 if individuals are attractive (ranks 1-6). Standard errors are reported in parenthesis.

Finally, we find no difference in updating behavior between attractive and the less attractive participants. The fraction of participants who update in the direction of feedback is the same when the feedback is negative (62% vs. 59% $\chi^2(1)=0.03$, p=0.859) and when it is positive (75% vs. 75%, $\chi^2(1)=0.00$; p=1.000).

RESULT 5: A majority of participants update their self-assessment in the direction of the feedback they receive. They update to a larger extent after positive feedback relative to negative feedback

Feedback should help individuals make more precise guesses. In order to explore whether receiving feedback helps the recipient, we investigate how individuals' precision in their self-assessment changes after feedback (stage four) as compared to before feedback (stage two). For the attractive participants the deviation changes from 0.07 to 0.08 (WS test, z=0.26, p=0.795) while for the less attractive individuals it changes from 3.66 to 3.98 (WS test, z=-0.73, p=0.467). Both changes are not significant. Hence, F2F feedback does not improve the accuracy of the assessment neither for the attractive nor the less attractive participants.

RESULT 6: F2F Feedback does not change the accuracy of the self-assessments relative to no feedback

High Stakes — Comparing the results obtained in the Judgment treatment to those obtained in the F2F Feedback treatment shows that individuals are reluctant to provide negative feedback to less attractive participants. Although the senders might intend to "be nice" by giving more flattering feedback, such kindness comes at a monetary cost to themselves and to their counterpart. As with self-confidence, while people might be willing to provide an overly nice feedback when the price for doing so is $10, they might be less willing to do so when the price is $50.

We find that higher stakes do not induce participants to send more honest messages (mean rank 4.22, N=100). Only 7% of the participants in High Stakes told their counterpart that they are "less-attractive" (Figure 4B). The distribution of feedback in this treatment does not differ from the distribution of F2F-Feedback with low stakes ($\chi^2(8)=5.30$, p=0.725), while it significantly differs from the distribution of ranks observed in Judgment ($\chi^2(9)=19.07$, p=0.025). Again, the difference in the

distribution of ranks is driven by the feedback to the less-attractive individuals ($\chi^2(8)=17.77$, $p=0.023$). For the attractive individuals the two distributions do not differ ($\chi^2(8)=8.62$, $p=0.375$).

The results on the deviation between participant's actual ranks and the feedback provided to them in this treatment are almost identical to those observed in the F2F Feedback with low stakes (Figure 3C). For participants matched with attractive participants the average deviation is -0.23 (vs. -0.15 in F2F Feedback) while for those matched with the less attractive participants it is 3.55 (vs. 3.48 in F2F Feedback). As in the F2F treatment, when participants are matched with a less attractive participant, the average deviation is significantly larger in the High Stakes treatment than in the Judgment treatment (MW test, $z=-4.15$, $p<0.001$), while we do not find differences for participants matched with an attractive participant (MW test, $z=-0.26$, $p=0.796$)

The probit regression results reported in Table 2 show that participants in the High Stakes treatment are 29 percentage points less likely to assign negative ranks to less attractive individuals than participants in the Judgment treatment. A comparison between the coefficients of the High Stakes and F2F Feedback treatment shows that the likelihood of providing honest negative feedback to the less attractive participants does not differ between these two treatments ($\chi^2(1)=0.11$, $p=0.746$).

The additional analyses on the participants matched with attractive individuals reported in Appendix A show that participants are 6 percentage points more likely to evaluate their attractive counterpart as attractive (ranks 1-6) in High Stakes than in the the Judgment treatment.

RESULT 7: Increasing the incentives does not increase the accuracy of face to face feedback towards the less attractive individuals

Similarly to the F2F Feedback treatment, participants updated in both directions depending on the content of the message received: 43% of those who were given negative feedback updated to a worse rank and 76% of those that were given positive feedback updated to a better rank. The difference in the fraction of people who update (in the direction of feedback) after receiving negative and positive feedback is statistically significant ($\chi^2(1)=8.71$, $p=0.003$). The probit regression in Table A2 (Appendix A) confirms that individuals are less likely to update in the direction of feedback when feedback is negative as compared to positive.

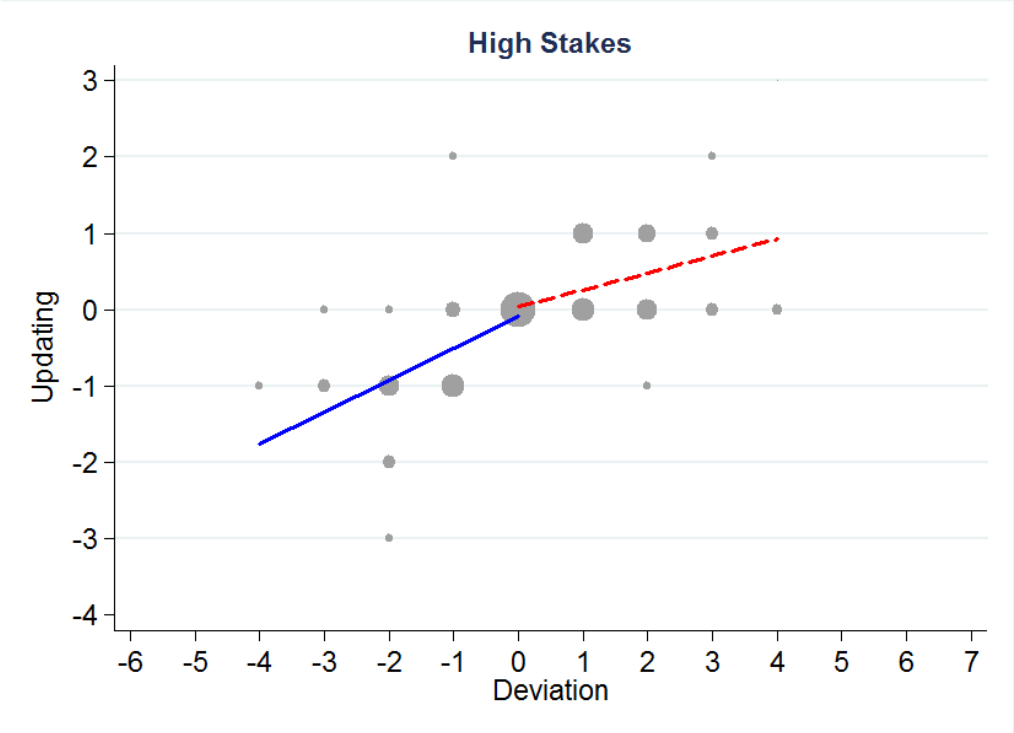

Figure 3.6 Updating Behavior after F2F High Stakes Feedback

Figure 6 illustrates the extent to which participants update after positive and negative feedback: as in the F2F feedback treatment with low stakes, we see that the regression line is less steep when individuals receive a negative signal. The OLS regression reported in Table 3 shows that participants update more the larger the difference between their initial self-assessment and the feedback they receive, both when the feedback is positive (β=.44, $p<0.001$) and when it is negative (β=.25, $p=<0.001$). However, the difference between the two coefficients is not statistically significant ($F(1,96)=2.24$, $p=0.136$).

Finally, looking at the precision of individuals' self-assessment in the updating stage we see that for the attractive participants the deviation between individuals' self-assessment and their actual rank does not significantly change after feedback (0.17 to 0.22, WS test, $z=0.37$, $p=0.709$), while for the less attractive participants it becomes significantly larger, changing from 3.38 to 3.65 (WS test, $z=-2.18$, $p=0.029$).

RESULT 8: F2F Feedback with high stakes lowers the accuracy of the self-assessments relative to no feedback

Anonymous Feedback — The reluctance to give honest feedback to less attractive individuals could be driven by the desire of protecting one's own image. Alternatively, participants may be unwilling to hurt another person's feelings by giving negative feedback due to altruistic preferences. To distinguish between these two mechanisms, we study the effect of anonymous feedback. If people are reluctant to hurt other people's feelings, we expect similar behavior in both feedback

treatments. Instead, if participants (also) care about being identified as the person who is providing the negative feedback, we expect more honest feedback when their identity is kept anonymous.

We find that feedback in the Anonymous treatment is less positive than in the F2F Feedback treatments (see Figure 7). The average rank is similar to the average observed in the Judgment treatment (mean rank 4.84, N=99). We find that the distributions of ranks provided in Anonymous and Judgment are not significantly different from each other ($\chi^2(9)=7.07$, $p=0.630$). On the other hand, the difference between the distribution of ranks in Anonymous and F2F Feedback is marginally significant ($\chi^2(8)=14.98$, $p=0.091$). Comparing the distributions of participants matched with an attractive and those matched with unattractive counterparts separately, however, yields insignificant results in both cases ($\chi^2(8)=12.38$ and 7.38, $p\geq0.135$). The difference between the distributions of ranks in Anonymous and High Stakes is not significant ($\chi^2(8)=12.88$, $p=0.168$).

However, the fraction of individuals who evaluated their partner as "less-attractive" in Anonymous is around three times higher than in F2F (7% vs. 20%, $\chi^2(1)=7.39$, $p=0.007$) and 3 times higher than in High Stakes (7% versus 20%).

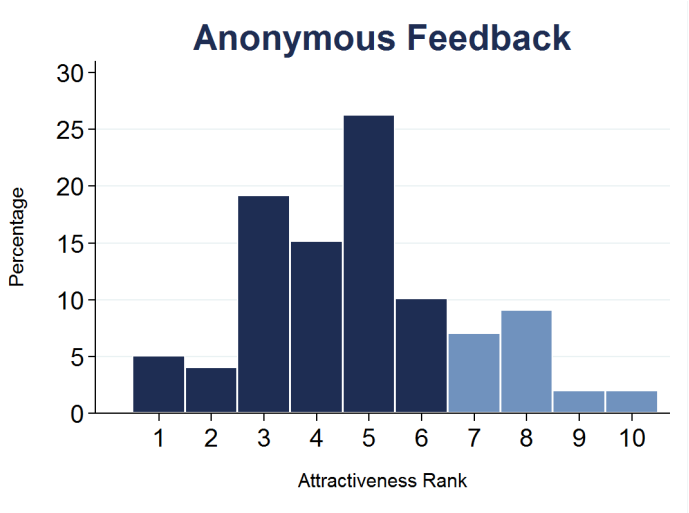

Figure 3.7 Distribution of Guesses of Others' Attractiveness

Figure 3D shows the average deviation between participants' actual rank and the anonymous message they receive. The average deviation between the feedback provided in stage three and the less attractive participants' actual ranks is 3.48 in the F2F Feedback treatment, while it is only 2.58 in the Anonymous treatment. The difference between treatments is statistically significant (MW, z=2.04, p=0.042). For the attractive individuals we find no significant difference; the average deviation is -0.15 in the F2F Feedback and -0.61 in the Anonymous treatment (z=0.95, p=0.340). When we compare the deviations between Anonymous and Judgment, we find that they do not differ for the attractive individuals (average deviation=-0.33 versus -0.61, MW test, z=0.61, p=0.540) and is marginally different for the less attractive participants (average deviation=1.9 versus 2.58, MW test, z=-1.73, p=0.084).

The probit regression reported in Table 2 shows that participants in the Anonymous treatment are equally likely to provide negative feedback to less attractive individuals as participants in the Judgment treatment. The comparison of the

coefficient estimates for the Anonymous Feedback treatment to those of the other treatments confirms that participants in the Anonymous treatment are significantly more likely to provide negative feedback than those in the F2F Feedback treatment ($\chi^2(1)=6.57$, $p=0.010$) and High Stakes treatment ($\chi^2(1)=05.18$, $p=0.023$).

The additional analyses on the participants matched with attractive individuals reported in Appendix A (Table A1) show that participants are equally likely to evaluate their attractive counterpart as attractive in the Anonymous and Judgment treatment.

RESULT 9: When feedback is anonymous individuals are more accurate in their assessment than when their identity is revealed to the receiver

Similarly to the other treatments, in the Anonymous treatment 70% of the participants among those who received feedback different from their self-evaluation updated their guess in the direction of feedback: 68% of those receiving negative feedback updated to a worse rank and 75% of those receiving positive feedback updated to a better one. The fractions of people who update in the direction of the message after receiving negative and positive feedback are not statistically different ($\chi^2(1)=0.36$, $p=0.551$). This result is confirmed by the probit regression reported in Table A2.

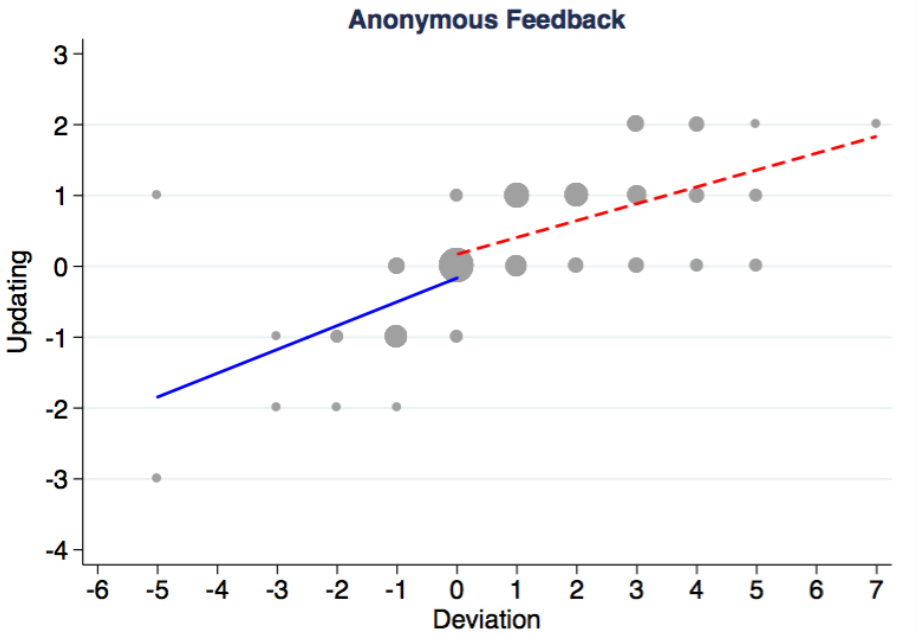

Figure 3.8 Updating Behavior after Anonymous Feedback

Taking into account the extent to which individuals update, Figure 8 illustrates the updating behavior after anonymous feedback. As in the other treatments, the OLS regression in Table 3 shows that the larger the difference between participants' initial self-assessment and the feedback they receive, the more they update, both when the feedback is positive (β=.41, $p<0.001$) and when it is negative (β=.28, $p<0.001$). As it can also be seen in the figure, the difference between the two coefficients is not statistically significant ($F(1,96)=1.44$, $p=0.234$).

Since individuals use the feedback received to update their self-evaluation and given that feedback is more honest in the anonymous treatment, we find feedback in this treatment to be beneficial to the participants, as it helps them correct their biased self-evaluation. In the updating stage, the average difference between participants'

self-evaluation and their actual rank significantly decreases from 3.7 to 3.4 (WS test, z=2.19, p=0.028) for the less attractive participants, while for the attractive ones it decreases from 0.43 to 0.2 (WS test, z=1.98, p=0.048).

RESULT 10: Anonymous feedback improves the accuracy of the self-assessment as compared to no feedback.

Our results show that the reluctance to provide negative face-to-face feedback to less attractive individuals is mainly driven by individuals' unwillingness to be identified as the messenger of bad news. Participants provide more accurate negative feedback when the anonymity of the feedback provider is guaranteed.

3.4 Concluding Remarks

Are there "must-lie" situations? In a novel experimental design we show that when asked to give face-to-face negative feedback people prefer to lie rather than provide an honest assessment. This aversion to providing negative feedback is costly because it slows down the learning process and possibly encourages superfluous investments of either time or money. Our results suggest that especially in cases where an individual's own perception is the furthest from the perception of others, a lack of honest feedback leads to greater distortion of self-perception and a worse outcome than without feedback.

The inflated feedback relates to the literature on deception in which people seem to have costs associated with lying (Charness and Dufwenberg, 2006, Dreber and Johannesson, 2008, Gneezy, 2005, Sutter, 2009, Erat and Gneezy, 2012). In our case, people have costs associated with telling the truth. These cost could be time, effort, money, or – as our results suggest – negative utility from openly delivering negative messages, which creates the "must lie situation." We show that increasing the stakes fivefold does not help to outweigh this cost.

This bias in the provision of information could have strong negative effect on, for example, transmission of information in organizations. In situations in which giving a negative feedback to a co-member of the organization may reflect badly on the sender, the organization may suffer from a slower learning process than is feasible.

The observation that the aversion to giving negative feedback is reduced when it is done anonymously suggests that guaranteeing the anonymity of the feedback provider is important for the accuracy and informational value of feedback. Consider the example of the peer-review process of journals. An obvious reason for making this process anonymous is the repeated game consideration, in which people would be reluctant to give negative feedback out of reciprocity concerns. Consequently, informing the author about the identity of the reviewers may lead to overly positive assessments of research articles and thus reduce the accuracy of evaluations. For example, if we were to ask our friends about their opinion regarding this paper, we might not receive honest feedback. In an anonymous referee process the same friend would be much more critical and thus, more helpful.

In an organizational context, in which the efficiency of team work critically hinges on honest communication, our results suggest that in addition to having an open feedback round, organizations should consider installing a 'feedback box' in which staff can anonymously provide feedback to their colleagues or bosses.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Gneezy, Uri, Christina Gravert, Silvia Saccardo, and Franziska Tausch, "A Must Lie Situation: Avoiding Giving Negative Feedback." The dissertation author was the co-primary investigator and author of this paper.

Appendix A. Additional Analyses

A1. Additional Tables

Table 3.A1 Probability of Giving Honest, Positive Feedback To Attractive Participants

Probability (Positive feedback)	(1)	(2)	(3)	(4)
F2F Treatment	.05*	.04*	.03	.03
	(.03)	(.02)	(.03)	(.03)
High Stakes Treatment	.06***	.06**	.05*	.06**
	(.02)	(.02)	(.03)	(.03)
Anonymous Treatment	.01	.01	-.01	-.00
	(.03)	(.03)	(.03)	(.03)
Female		.04	.03	.03
		(.03)	(.03)	(.03)
Age			-.00	-.00
			(.01)	(.01)
Ethnicity Control	N	N	N	Y
Observations	239	239	224	202

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Note: The table presents marginal effects estimated from Probit Regressions. Positive feedback is defined as feedback that categorizes the receiver as attractive (ranks 1-6). Probability (Positive Feedback) is the predicted probability of giving positive feedback in each of the three treatments as compared to judging an attractive individual positively. Therefore, the baseline treatment is the Judgment treatment. Female is a gender dummy and age is a continuous variable. All marginal effects are evaluated at the change of the dummy variable from 0 to 1. Specification 4 is missing 22 observations, as none of the Hispanic provides positive feedback to an attractive partner. Given the lack of variance in the dependent variable for the Hispanics the observations were omitted.

In this regression we investigate the likelihood of telling attractive individuals (ranks 1-6) that they are attractive. As displayed in the table, we find that individuals in the High Stakes treatment are more likely to provide positive feedback to an attractive individual as compared to Judgment. Also for the F2F Treatment the relation is marginally significant in column 1, but becomes insignificant when controlling for age, ethnicity and gender. No such effect exists for the Anonymous treatment.

Table 3.A2 Probability Of Updating In The Direction Of Feedback By Treatment

Probability	(1)	(2)	(3)	(4)	(5)	(6)
	F2F Feedbac	F2F Feedbac	High Stakes	High Stakes	Anonymou s	Anonymous
Negative Feedback	-.15	-.15	-	-.38***	-.08	-.08
	(.10)	(.10)	(.11)	(.11)	(.12)	(.12)
|Self-message	.05	.05	.07	.06	.02	.02
	(.04)	(.04)	(.07)	(.07)	(.04)	(.04)
Attractiveness		.04		-.14		.00
		(.11)		(.13)		(.11)
Observations	79	79	76	76	75	75

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Note: The table presents marginal effects estimated from Probit Regressions. Probability (Updating) is the probability of updating in the direction of feedback. The dependent variable is a dummy coded as 1 if people update in the direction of the feedback they receive, and 0 if they do not update or update in the opposite direction. The Negative Feedback variable is a dummy coded as 1 if individuals received negative feedback and zero if they received positive feedback. |Self-message received| is a variable that indicates the absolute difference between participants' self-evaluation and the message they receive from others. Attractiveness is a dummy coded as 1 if individuals are attractive (ranks 1-6). All marginal effects are evaluated at the change of the dummy variable from 0 to 1. For this analysis we only consider individuals that received feedback that is different from their self-evaluation.

With this probit regression we investigate the probability of updating in the direction of feedback. We regress a dummy indicating whether people update in the direction of feedback on the distance between individuals' self-assessment and the message that they receive. We find that in the F2F treatment participants who receive negative feedback are equally likely to update in the direction of the feedback as those who receive positive feedback (p=0.142). In the High Stakes treatment participants who receive negative feedback are 35 percentage points less likely to update in that direction than those who receive positive feedback (p=0.001). In the Anonymous treatment participants who receive negative feedback are equally likely to update negatively as compared to those who receive positive feedback (p=0.516). The result doesn't change if we control for participants' attractiveness.

A2. Additional analysis on the precision of Stage three evaluations

We further investigate the precision of participants' assessment of others during stage three by looking at the percentage of evaluations that match the counterpart's actual attractiveness rank. Since guessing someone's rank exactly right is difficult and involves some degree of luck, in our analysis we investigate the number of correct guesses allowing a one-rank deviation from participants' actual rank in both directions. This analysis provides further support to the findings on how the precision of feedback varies depending on the attractiveness of participants' counterparts. In particular, in the F2F Feedback treatment participants are more likely

to provide a correct evaluation when matched with an attractive then when matched with a less attractive participant. Of all the participants matched with attractive participants, 58% evaluate their counterpart within one rank of their actual attractiveness in the Judgment treatment, while a similar fraction of participants (62%) does so in the F2F Feedback treatment ($\chi^2(1) = 0.14$, $p = 0.709$). Of all the participants matched with less attractive counterparts, this fraction is 40% in the Judgment treatment and only 15% in F2F Feedback treatment, ($\chi^2(1) = 6.27$, $p = 0.012$).

In the High Stakes treatment the results are very similar. 65% of the participants matched with attractive counterparts assessed them within one rank from their actual attractiveness. This fraction is smaller (5%) when participants are matched with less attractive counterparts ($\chi^2(1)=35.72$, $p<0.001$), it is significantly different from the 40% observed in the Judgment treatment ($\chi^2(1) = 14.05$, $p<0.001$) and does not differ from the 15% observed in the F2F Feedback treatment ($\chi^2(1) = 2,22$, $p = 0.136$).

In the Anonymous treatment, we find that the fraction of correct guesses when participants are matched with attractive individuals is similar to the fraction we observe in all the other treatments (59%). Instead, when participants are matched with less attractive counterparts this fraction is 33%, which does not differ from the respective fraction observed in the Judgment treatment ($\chi^2(1)=0.49$, $p=0.485$) and is significantly larger than the fraction of correct guesses in both F2F Feedback treatments (F2F: $\chi^2(1)=3.38$, $p=0.066$, High Stakes: $\chi^2(1)=9.92$, $p=0.002$). This result confirms that individuals are unwilling to provide negative feedback when feedback is provided face to face. When feedback is anonymous instead, individuals are much more precise.

A3. Robustness check

In this section we investigate whether the results reported in the main text are robust to changing the cut-off rank that we use as a threshold between attractive and less attractive individuals. We conduct two robustness checks with respect to the rank threshold on all estimations where the threshold was relevant. First, we conduct a median split and redefine as “attractive” all participants with a rank between 1-5, and as “less attractive” all participants with ranks 6-10. Second, we look at the upper 30% of the distribution and redefine as “attractive” participants with ranks 1-7, and as “less attractive” participants with ranks 8-10. Overall, the results remain robust. Exceptions are clearly noted below.

Median split. The self-assessment analysis yields qualitatively the same results as with a cutoff of 7. In the self-evaluation phase with the $10 incentives (N=298), 17% of the participants guessed that they were ranked among the less attractive in the group, as opposed to the expected 50% (test for proportions, $z=11.24$, $p<0.001$). Looking at the extent to which participants’ self-assessments deviate from their actual rank, we see that on average the self-assessment of the “less attractive” individuals is more biased (average deviation=3.16, SD=2.07) than that of the “attractive” individuals (average deviation=0.13, SD=1.47) (MW test, $z=11.44$,

$p<0.001$). In the High Stakes Treatment only 19% of the participants evaluate themselves as less attractive (ranks 6-10). This percentage significantly differs from the expected 50% (test for proportions, $z=6.03$, $p<0.001$). When evaluating themselves, attractive individuals deviate from their actual rank by -0.16 on average (SD=1.72), whereas for the less attractive ones the deviation is higher with a value of 3.12 (SD=1.64, MW test, $z=7.40$, $p<0.001$).

In the Judgment treatment we analyze the attractiveness assessment of participants matched with attractive versus less attractive counterparts and we see that as in the main analysis this deviation is closer to zero for the former group (average deviation=-0.80, SD=1.94, WS test: different from zero, $z=-2.64$, $p=0.008$) than for the latter (average deviation=1.92, SD=1.72, WS test: different from zero, $z=5.51$, $p<0.001$). The difference in deviations between the two aforementioned groups is again statistically significant (MW test, $z=6.13$, $p<0.001$). When we then compare how less attractive individuals assess themselves to how another same sex participant assesses them we find that in the latter case the deviation from the actual rank is significantly smaller (2.96 versus 1.92, MW test, $z=-2.913$, $p=0.004$). All results stay robust to the new threshold.

Next we test whether the different threshold leads to any differences in the F2F Feedback treatment. Again, the distribution of ranks provided by the participants matched with a less attractive individual significantly differs between Judgment and F2F Feedback ($\chi^2(8)=17.39$, $p=0.043$). As before, the distributions of ranks provided by participants matched with attractive individuals do not differ ($\chi^2(8)=8.28$, $p=0.407$). While 39% of the participants in the Judgment treatment evaluated their counterpart as "less attractive," only 22% in the F2F Feedback treatment did so ($\chi^2(1)=6.82$, $p=0.009$).

The average deviation between the F2F feedback provided and participants' actual ranks for the participants who were matched with an attractive individual is -0.42 (SD=1.65), as compared to -0.8 (SD=1.94) in the Judgment treatment (MW test, $z=-0.71$, $p=0.479$). As in the analysis we report in the main text where we use 7 as a threshold, the difference is not significant. When matched with a less attractive participant, the average deviation is significantly larger in the F2F feedback treatment than in the Judgment treatment (3.02 (SD=1.92) vs. 1.92 (SD=1.72) respectively (MW test, $z=-2.82$, $p=0.005$). We further examine the likelihood that less attractive individuals are ranked as less attractive using a regression framework. We re-run the probit regressions introduced in the main text using the new threshold of 6 (see Table C1). The estimates for the F2F Feedback treatment become smaller than with a threshold of 7, but the results remain significant at the 10 percent level.

Table 3.A3 Likelihood of Providing Negative Feedback To The Less Attractive Participants (Median Split)

Probability (Negative feedback)	(1)	(2)	(3)	(4)
F2F Treatment	-.17*	-.17*	-.17*	-.15
	(.09)	(.09)	(.09)	(.10)
High Stakes Treatment	-.28***	-.29***	-.28***	-.26***
	(.09)	(.08)	(.09)	(.09)
Anonymous Treatment	-.12	-.12	-.11	-.10
	(.09)	(.09)	(.10)	(.10)
Female		-.04	-.05	-.06
		(.07)	(.07)	(.07)
Age			-.00	-.00
			(.02)	(.02)
Ethnicity Control	N	N	N	Y
Observations	200	200	190	190

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Note: The table presents marginal effects estimated from Probit Regressions. Probability (Negative Feedback) is the predicted probability of giving negative feedback (ranks 6 to 10) in each of the three treatments compared to judging a less attractive individual negatively. The baseline treatment is the Judgment treatment. Female is a gender dummy and age is an ordinal variable. All marginal effects are evaluated at the change of the dummy variable from 0 to 1.

Next we see if there are any differences in the updating results with the new threshold. We find that the results are qualitatively the same as in the estimation reported in the main text. For the attractive participants the deviation between the initial self-assessment and the self-assessment after feedback changes from -0.20 to -0.18 (WS test, z=0.17, p=.865) while for the less attractive individuals it changes from 3.19 to 3.46 (WS test, z=-0.54, p=0.587).

In the High Stakes treatment 18% gave their counterparts feedback that they are among the less attractive participants. Again, the difference in the distribution of ranks as compared to the Judgment treatment is driven by the feedback to the less-attractive individuals ($\chi^2(8)=17.74$, $p=0.038$).

For participants matched with attractive participants the average deviation between actual rank and feedback provided is -0.56 (vs. -0.42 in F2F Feedback) while for those matched with the less attractive participants it is 3.12 (vs. 3.02 in F2F Feedback). We can confirm that the average deviation for those matched with a less attractive participant is significantly larger in the High Stakes treatment than in the Judgment treatment (MW test, z=-3.29, p=0.001), while we do not find differences for participants matched with an attractive participant (MW test, z=-0.47, p=0.639). Table C1 shows that the probit results with the alternative threshold for High Stakes remain robust and significant.

Looking at the precision of individuals' self-assessment in the updating stage we see that for the attractive participants the deviation between individuals' self-assessment and their actual rank does not significantly change after feedback (-0.16 to -0.22, z= 0.590, p=0.555), while for the less attractive participants it becomes

significantly larger, changing from 3.12 to 3.4 (z=-2.25, p=0.024). This is in line with the findings using 7 as the attractiveness threshold.

The first difference we find with the threshold at the median is in the Anonymous treatment. While with a threshold of 7 we find that significantly more individuals evaluated their partner as "less-attractive" in Anonymous as compared to F2F, this difference is not significant anymore under the new threshold (22% vs. 30%, $\chi^2(1)=1.78$, p=0.183). However, importantly the results on the precision of feedback remain robust. The average deviation between the feedback provided in stage three and the less attractive participants' actual ranks is 3.02 in the F2F Feedback treatment, while it is only 2.26 in the Anonymous treatment (MW, z=1.84, p=0.065). For the attractive individuals we again find no significant difference; the average deviation is -0.42 in the F2F Feedback and -0.94 in the Anonymous treatment (z=1.05, p=0.295). When we compare the precision of feedback in the Anonymous Feedback treatment to the precision of evaluations in the Judgment treatment, we find that, similarly to the main analysis, the average deviation does not differ for the attractive individuals (average deviation=-0.80 versus -0.94, MW test, z=-0.38, p=0.701) and with the median threshold it also does not differ for the less attractive participants (average deviation=1.92 versus 2.26, MW test, z=-1.03, p=0.304).

Table C1 reports the results on the likelihood that less attractive individuals are ranked as less attractive are reported in Table C1. We find that participants in Anonymous are equally likely to provide negative feedback to less attractive individuals as participants in the Judgment treatment. This supports the result that when ask to provide anonymous feedback, subjects matched with a less attractive individual are more honest than when asked to provide F2F feedback.

As in the main analysis we find feedback in this treatment to be beneficial to the participants, as it helps them correct their biased self-evaluation. In the updating stage, the average difference between participants' self-evaluation and their actual rank significantly decreases from 3.32 to 3.02 (WS test, z=2.20, p=0.028) for the less attractive participants, while for the attractive ones it decreases from 0.18 to -0.6 (WS test, z=1.91, p=0.056).

Taken together, the results from the robustness check with the median split provide support to the results we observe when using 7 as a threshold.

70th Percentile Split. The self-assessment analysis using a threshold at the 70th percentile (rank 8) yields qualitatively the same results as the ones we find using rank 7 as a threshold. In the self-evaluation stage with the $10 incentive, 4% of the participants guessed that they were ranked among the less-attractive in the group as opposed to the expected 30% (test for proportions, z=-9.651, p<0.001). Looking at the extent at which participants' self-assessments deviate from their actual rank, we see that on average the self-assessment of the "less attractive" individuals is more biased (average deviation=3.88, SD=1.96) than that of the "attractive" individuals (average deviation=0.69, SD=1.79). The difference is statistically significant (MW test, z=10.38, p<0.001). In the High Stakes Treatment only 5% of the participants evaluate themselves as less attractive. This percentage significantly differs from the expected 30% (test for proportions, z=-5.47, p<0.001). When evaluating themselves, attractive individuals deviate from their actual rank by 0.58 on average (SD=2.06), whereas for

the less attractive ones the deviation is higher with a value of 3.5 (SD=1.59). The difference is statistically significant (MW-test, $z=5.95$, $p<0.001$).

In the Judgment stage we analyze the behavior of participants matched with attractive versus less attractive counterparts and we see that as in the main analysis this deviation is closer to zero for the former group (average deviation=-0.086, SD=2.15, WS test: not different from zero, $z=-0.12$, $p=0.900$) than for the latter (average deviation=2.07, SD=1.86, WS test: different from zero, $z=4.25$, $p<0.001$). The difference in deviations between the two aforementioned groups is again statistically significant (MW test, $z=4.27$, $p<0.001$). Comparing how less attractive individuals assess themselves to how another same sex participant assesses them we find that in the latter case the deviation from the actual rank is significantly smaller (3.43 versus 2.07, MW test, $z= -2.78$, $p=0.005$). Thus, the results on the self-assessment stay robust to the alternative threshold.

Next we investigate whether there is any difference between subjects assessment of others in the Judgment and F2F feedback treatment. Again, the distribution of ranks provided by the participants matched with a less attractive individual significantly differs between Judgment and F2F Feedback ($\chi^2(8)=13.84$, $p=0.054$). While 16% of the participants in the Judgment treatment evaluated their counterpart as "less attractive," only 3% in the F2F Feedback treatment did so ($\chi^2(1)=9.83$, $p<0.002$).

The average deviation between the F2F feedback received and an individual's actual rank is 0.30 for the participants who were matched with an attractive individual (SD=2.02) as compared to -0.09 (SD=2.15) in the Judgment treatment (MW test, $z=-0.74$, $p=0.458$). When matched with a less attractive participant, the average deviation is significantly larger in the F2F feedback treatment than in the Judgment treatment (3.63, SD=1.83 vs. 2.07, SD=1.86) respectively (MW test, $z=-2.96$, $p=0.003$).

We confirm our finding with the probit estimation in Table C2. The estimates remain similar in size and the levels of significance are identical to the analysis conducted using a threshold of 7.

Table 3.A4 Likelihood of Providing Negative Feedback To The Less Attractive Participants (70th Percentile Split)

Probability (Negative feedback)	(1)	(2)	(3)	(4)
F2F Treatment	-.23***	-.23***	-.22***	-.21***
	(.06)	(.06)	(.07)	(.07)
High Stakes Treatment	-.27***	-.27***	-.25***	-.24***
	(.06)	(.06)	(.06)	(.07)
Anonymous Treatment	-.09	-.09	-.08	-.07
	(.08)	(.08)	(.08)	(.08)
Female		-.01	.02	.01
		(.07)	(.08)	(.08)
Age			-.01	-.02
			(.02)	(.02)
Ethnicity Control	N	N	N	Y
Observations	120	120	114	114

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Note: The table presents marginal effects estimated from Probit Regressions. Probability (Negative Feedback) is the predicted probability of giving negative feedback (ranks 8 to 10) in each of the three treatments compared to judging a less attractive individual negatively. The baseline treatment is the Judgment treatment. Female is a gender dummy and age is an ordinal variable. All marginal effects are evaluated at the change of the dummy variable from 0 to 1.

Next we investigate whether there are any differences in the updating results. For the attractive participants the deviation between the initial self-assessment and the self-assessment after feedback changes from 0.40 to 0.47 (WS test, z=-0.131, p=0.900) while for the less attractive individuals it changes from 4.11 to 4.37 (WS test, z=-0.29, p=0.772). In line with the results we find when using 7 as a threshold, both changes are not significant.

In the High Stakes treatment 3% provide feedback to their counterpart implying that the person is among the less attractive participants. Again, the difference in the distributions of ranks between the Judgment and the High Stakes treatment is driven by the feedback to the less-attractive individuals ($\chi^2(8)=17.29$, p=0.027).

For participants matched with attractive counterparts the average deviation between subjects' actual rank and the feedback that is provided to them is 0.23 (vs. 0.30 in F2F Feedback), while for those matched with the less attractive participants it is 3.73 (vs. 3.63 in F2F Feedback). We can confirm that the average deviation under this threshold is significantly larger in the High Stakes treatment than in the Judgment treatment (MW test, z=-3.28, p=0.001), while we do not find differences for participants matched with an attractive participant (MW test, z=-0.64, p=0.521). Table C2 shows that the probit results for the High Stakes treatment remain robust and significant with the alternative threshold.

Looking at the precision of individuals' self-assessment in the updating stage we see that for the attractive participants the deviation between individuals' self-assessment and their actual rank does not significantly change after feedback (0.58 to 0.67, z= -0.35, p=0.726). In contrast to the threshold of 7 and the threshold of 6, in this

specification the deviation does not become significantly larger after feedback for the less attractive individuals (3.5 to 3.73; z=-1.48, p=0.140). However, the direction of the change in the average deviation stays the same however.

As with a threshold of 7, significantly more individuals evaluated their partner as "less-attractive" in Anonymous compared to F2F (13% vs. 3%, $\chi^2(1)=6.90$, p=0.009). Comparing the distributions of evaluations separately for attractive and less attractive individuals like in the main analysis yields insignificant results ($\chi^2(1)=11.32$; 8.75, $p\geq0.184$).

The average deviation between the feedback provided in stage three and the less attractive participants' actual ranks is 3.63 in the F2F Feedback treatment, while it is only 2.87 in the Anonymous treatment (MW, z=1.53, p=0.126). Due to the lower number of observations this difference is marginally significant, which is in contrast with the significant difference that we find when using a threshold of 7. For the attractive individuals we again find no significant difference; the average deviation is 0.30 in the F2F Feedback and -0.28 in the Anonymous treatment (z=1.24, p=0.213). When we compare the precision of feedback in the Anonymous Feedback treatment to the precision of evaluations in the Judgment treatment, we find that the average deviation does not differ when participants are matched to attractive individuals (average deviation=-0.28 versus -0.09, MW test, z=0.538, p=0.591) and it marginally differs when participants are matched with the less attractive individuals (average deviation=2.87 versus 2.07, MW test, z=-1.687, p=0.092). These findings are in line with those based on a threshold of 7.

In the updating stage, the average difference between participants' self-evaluation and their actual rank decreases from 4.10 to 3.87 (WS test, z=1.22, p=0.221) for the less attractive participants, while for the attractive ones it decreases from 0.74 to 0.46 (WS test, z=2.71, p=0.007). While when defining attractiveness based on a threshold of 7 both groups had a significant improvement in their self-assessment, with a threshold at the 70^{th} percentile of the rank distribution only the attractive individuals significantly benefit from feedback. However, the direction of the differences is the same for less attractive individuals. Overall, changing the threshold of the classification between "attractive" and "less attractive" to either 6 or 8 has no considerable effect on our main results.

Appendix B. Instructions

Instructions Judgment treatment

Your ID code is FA

Welcome and thank you for participating in our study!

This experiment will take about 15 minutes. If you read the instructions carefully you can earn a considerable amount of money depending on your decisions and the decisions made by other participants. The money will be paid to you privately and in cash in class next week. **Importantly, any choice you make in this experiment will be treated confidentially.** Other participants in the experiment will never know what decisions you made. Participation will not affect your grading in any class at Rady.

You have been assigned to a group of 10 participants. The members of your group, **Group F**, have IDs FA, FB, FC,…,FJ.

Please do not turn the page until instructed by the experimenter.

------- page break --------

Stage 1

In the group in front of you there are also 10 participants; they are called **Group M**. Each of the 10 participants in Group M is wearing a sticker that indicates that person's ID (MA, MB, MC,…, MJ).

Your first task in this experiment is to rank the 10 members of Group M according to attractiveness.

The most attractive person should get a ranking of 1, the second most attractive a ranking of 2, and so on—rank 10 should be assigned to the least attractive person in that group. Please note that each of the 10 members of the other group should get a different number.

We ask all the members in your group to do the same ranking task. Based on all the rankings provided by your group, we will create an aggregate ranking (1-10) of **Group M**. We will pay you $10 if your ranking of at least 5 members of Group M matches the aggregate ranking. (In the unlikely case that in the overall ranking two individuals are equally ranked, we will randomly determine which of them is assigned the higher rank.) To mark your choice, please write the participant's ID next to the number ranking below. It is important to note that your ranking and the aggregate ranking will remain totally confidential, and will never be revealed to any of the

participants. Please make sure that no one else in the room will be able to see your ranking.

Ranking
1____
2____
3____
4____
5____
6____
7____
8____
9____
10___

Please do not turn the page until instructed to do so.

------- page break --------

Stage 2

In Stage 1 we asked the members of Group M to rank the members of your group in the same way you did the ranking for them. We will use this ranking to compute an aggregate ranking for your group (1-10).

Please indicate below your guess about your rank in this aggregate ranking**.** That is, we ask you to guess how attractive the members of Group M said you are relative to the other members of your group. Please remember that 1 is the most attractive person in the group and 10 the least attractive.

If your guess about your own rank matches your actual position in the aggregate ranking made by Group M, you will receive 10 dollars.

My guess about my rank in the aggregate ranking made by Group M (1-10): _____

Please turn the page and continue with stage 3.

------- page break --------

Stage 3

This is the third stage of the experiment. We ask you, in a similar way to your guess in stage 2, to guess the ranking of participant FB in your group.

As in stage 2 above, we ask you to guess how attractive the members of Group M said participant FB is relative to the other members of your group. Please remember that 1 is the most attractive person in the group and 10 the least attractive.

If your guess about the rank of FB matches the actual position of this participant in the aggregate ranking made by Group M, you will receive 10 dollars.

My guess about participant FB's position in the aggregate ranking (1-10):

_____

Please do not turn the page until instructed to do so.

------- page break --------

Stage 4

We now ask you to do a complete ranking of your own Group F.

We will pay you $10 if your ranking of at least 5 members of Group F matches the aggregate ranking provided by Group M. (In the unlikely case that in the overall ranking two individuals are equally ranked, we will randomly determine which of them is assigned the higher rank.) To mark your choice, please write the participant's ID next to the number ranking below. It is important to note that your ranking and the aggregate ranking will remain totally confidential, and will never be revealed to any of the participants. Please make sure that no one else in the room will be able to see your ranking.

Ranking
1____
2____
3____
4____
5____
6____
7____
8____
9____
10___

Please wait until an experimenter comes to pick up your sheet.

Thank you for participating in our experiment! We will notify you via e-mail about your earnings and inform you where to pick them up.

Instructions F2F Feedback

Your ID code is FA

Welcome and thank you for participating in our study!

This experiment will take about 20 minutes. If you read the instructions carefully you can earn a considerable amount of money depending on your decisions and the decisions made by other participants. The money will be paid to you privately and in cash. **Importantly, if not indicated otherwise, any choice you make in this experiment will be treated confidentially.**

You have been assigned to a group of 10 participants. The members of your group, **Group F**, have IDs FA, FB, FC,…,FJ.

Please do not turn the page until instructed by the experimenter.

------- page break --------

Stage 1

In the group in front of you there are also 10 participants; they are called **Group M**. Each of the 10 participants in Group M is wearing a sticker that indicates that person's ID (MA, MB, MC,…, MJ).

Your first task in this experiment is to rank the 10 members of **Group M** according to attractiveness.

The most attractive person should get a ranking of 1, the second most attractive a ranking of 2, and so on—rank 10 should be assigned to the least attractive person in that group. Please note that each of the 10 members of the other group should get a different number.

We ask all the members in your group to do the same ranking task. Based on all the rankings provided by your group, we will create an aggregate ranking (1-10) of **Group M**. We will pay you $10 if your ranking of at least 5 members of Group M matches the aggregate ranking. (In the unlikely case that in the overall ranking two individuals are equally ranked, we will randomly determine which of them is assigned the higher rank.) To mark your choice, please write the participant's ID next to the number ranking below. It is important to note that your ranking and the aggregate ranking will remain totally confidential, and will never be revealed to any of the participants. Please make sure that no one else in the room will be able to see your ranking.

Ranking

1____
2____
3____
4____
5____
6____
7____
8____

9____
10___

Please do not turn the page until instructed to do so.

------- page break --------

Stage 2

In Stage 1 we asked the members of Group M to rank the members of your group in the same way you did the ranking for them. We will use this ranking to compute an aggregate ranking for your group (1-10).

Please indicate below your guess about your rank in this aggregate ranking**.** That is, we ask you to guess how attractive the members of Group M said you are relative to the other members of your group. Please remember that 1 is the most attractive person in the group and 10 the least attractive.

If your guess about your own rank matches your actual position in the aggregate ranking made by Group M, you will receive $10.

My guess about my rank in the aggregate ranking made by Group M (1-10):

_____

Please turn the page and continue with stage 3.

------- page break --------

Stage 3

In this stage you are matched with participant **FB** from your own group F. We now ask you to send participant **FB** a message with your guess about her rank in the aggregate ranking made by group M. Participant **FB** will know that the message is sent by you.

After receiving your message, **FB** will have the opportunity to update her guess about her own position in the ranking.

If **FB**'s updated guess about her rank will match her actual rank both you and participant **FB** will receive $10. If **FB**'s guess does not match her actual rank, you will both receive $0 for this part.

Please indicate your guess about participant **FB**'s rank in the aggregate ranking made by Group M on the sheet of paper that was given to you.

Please remember that 1 is the most attractive person in your group and 10 the least attractive.

Please raise your hand when you are ready. Your message will be collected by the experimenter and will be given to participant **FB**.

------- page break --------

Stage 4

Just like the guess you just provided, participant **FJ** was asked to send you a message indicating her guess about your rank in the aggregate ranking.

You now have the option to update the guess you provided in stage 2 about your own rank in the aggregate ranking. Please remember that you will earn $10 if you guess your own rank correctly in this experiment. If you change your mind regarding your ranking, and now give a different number than the one you gave in Stage 2, we will only use this new number.

My final guess about my rank in the aggregate ranking made by Group M (1-10):

------- page break --------

Gender: _____
Ethnicity: _____
Age:_____

Please wait until an experimenter comes to pick up your sheet.

Thank you for participating in our experiment! We will notify you via e-mail about your earnings and inform you where to pick them up.

Message sheet F2F Feedback (10$) and High Stakes (50$) treatment

My guess about participant **FB**'s position in the aggregate ranking (1-10): ___

On the outside of the folded sheet it said: From FA to FB

References

Balafoutas L, Kerschbamer R, Sutter M (2012). Distributional preferences and competitive behaviour. *Journal of Economic Behavior and Organizations* 83(1): 125-135.

Burks SV, Carpenter JP, Goette L, Rustichini A (2013). Overconfidence and social signaling. *Review of Economic Studies* 80(3): 949-983.

Charness G, Dufwenberg M (2006). Promises and Partnership. *Econometrica* 74(6): 1579–1601.

Crawford V, Sobel J (1982). Strategic Information Transmission, Econometrica, 50(6): 1431-1451.

Dreber A, Johannesson M (2008). Gender Differences in Deception. *Economic Letters* 99 (1): 197-199.

Eil D, Rao JM (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics,* 3(2): 114-138.

Erat S, Gneezy U (2012). White lies. *Management Science* 58(4): 723-733.

Gneezy U. (2005). Deception: The role of consequences. *American Economic Review,* 95(1): 384-394.

Freud S. (1991). *On Metapsychology* (PFL 11) p. 454-5.

Levitt S., Snyder C. (1997). Is No News Bad News? Information Transmission and the Role of "Early Warning" in the Principal-Agent Model. *Rand Journal of Economics,* 28(4): 641-661.

Moore D., Healy P. (2008). The trouble with overconfidence. *Psychological Review,* 115(2): 502–517.

Morris S. (2001). Political correctness. *Journal of Political Economics*, 109(2): 231-265.

Olszewski W. (2004). Informal communication. *Journal of Economic Theory,* 117(2): 180-200.

Ottaviani M., Sørensen P., (2005a). Professional advice. *Journal of Economic Theory,* 126(1): 120-142.

Ottaviani M, Sørensen P (2005b). Reputational cheap talk. *Rand Journal of Economics* 37(1): 155-175.

Prendergast C (1993). A theory of "yes men". *American Economics Review* 83(4): 757-770**.** Solnick S, Schweitzer M (1999) The Influence of Physical Attractiveness and Gender on Ultimatum Game Decisions. *Organizational Behavior and Human Decision Process*, 79(3):199-215.

Sutter M (2009). Deception through telling the truth? Experimental evidence from individuals and teams, *The Economic Journal*, 119(534): 47-60.

4. Discrimination in Disguise

Abstract

We present evidence of prejudice-based ethnic discrimination. We show that discrimination is absent in contexts in which individuals cannot justify it but occurs in contexts in which it can be disguised behind conformity to social or moral norms. In a binary dictator game where discriminative acts cannot be justified, we find no evidence of discrimination. Subjects are equally likely to choose prosocial payoff allocations for individuals of their own or of a different ethnicity. However, in an experiment in which the prosocial payoff allocation can be obtained by telling an altruistic lie, subjects are predominantly honest only when an individual of a different ethnicity benefits from the lie. Hence, they disguise discrimination behind adherence to the norm of honesty. Further, we show that subjects disguise discrimination behind endorsement of a fairness norm in an ultimatum game. Responders demand higher offers from individuals of a different ethnicity as opposed to their own. We conclude that taste-based discrimination is still present in modern societies. Although it is suppressed in contexts in which it cannot be disguised, discrimination appears when it can be rationalized as compliance to a virtuous norm.

4.1 Introduction

Instances of hostility, bigotry and injustice on the grounds of race, ethnicity, gender or religion have been ubiquitous over the course of human history. Behaviors aimed at establishing the superiority of a particular group were formerly not only commonly accepted and embedded in everyday behavior, but also legally institutionalized in many societies. Some remarkable examples are the segregation of African Americans in the United States following centuries of legalized slavery, the anti-Semitic movements in Europe and the US, and the denial of political and social rights to women. Often such acts originated from prejudice, i.e., feelings of hatred or dislike of the targeted group. These types of behaviors arising from preference-based prejudice are known in the economic literature as "taste-based" discrimination (Becker, 1957).

Over the second half of the Twentieth Century, the Western world has made substantial advances in eliminating institutional, sanctioned discrimination. With the progress achieved by civil rights movements, modern democracies began to embrace principles of justice and egalitarianism as well as to enforce various legislation targeting discrimination (Pinker, 2002). Today, Western societies are more multicultural, diverse and tolerant than ever before. Citizens are granted equal legal rights regardless of their gender or ethnicity. Open expression of prejudice has drastically declined (Mdon, 2001; Devine and Elliot, 1995). Society has become more favorable to interracial marriage and more supportive of affirmative policies in favor

of minorities.[1] These encouraging observations suggest that instances of taste-based discrimination should now occur less often.

Yet economic inequalities on the grounds of race, gender and ethnicity still persist in modern societies (Adida, Laitin and Valfort, 2010; Fryer and Levitt, 2004). Racial gaps characterize labor, credit and consumer markets, contributing to disparities in earnings and overall wealth (Ayres, Banaji, Jolls, 2011; Blanchflowr, Levine and Zimmermann, 2003; Carlsson and Rooth, 2007). For example, in OECD countries, educated foreigners are less likely to be employed than their native-born counterparts (77% versus 84% as of 2013). Further, when employed they are twice as likely to be over-qualified.[2] Similarly, whereas gender differences in level of education are disappearing, women still remain a minority in top corporate positions (Bertrand, 2009; Wolfers, 2006). Although historical factors and individuals' preferences might also contribute to such inequalities, empirical studies suggest that discrimination persists in many economic domains (Bertrand and Mullainathan, 2004; Fong and Luttmer, 2011; Kaas and Manger, 2012). However, recent experimental research contends that most of today's discrimination is a consequence of profit-maximization motives and stereotypical expectations about average behavior of the discriminated group rather than prejudice (Ewens, Tomlin and Wang, 2013; Gneezy,

---

[1] See also Gallup (2013) Poll 163687, June 13 - July 5. Available at: http://www.gallup.com/poll/163697/approve-marriage-blacks-whites.aspx

[2] OECD (2014) *International Migration Outlook 2014* (OECD Publishing) Available at: http://www.oecd-ilibrary.org/social-issues-migration-health/international-migration-outlook-2014_migr_outlook-2014-en [Accessed May 13, 2015].

List and Price, 2012; List, 2004). This type of discrimination is also known as statistical discrimination (Arrow, 1998; Phelps, 1972).

In this paper we show that prejudice remains a significant driver of discrimination, but only when it can be disguised. In a series of incentivized laboratory experiments, we show that individuals do not discriminate when behavior can be easily attributed to prejudice or feelings of dislike. However, when individuals can disguise it behind adherence to certain norms or values, discrimination emerges. In particular, we find that individuals endorse certain norms to a greater extent when doing so allows them to discriminate. Our results are in line with research in social psychology suggesting that changes in modern societies have made prejudice increasingly subtle (Dovidio and Gaertner, 1998; Devine, 1989; Crosby, Bromley and Saxe, 1980). Over the past decades, the social norms that favored open expression of prejudice have evolved into new pervasive norms of political correctness that condemn any verbal or behavioral expression of hatred (Fiske, 1998). This societal change may have not eradicated prejudice but instead driven individuals to disguise it. If this is the case, taste-based discrimination should be limited to contexts where it cannot be recognized as such, for example, in the presence of ambiguity (McConahay, 1986) or conflicting social norms. We add to this literature by demonstrating that prejudice is most likely to affect behavior in settings with conflicting social norms by showing that adherence to norms and values can provide individuals with a way to disguise taste-based discrimination.

Importantly, our experiments consider non-strategic environments such that any disparate treatment can be explained only by expression of prejudice (taste-based

discrimination), isolating it from the payoff-maximizing concerns that govern statistical discrimination. In many cases, statistical and taste-based discrimination result in identical observable outcomes, and thus understanding whether prejudice is a determinant of observed economic disparities is nontrivial and hard to observe in empirical data. However, identifying the nature of discrimination is necessary for designing effective policy interventions. Thus far, a typical experimental design used to isolate taste-based from statistical discrimination is the dictator game. Several experiments based on dictator games have found that giving behavior is not affected by the receiver's group membership, concluding that discrimination in markets is largerly free from prejudice (Castillo, Petrie, Torero and Vesterlund, 2013; Fershtman and Gneezy, 2001; List, 2004). Our results suggest the dictator game might not be the appropriate method to capture taste-based discrimination, as it does not leave room to disguise the expressed prejudice.

In our experiments, we matched subjects with individuals of either their own or of a different ethnicity. We then asked them to make choices that affected both their payoffs and the payoffs of their respective counterparts. In a dictator game in which subjects were asked to directly choose the payoffs for both participants, we find that subjects are equally prosocial toward individuals of their own and different ethnicities. This was expected given that subjects could not disguise discrimination in this setting.

In two additional experiments, we consider settings in which taste-based discrimination can be disguised. In one experiment, instead of requiring a simple direct choice of payoffs as in the dictator game, a prosocial outcome entails telling a white lie – a violation of a moral norm. We find a strikingly different pattern of

behavior than that observed in the dictator game. When the lie benefits a member of a different ethnicity, individuals are significantly less likely to tell a white lie then they are when lying helps an individual of their own ethnicity. These results provide evidence of taste-based discrimination disguised behind honesty. In another experiment, we investigate fairness attitudes of responders in an ultimatum game. Similar to moral norms of honesty, norms of fairness can be used to disguise taste-based discrimination. Indeed, we find that receivers' preferences regarding the minimum acceptable split of a total pie depend on the ethnicity of their counterparts. While individuals request high minimum offers from counterparts of a different ethnic group, the requested minimum offers are substantially lower when interacting with their own ethnicity. Again, we find evidence for taste-based discrimination, which in this case is disguised behind fairness.

4.2 The Experiments

Our empirical investigation was carried out in Germany and tested how German subjects behave towards individuals with native-sounding names as opposed to individuals with Turkish- or Arabic-sounding names. Turks are the largest ethnic minority in Germany. In spite of more than 50 years of living in Germany, their integration is still quite poor. They are more likely to be unemployed (Kaas and Manger, 2012), less likely to succeed in school (Ross, 2009), and more likely to have lower income than native Germans (Bruder, Neuberger and Räthke- Döppner, 2011).

. While these statistics suggest discrimination might contribute to the economic gap between Turkish and German natives, it is not clear whether prejudice toward Turks is part of this problem. Further, we also chose to focus on this minority due to cultural, etymological, and religious differences. Turkish names also sound very distinct from common German names.

We invited male subjects (N=329) with German sounding names from the University of Cologne to participate in our experiments in exchange for €2.50 and assigned them to the role of the decision-maker (DM). We informed them that the decision of one out of every 10 randomly determined participants would be implemented, resulting in additional earnings for himself and another individual – the receiver. Another 32 subjects were recruited to the study and assigned to the role of the receiver. In half of the cases, receivers had German-sounding first names; in the other half they had Turkish-sounding names (see SI Materials and Methods for more details). A manipulation check confirmed that individuals from the same subject pool consider the Germans-sounding names more likely to belong to people of German ethnicity than the Turkish-sounding names (see SI Material and Methods).

4.2.1 A Dictator Game.

DMs (N=83) played a binary dictator game. They were randomly matched with one of eight receivers. Receivers' first names were either German (named Bernd, Dirk, Ingo, Johannes) or Turkish (named Baris, Emrah, Ismail, and Mustafa). Both the DM and the receiver were informed about the first name of their matching partner (See SI for more details). We then asked DMs to determine their own earnings and the

earnings of the receiver by choosing one of two payoff allocations. Option 1 resulted in a payment of €10 for both players; Option 2 resulted in an altruistic and more efficient payoff allocation of €9 to the DM and €15 to the receiver. DMs also knew that receivers were not informed about the payoffs associated with either option and made no decisions.

We study prosocial behavior toward the receiver as expressed by DMs' willingness to sacrifice one euro in order to give their counterpart five extra euros. In this context, a discriminatory behavior cannot be attributed to factors other then prejudice. We, therefore, expected similar average rates of prosocial choices (Option 2) regardless of the ethnicity of the receiver. Further, given that in this context strategic considerations are not present, any disparity in behavior towards German or Turkish receivers can be explained by taste-based discrimination.

4.2.2 Results

In line with our hypothesis and previous findings in dictator games, the rate of prosocial and welfare-maximizing choices was indeed almost identical between the German Receiver (38.1%) and Turkish Receiver (36.6%) treatments (test of proportions: $Z=0.14$, $p=0.8869$, see Figure 1 panel A). [3]

Is prejudice toward Turks not present or does conformity to anti-discriminative norms prevent individuals from expressing it? To answer this question, we compare the results of our dictator game to the results of the next experiment, in which the altruistic and efficient payoff can only be reached by telling a lie.

[3] All reported tests are two-sided.

4.2.3 A Prosocial Lies Experiment

In the prosocial lies experiment, we designed a decision situation with two conflicting norms. DMs (N=83) could reach the same altruistic payoff allocation of the dictator game by sending a false message –a lie – to the receiver. Previous research has shown that engaging in unethical behavior results in moral costs to individuals' self-image (Gneezy, 2005; Fischbacher and Follmi-Heusi, 2013; Mazar, Amir and Ariely, 2008). All else equal, people who experience such costs prefer an outcome that is reached without telling a lie to an outcome reached through dishonesty. The possibility of appearing to comply with the norm of honesty might provide DMs with an excuse not to favor the altruistic payoff allocation. Importantly, we hypothesized that individuals would selectively choose whether to favor compliance to honesty over prosocial behavior depending on the ethnicity of the receiver. While DMs might be willing to bear a moral cost in order to be prosocial when matched with receivers from their own ethnicity, they might not be willing to do something unethical to favor an ethnic minority. In this context, lack of prosocial lying toward a Turkish counterpart can be ascribed to something other than prejudice, a commitment to honest behavior, which allows individuals to disguise discrimination.

DMs with German first names who did not participate in the dictator game (N=83) took part in this experiment. We randomly assigned half of the subjects to one out of four receivers with a German first name and the other half to one out of four receivers with a Turkish first name. The receivers invited to this experiment had the same first names as those who took part in the dictator game. We kept the monetary incentives identical to those of the dictator game. Instead of making a choice about

payoffs, subjects played a deception game (Erat and Gneezy, 2012; Gneezy, 2005). DMs were informed about the outcome of a dice roll and asked to communicate this to their receiver via a message ("The outcome of the dice roll is ___", where X was a number between one and six). Here the DMs could communicate any possible dice roll number (either true or not).

After the receiver saw the message, he was asked to guess the actual dice roll outcome. If his guess was correct, both players were paid €10. If the guess was incorrect, the DM received €9 and the receiver got €15. Importantly, the DMs knew that receivers were not informed about the actual outcome of the dice roll and did not know the payoffs associated with either option (see SI Appendix for full instructions). Therefore, assuming the receiver follows the message, DMs could increase the receivers' payoffs by sending a false message.[4]

4.2.4 Results

Are individuals willing to lie and sacrifice one euro to benefit the receiver? In line with our prediction, a substantial fraction of DMs who were matched with the German receivers lied to benefit their counterpart (23.8%). However, when matched with Turkish receiver, the percentage of subjects who lied dropped to 7.3%. The difference between the two treatments is statistically significant at the 5% level (test of proportions: Z=2.07, p=0.0387). This finding is in contrast with the results of the dictator game (see Figure 1 for comparison), where participants did not display any disparate behavior between treatments. A regression analysis confirms that individuals

[4] Note that in similar experiments the majority of receivers in this game follow the message.

were less likely to act prosocially in the "Prosocial Lies Turkish Receiver" case than in any of the other cases (see SI Material and Methods).

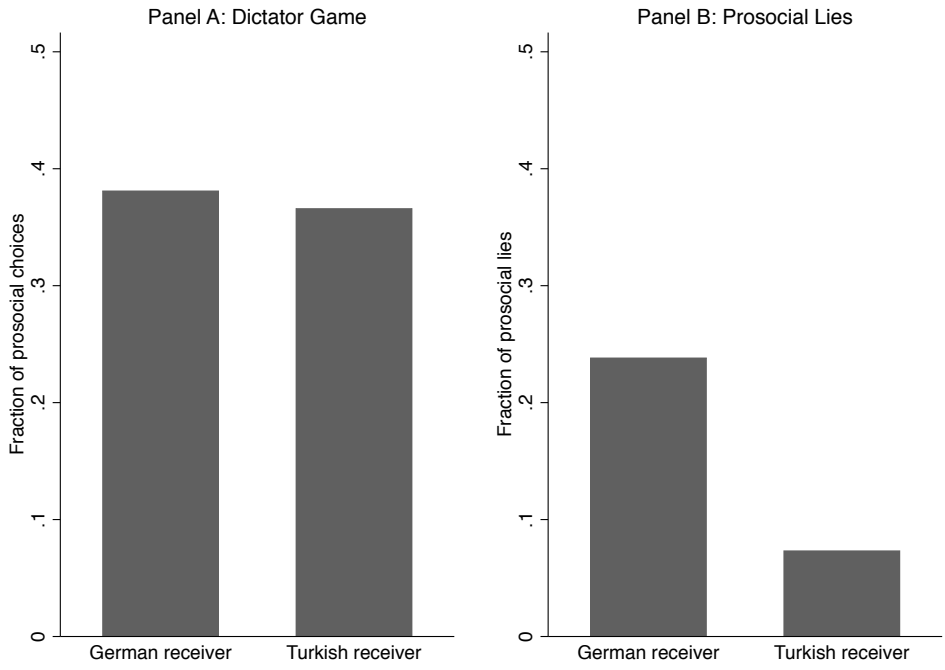

Figure 4.1 Fraction of Prosocial Choices

Note: Prosocial choices in the dictator game (Panel A) and prosocial lies experiment (Panel B): The fraction of prosocial decisions doesn't depend on the ethnicity of the receiver in the dictator game. However, in the prosocial lies experiment, significantly more subjects engage in prosocial lying when they were matched with a German receiver as compared to the treatment with Turkish receiver. The sample size was as following: Dictator game "German receiver" condition: N=42; Dictator game "Turkish receiver" condition: N=41; Prosocial lies experiment "German receiver" condition: N=42; Prosocial lies experiment "Turkish receiver" condition: N=41.

There are two alternative explanations for the disparity in behavior towards German and Turkish receivers. First, DMs' beliefs about Germans' and Turks' reaction to the message could differ. For example, DMs' prior beliefs could be that Turkish participants will be less likely to follow the message. Therefore, DMs may strategically tell the truth to increase the other party's payoff (Sutter, 2009). Second,

the individual moral cost of lying may vary with the ethnicity of the receiver. Individuals may consider lying to someone of the same ethnic group as appropriate but may feel bad about lying to a person of a foreign origin. Both these arguments could have led to more honesty towards Turkish receivers in the prosocial lies experiment.

To rule out these two alternative explanations, we ran an additional experiment in which DMs (N=82) could forgo monetary profits by sending a truthful message. In this experiment, an incorrect guess by the receiver resulted in €15 for the sender and €9 for the receiver. A correct guess led to a payment of €10 for each subject. Hence, in this context, honest behavior by the DM decreased his payoff by €5 and increased the receiver's payoff by €1. Given that in this game the norm of honesty and altruistic behavior result in the favorable outcome for the receiver, discrimination cannot be disguised behind honesty. Therefore, we did not expect to observe any effect of the ethnicity of receivers. Note that, considering the payoff structure, we expected the overall honesty rates in this game to be lower than in the prosocial lies game.

In line with our predictions, in this treatment only 26.2% of the participants matched with Turks were honest, while the rest of the subjects lied to maximize their payoffs. This fraction is lower but not statistically different from the fraction of honest subjects matched with Germans (37.5%, test of proportions: $Z=1.10$, $p=0.2713$). In both cases, the honesty rate is significantly lower than in the prosocial lies experiment (test of proportions, $p<0.001$). This rules out the alternative explanations mentioned above, showing that subjects did not tell the truth more often when matched with Turks. Thus, the differential behavior we observe in the prosocial lies experiment

cannot be explained by differences in expectations about the likelihood that the message will be followed or by differences in lying costs towards the minority.

Taken together, the three experiments provide clear evidence that taste-based discrimination occurs, but only in a context where it can be disguised. The results suggest that although prejudices still affect behavior, they come to light only in contexts in which discrimination cannot be directly interpreted as such.

Two further questions remain: Are individuals willing to forgo profit to satisfy their prejudice? Is the observed effect limited to moral norms of honesty, or does it extend to other types of norms? To answer these questions, we investigate whether individuals disguise a prejudiced taste behind the norm of fairness in negotiations.

4.3 A Bargaining Experiment

In the next experiment, we study responders' behavior in an ultimatum game (Guth, Schimttberger, and Schwarze, 1982). In this game, a proposer is given an endowment and asked to propose how to divide it between himself and an unknown responder. The responder decides whether to accept or reject the proposal. In case of acceptance, the money is split accordingly. If the responder rejects, both players receive zero. A large body of research has shown that responders tend to reject low positive offers, considering them as "unfair" (Kahneman, Knetsch and Thaler, 1986; Ochs and Roth, 1989; Thaler, 1988). In ultimatum games, rejections can be seen as a punishment for proposers who offered an unfair split (Camerer and Thaler, 1995). However, individual standards about what constitutes a fair offer may vary. In the literature, disagreements still exist on what drives individuals' fairness views given

that they seem to be malleable, context dependent and subject to self-serving biases (Babcock and Loewenstein, 1997; Konow, 2000; Konow, 2003). As a result, in our context responders may self-servingly use the proposer's (lack of) fairness as an excuse to discriminate toward a different ethnicity (by imposing punishment more often than toward proposers of their own ethnicity).

Our experiment builds on evidence that individuals adopt fairness arguments to justify behavior. We therefore hypothesized that in a bargaining context individuals would be more likely to reject unfair offers from Turks than from Germans. We focus on the behavior of the ultimatum game responders (N=81). We manipulate the ethnicity of proposers and use experimental procedures in the same way as in the previous experiments. We randomly matched responders with one of eight proposers. Again, four subjects in the role of proposers had Turkish sounding names (Ali, Huseyin, Ismail and Murat), and the other four subjects had German sounding names (Andreas, Dirk, Florian and Tobias). We informed responders that the proposer was endowed with €20 and had to offer how to split this amount between herself and the responder. Responders had to indicate the minimum offer they were willing to accept. Both responder and proposer made their decisions before knowing the decision of the partner. Any offer that was equal to or greater than the amount indicated by the responder was implemented. Conversely, any offer below the threshold was automatically rejected, resulting in zero earnings for both parties.

In line with previous findings and norms of fairness, we hypothesized that responders would reject low offers from proposers. However, on top of that we predicted the acceptance rate of low offers to depend on the name origin of the

proposer, with subjects requesting more money from Turkish proposers. Such requests cannot be explained by any strategic considerations or payoff-maximizing beliefs. Thus, when observed, it allows us to disentangle taste-based from statistical discrimination.

4.3.2 Results.

We find that, on average, individuals reject offers below €7.15 (Median=8, SD=3.85). However, when matched with Germans proposers, on average responders were willing to accept offers that were 21.6% lower than when matched with Turkish proposers (Mean=6.28, Median=6.45, SD=3.94 from Germans vs. Mean=8, Median=8, SD=3.61 from Turks). The difference between the two distributions is statistically significant (MWU test: U=2.26, p=0.0239). As shown in Figure 2, the difference is driven by a larger fraction of participants requesting a minimum offer of €10 – the equal split – in the Turkish treatment. Indeed, while 80% of the subjects matched with Germans demanded less than €10, only 58.5% of those matched with Turks did so (test of proportions: Z=-2.09, p=0.0366).

To test whether this result is driven by different expectations about proposer behavior of the two ethnicities, we surveyed another 53 male subjects regarding their beliefs about the amount proposers with one of the names used in the experiment would offer. We find no difference in the beliefs (Mean=7.28, Median=8, SD=2.89, N=25 from German proposers vs. Mean=7.82, Median=9, SD=3.62, N=28 from Turkish proposers, MWU test: U=-0.476, p=0.669). Further, we find that the large

majority of the subjects believe that in Germany native Germans have higher average income than the population with a Turkish background (71.7%), suggesting that beliefs about subjects' income also cannot explain the effect we observe.

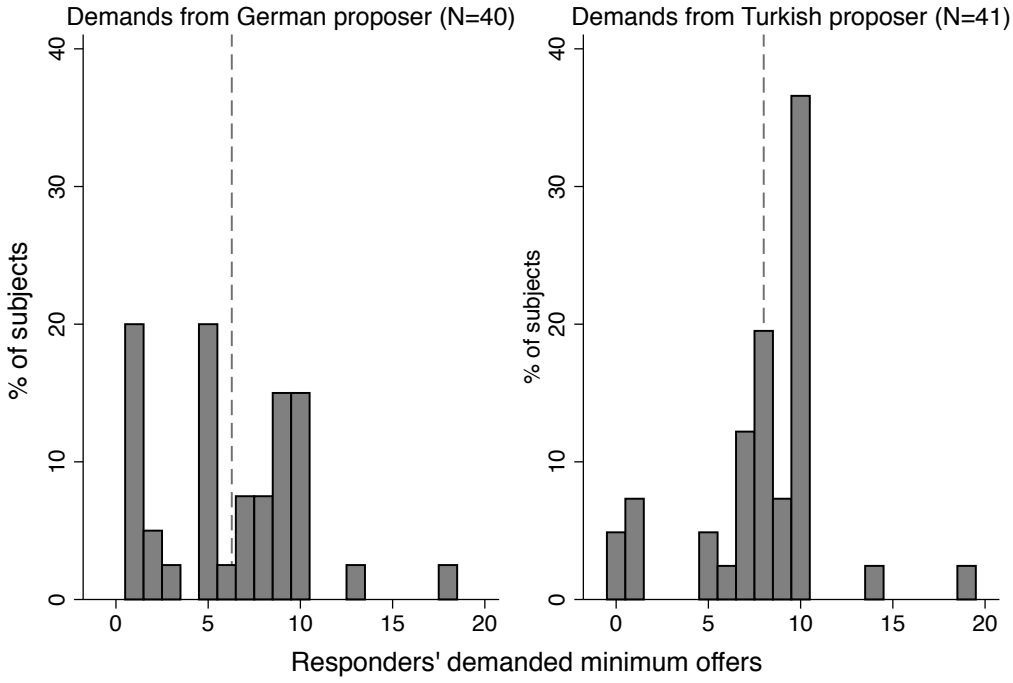

Figure 4.2 Distribution of Responders' Demanded Minimum Offers.

This result suggests provide evidence for taste-based discrimination. Subjects matched with Turkish counterparts were more willing to forgo positive profits in order to punish the "unfair" behavior of proposers. In term of monetary payoffs, this type of discrimination harms both the responder and the proposer, suggesting individuals experience more disutility from the disparity in allocation between themselves and a Turkish counterpart than from a disparity in allocation between themselves and a German counterpart. Similar behavior in the real world could lead to a bias toward

minorities in take-it-or-leave-it negotiations, with considerably inefficient economic consequences.

4.4 Conclusion

Understanding the nature of discrimination can help policy makers to design interventions aimed at fighting inequalities in markets. For this purpose, it is important to understand whether taste-based discrimination still contributes to today's inequalities. By incorporating insights from economics and social psychology, we provide evidence for subtle, disguised taste-based discrimination. When behavior can easily be ascribed to prejudice, discrimination is less likely to emerge, as the strong anti-discrimination norms and legislation that govern behavior in today's societies prevent individuals from openly expressing those attitudes. However, taste-based discrimination does emerge when it can be attributed to other factors.

In the context of prosocial behavior, individuals selectively adhere to the norm of honesty to abstain from being prosocial toward a minority counterpart but are less concerned about this norm when the prosocial act benefits somebody of their own ethnicity. In a bargaining context, they embrace different fairness standards depending on the ethnicity of their counterpart. Individuals require higher offers from minority proposers than from proposers of their ethnicity, even though they are aware that the minority might have a lower income. Taken together, these findings support the hypothesis that individuals tend to disguise taste-based discrimination.

Our findings inform the debate around the economic theories of discrimination. First, they suggest that failure to detect taste-based discrimination in previous literature could have been an artifact of the particular research design used. Second, they suggest that individuals might disguise discrimination not only behind norms of honesty or fairness, but also behind selective adherence to law, regulations or other principles. Given the complexity and numerous norms that characterize today's markets, individuals might have great scope for disguising taste-based discrimination behind an excessive endorsement of rules, regulations or of certain norms. Consider, for example, the case of white US policemen's behavior toward African American. In the US, African Americans are more likely to be stopped and searched by the police, more likely to receive citations and to be arrested after being stopped (Gelman, Fagan and Kiss, 2007). The over-targeting of African Americans by police officers sometimes results in unnecessary shootings.

In a 2014 speech condemning the rise of such episodes of violence, US President Barack Obama stated: "We have made enormous progress in race relations over the past several decades. But there are still problems, and communities of color aren't just making these problems up. Separating that from this decision, there are issues in which the law too often feels as if it is being applied in a discriminatory fashion."

Nevertheless, while multiple factors can contribute to this problem, it also is possible that such behavior is a result of prejudice disguised behind laws and statistics regarding average African-American behavior. Similar to participants of our experiments, officials may disguise prejudice toward African-Americans behind over-

compliance to justice and fire weapons more often towards ethnic minorities. Future research should explore this possibility.

From a policy perspective, reducing the complexity of an environment and minimizing the prevalence of conflicting norms in markets could prove successful in limiting the scope for taste-based discrimination. Preventing individuals from hiding their prejudice behind other forms of "appropriate" behavior could help to reduce inequalities in society.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Danilov, Anastasia and Silvia Saccardo, "Discrimination in Disguise". The dissertation author was the co-primary investigator and author of this paper.

Appendix A. Procedures and Additional Analyses

Treatment Manipulation. All participants learned the first name of the person they were matched with. All subjects at the focus of this study were males, German native speakers with typical German names. They were referred as Participants A in the instructions. The first name origin of experimental opponents (referred as Participant B in the instructions) varied between experimental conditions. In the German condition (GER), Participants B had common German male names. In the Turkish condition (TUR), Participants B had Turkish-sounding male names that signaled their Turkish ethnic background.

Pronounced residential ethnic segregation as a result of gaps in socioeconomic status hampers integration (De Groot and Sager, 2010; Schelling, 1969). This narrows down interaction between the German and Turkish population, and hence, many prejudices and stereotypes persist (Kuhnel, Leibold and Mays, 2013).

Since Turks in Germany marry mostly within their own ethnicity, the vast majority of children bear Turkish names, even in the second and third generations. Due to cultural, etymological, and religious differences, these names sound very distinct from common German names. Turkish origin, even of those who were born in Germany or have German citizenship, can therefore be easily detected from their names. The detrimental effect of ethnically differently sounding names has also been used in other studies of discrimination (Bertrand and Mullainathan, 2004; Kaas and Manger, 2012).

Manipulation Check. To ensure the validity of the GER/TUR manipulation, we ran a survey that confirmed that the first names used in the TUR condition are perceived as being less likely to belong to a German than names used in GER. A group of 220 individuals from the subject pool of the Cologne Laboratory for Experimental Research was presented with one of the two lists of 10 first names sorted in a random order. Subjects were asked how likely is it that a person with a particular first name is of German origin. For each name we had a total of 110 answers. Each answer was given on a 5-points Likert scale with 1 = “definitely not” to 5 = “for sure” with an additional option “do not know”. See Table S1 for average name evaluations. An ordered probit regression of the estimated likelihood on the dummy of the name being used in the TUR conditions with standard errors clustered on individual levels provides a significant negative correlation (β = -2.74, z=-16.61, $p < 0.001$).

Table 4.A1 Name evaluation survey

Condition	Name	Used in the experiment	Mean estimated likelihood of a name to belong to a person of German origin (1 = "definitely not" to 5 = "for sure")	Number of evaluations ("Don't know" answers)
GER	Andreas	UG	4.24	108 (2)
	Bernd	DG, PL, SL	4.50	109 (1)
	Dirk	DG, PL, SL, UG	4.44	109 (1)
	Florian	UG	4.31	108 (2)
	Ingo	DG, PL, SL	4.20	109 (1)
	Johannes	DG, PL, SL	4.25	108 (2)
	Tobias	UG	4.22	108 (2)
TUR	Ali	UG	2.44	107 (3)
	Baris	DG, PL, SL	2.35	103 (7)
	Huseyin	UG	2.08	108 (2)
	Emrah	DG, PL, SL	1.93	108 (2)
	Ismail	DG, PL, SL, UG	2.28	108 (2)
	Murat	UG	2.29	108 (2)
	Mustafa	DG, PL, SL	2.12	109 (1)

Note: DG denotes Dictator Game, PL denotes Prosocial Lies experiment, SL denotes selfish lies experiment, UG denotes Ultimatum Game.

Experimental Procedures. The experimental procedures were identical in all four experiments. Participants A were recruited from the subject pool of the Cologne Laboratory for Experimental Research. To control for the nationality, participants were asked at the end of the experiment about their mother tongue and place of birth. Indeed, all Participants A indicated German as their mother tongue. A very small fraction (4%) indicated an additional native language (such as English, French, Italian, Polish, Filipino or Russian). Excluding these participants from the analysis does not change our results. Individuals with German or Turkish names were recruited for the role of Participants B.

At the very beginning of the experiment, all participants were asked to agree to disclose their first names to the experimental opponent and type it in the respective field. In the event that subjects did not agree to the name disclosure request, they were given the option to close the web browser window and terminate their participation in the experiment (only 0.8% of all subjects did so). Hence, all individuals who participated were informed of the first name of the person they were matched with.

We limited the number of enrolled Participants A to 84 per experiment (assuming a 5% attrition rate, we aimed to have 40 independent observations in each condition). The subjects were assigned randomly to an opponent from GER or TUR conditions and participated only once.

All names of Participants B used in the experiment result from individuals who signed up for the experiment (see Table S1 for an overview). We kept these names constant in the dictator game, prosocial lies and selfish lies experiments to allow the highest possible comparability between experiments. Therefore, some of the Participants B (especially in the TUR condition) took part in more than one experiment. Since we do not study the behavior of Participants B, this detail has no impact on our results.

Our original intention was also to use the same names of the Participants B in the Ultimatum Game as in the other three experiments. However, we did not succeed in finding enough Participants B with the same names used in the other experiments. Thus, some Participants B had different names in the Ultimatum Game as in the previous three experiments.

Each participant was compensated with a participation fee of €2.50. In addition, Participants A were informed that the decision of one out of every 10 individuals – determined at random – would be implemented in accordance with the instructions and would result in additional earnings for the selected individuals. All payments were conducted online via Paypal or bank transfer or with amazon.de gift vouchers.

Belief elicitation about the ultimatum game. We surveyed a total of 134 German participants (53 males) from the subject pool of the Cologne Laboratory for Experimental Research. Subjects were given the description of the ultimatum game we conducted before. They were asked to estimate the average amount offered by proposer (who had one randomly chosen first names that had been used in our experiment) to the receiver. The second question asked their opinion about the income of Turkish immigrants in Germany ("higher/equal/lower than native German population").

The analyses for the male subjects are reported in the main text. Looking at the full sample, we do not observe any gender difference in responses. Over all subjects, we also find no difference in the average offer that subjects expect from proposers with German vs. Turkish names beliefs (Mean=8.08, Median=8, SD=2.97, N=67 from German proposers vs. Mean=8.21, Median=10, SD=2.93, N=67 from Turkish proposers, two-sided MWU test: U=-0.405, p=0.686). Further, we find that also in the full sample 67.16% of the subjects believe that in Germany native Germans have on average a higher income than population with Turkish background, only 2.24% believes the opposite, and 30.60% believe that the incomes are equal.

Additional Analysis

Table 4.A2 Overview of descriptive statistics

	Decision	German names condition	Turkish names condition	Proportion test	Chi-square test
Dictator Game	Selfish choice	26 (61.9%)	26 (63.4%)	Z=0.14 p=0.8869	^{2}(1)=0.02 p=0.887
	Altruistic choice	16 (38.1%)	15 (36.6%)		
Prosocial Lies	Truth	32 (76.2%)	38 (92.7%)	Z=2.07 p=0.0387	^{2}(1)=4.27 p=0.039
	Altruistic lie	10 (23.8%)	3 (7.3%)		
Selfish Lies	Truth	15 (37.5%)	11 (26.2%)	Z=1.10 p=0.2713	^{2}(1)=1.21 p=0.271
	Selfish Lie	25 (62.5%)	31 (73.8%)		

Table 4.A3 Regression analysis of the TUR condition effect.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Experiment	Dictator Game			Prosocial Lies			Selfish Lies		
Dependent variable (0/1)	Altruistic choice			Altruistic lie			Selfish lie		
TUR condition (0/1)	-0.02	-0.02	-0.04	-0.16	-0.16	-0.16	0.11	0.14	0.15
	(0.888)	(0.867)	(0.735)	(0.042)	(0.040)	(0.049)	(0.275)	(0.181)	(0.16)
Socio-demographic controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Bilingual individuals included	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No
Pseudo R-squared	0.000283	0.028583	0.033182	0.062283	0.133183	0.129678	0.011882	0.048782	0.057077
Observations									

Probit regressions (marginal effects). p-values are reported in parentheses. Variable "TUR condition" is a binary variable equal to 1 in TUR conditions and 0 in GER condition. Socio-demographic controls include age and field of study.

Table 4.A4 Pooled regression analysis

	(1) Probit marginal effects	(2) Linear probability model	(3) Probit marginal effects	(4) Linear probability model
Pooled data from...	Prosocial Lies and Dictator Game		Prosocial Lies and Selfish Lies	
Dependent variable (0/1)	Altruistic decision"		Lie	
TUR condition in Dictator Game (0/1)	-0.02	-0.02		
	(0.842)	(0.874)		
GER condition in Prosocial Lies (0/1)	-0.12	-0.15	-0.34	-0.38
	(0.140)	(0.149)	(0.001)	(<0.001)
TUR condition in Prosocial Lies (0/1)	-0.27	-0.30	-0.51	-0.55
	(0.002)	(0.001)	(<0.001)	(<0.001)
TUR condition in Selfish Lies (0/1)			0.13	0.11
			(0.260)	(0.281)
Constant		0.80		0.93
		(<0.001)		(0.002)
(Pseudo) R-squared	0.1005	0.1021	0.2642	0.3206
Observations	166	166	165	165

Notes. p-values are reported in parentheses. The reference group in models (1) and (2) is GER condition of the dictator game. The reference group in models (3) and (4) is GER condition in the selfish lies experiment. All regressions include socio-demographic control variables such as age and field of study.

Table 4.A5 Overview of descriptive statistics and non-parametric analyses of Ultimatum Game

Condition	N	Mean	Median	Standard deviation	Mann-Whitney-U test
GER	40	6.28	6.5	3.94	U=2.26
TUR	41	8	8	3.61	p=0.0239

Table 4.A6 Regression analysis of Ultimatum Game.

Dependent variable: Requested minimum transfer	(1)	(2)
TUR condition (0/1)	1.72	1.76
	(0.043)	(0.036)
Socio-demographic controls	No	Yes
Constant	6.28	8.25
	(0.62)	(0.003)
R-squared	0.0508	0.09
Observations	81	81

Notes. Variable "TUR condition" is a binary variable equal to 1 in TUR condition and 0 in GER condition. Socio-demographic controls include age and field of study.

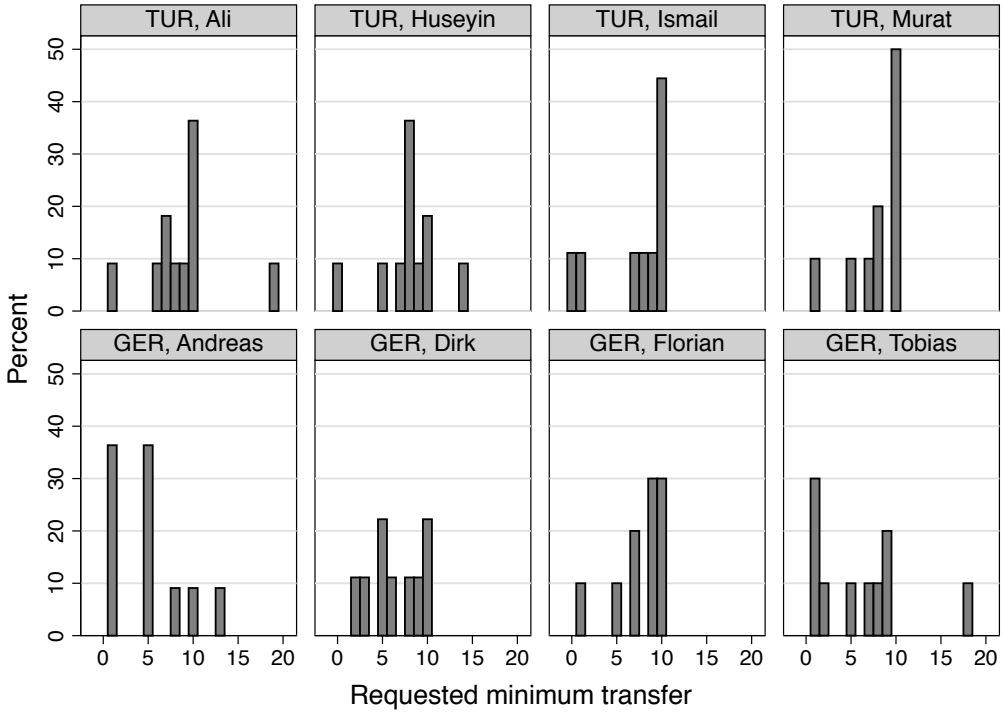

Figure 4.A1 Distribution of the requested minimum transfers in ultimatum game.

Appendix B. Instructions

Instructions.

Experimental Instructions.

Below we report the instructions for subjects in the Role of Dictator/Sender in the Dictator Game, Prosocial Lies and Selfish Lies Experiments (Translation from German)

Screen 1 – In all three experiments

Welcome to this online experiment.

Please consider the following: This experiment is designed in such a way that your first name will be reported to another participant of this experiment. Nevertheless, your decisions remain fully anonymous. By participating in this experiment, you agree with the rule that another participant will be notified of your first name.

If you don't agree with the surname disclosure rule, please close the browser window. By doing this you will opt out of this experiment.

The experiment will take about 15 minutes of your time. Please read these instructions carefully. You may earn a considerable sum of money. At the end of the experiment we will randomly choose 1 participant out of 10 and pay this person according to the instructions below. Additionally, each participant will receive €2.50 as a participation fee. At the end of the experiment, you can choose your preferred method of payment. It can be either an Amazon voucher, or a PayPal or bank transfer.

The instruction on the next page describes what the procedure will be should you be chosen.

Screen 2 – In the Dictator Game

Please note that your decisions are fully anonymous. Neither the other participant nor the experimenter will find out your full name and identity.

You are randomly matched with another participant. You are called Participant A and the other person you are matched with is called Participant B. Neither of you will know the full identity of the other. You and Participant B will know only the first names of each other:

The name of Participant B, with whom you are matched, is **FIRST_NAME**.

Also, Participant A will know your first name. Please type your first name here: _____

Your only decision today is to choose between two options that impose different payments on you and Participant B.

Please choose one of two options:

- **Option 1:** You receive €10 and Participant B receives €10.
- **Option 2:** You receive €9 and Participant B receives €15.

Only you know the particular monetary values connected to each option. Participant B will not be informed about them. However, Participant B will be told that you have been informed of the monetary value connected to each option.

Which option do you choose?

○ Option 1

○ Option 2

Screen 2 – In the Prosocial Lies Experiment

Please note that your decisions are fully anonymous. Neither the other participant nor the experimenter will find out your full name and identity.

You are randomly matched with another participant. You are called Participant A and the other person you are matched with is called Participant B. Neither of you will know the full identity of the other. You and Participant B will know only each other's first names:

The name of Participant B, with whom you are matched, is **FIRST_NAME**. Also, Participant A will know your first name. Please type your first name here: _____

Before starting this experiment, we have rolled a 6-sided dice.

The outcome of the dice roll is:  (in words: TWO).

Participant B is not informed about the outcome of the dice roll. However, he knows that you have been informed about the outcome of the dice roll.

We now ask you to send a message to Participant B. The message is:

"The outcome from the roll of the 6-sided dice is (i)."

where (i) corresponds to a number from 1 to 6.

Your message will be shown to Participant B, and then he will be asked to choose a number between 1 and 6. This is the only information Participant B knows.

The number participant B chooses will determine the payments in the experiment. Participant B will not be informed of these monetary values. But you know them:

- Option 1: If Participant B chooses the **correct outcome** of the roll of the dice, then **you** will receive **€10** and Participant **B** will receive **€10**.
- Option 2: If Participant B chooses a number **different than the actual outcome**, **you** will receive **€9** and Participant **B** will receive **€15**.

Only you know the particular monetary values connected to each option. Participant B will not be informed either of the dice-roll outcome or of the options' monetary values. However, Participant B will be told you have been informed of the monetary value connected to each option.

Which message do you want to send?

The message I will send is "The outcome from the roll of the 6-sided dice is <Scroll 1-6>."

Screen 2 – In the Selfish Lies Experiment

Please note that your decisions are fully anonymous. Neither the other participant nor the experimenter will find out your full name and identity.

You are randomly matched with another participant. You are called Participant A and the other person you are matched with is called Participant B. Neither of you will know the full identity of the other. You and Participant B will know only each other's first names:

The name of Participant B, with whom you are matched, is **FIRST_NAME**.

Also, Participant A will know your first name. Please type your first name here: _____

Before starting this experiment, we have rolled a 6-sided dice.

The outcome of the dice roll is:  (in words: TWO).

Participant B is not informed about the outcome of the dice roll. However, he knows that you have been informed about the outcome of the dice roll.

We now ask you to send a message to Participant B. The message is:

"The outcome from the roll of the 6-sided dice is (i)."

where (i) corresponds to a number from 1 to 6.

Your message will be shown to Participant B, and then he will be asked to choose a number between 1 and 6. This is the only information Participant B knows. The number participant B chooses will determine the payments in the experiment. Participant B will not be informed of these monetary values. But you know these:

- Option 1: If Participant B chooses the **correct outcome** of the roll of the dice, then **you** will receive **€10** and Participant **B** will receive **€10**.
- Option 2: If Participant B chooses a number **different than the actual outcome**, **you** will receive **€15** and Participant **B** will receive **€9**.

Only you know the particular monetary values connected to each option. Participant B will not be informed either about the dice-roll outcome or about the options' monetary values. However, Participant B will be told that you have been informed of the monetary value connected to each option.

Which message do you want to send?

The message I will send is, "The outcome from the roll of the 6-sided dice is <Scroll 1-6>."

Screen 3 – In all three experiments

Now we would like to ask you to fill in a short questionnaire. After the questionnaire, you will be asked to choose your preferred method of payment.

Please indicate your gender: O male / O female

How old are you? ___

What is your place of birth? ____

What is your field of study? ____

Please indicate your mother tongue (more than one choice is possible):
○ German,
○ English,
○ French,
○ Italian,
○ Polish,
○ Russian,
○ Spanish,
○ Turkish,
○ Other ___

Screen 4 – In all three experiments

You can get paid through Paypal, an amazon.com voucher, or by check. Please provide your PayPal email, the email where you wish to receive your Amazon voucher, or a physical address to which we can send a check between ___ and ___.

○ Paypal
○ Amazon.com
○ Bank transfer

Screen 5-A

Please enter your e-mail address for Paypal: _____

Screen 5-B

Please enter the e-mail address where the Amazon gift card should be sent: _____

Thank you for your participation in our online experiment!

Experimental Instructions for Subjects in the Role of Receivers in the Ultimatum Game (Translation from German)

Stage 1

Screen 1

Welcome to our online experiment. The experiment consists of two stages.

In the **first** stage (today), you are only required to confirm that you accept the rules of the experiment.

In the **second** stage (over the course of the next two days), you will make a decision. This will take up to 10 minutes. Here you have the opportunity to earn a significant amount of money, depending on the decision you make during the experiment. You have two options for receiving the amount earned: Paypal or an Amazon.de gift card.

Screen 2

Please note that this experiment is designed such that your first name will be shared with one other participant in the experiment. However, your decision will remain completely anonymous. With your participation in this experiment you agree that your first name will be shared with one other participant. If you agree to this terms, please click on "next." If you do not agree to this terms, please close your browser window. Your participation in this experiment will be canceled.

Screen 3

The first stage has been completed. In the next few days, you will receive a link to the second stage of the experiment. For this purpose, we ask that you enter the e-mail address here where the link should be sent.

E-mail address: _____

Screen 4

Stage 1 has been completed. You may close the browser window.

Stage 2

Screen 1

Welcome to the second stage of the online experiment.

The experiment will take approximately 10 minutes of your time (including reading the instructions). Please read the instructions carefully.

You have the opportunity to earn an amount of money that is dependent on the decisions made by you and the other participants. At the end of the experiment, one out of ten participants will be randomly chosen and paid their respective earnings.

Independent of that, each participant will receive a participation fee of €2.50.

At the end of the experiment, you can choose whether you would like to receive your earnings through Paypal or in the form of an Amazon.de gift card.

The following instructions describe the process should you be selected.

Screen 2

Please note that your decisions will be treated in a **completely anonymous** manner. Neither the other participants nor the experimenter will be informed of your full name or your identity.

You will be randomly paired with one other participant in the experiment. You will take on the role of Participant A, and the other person will play the role of Participant B. Neither of you will find out the identity of the other person. You and Participant B will only see each other's first name. The first name of Participant B matched with you is **Ali**.

Participant B will also be informed of your first name, but only after he has made his decision.

Please enter your first name here: _____

The purpose is to divide €20 between Participant A (you) and Participant B.

Participant B will make an offer for how the €20 should be divided between you and him. That means that Participant B suggests how much you should receive and how much is left over for him.

You decide whether you want to accept or reject his offer. If you accept the offer, you will receive the amount offered by Participant B. If you reject the offer, neither you nor Participant B will receive any amount.

Please choose the smallest amount that you would be willing to accept:

(slider with no default option)

When both you and Participant B have made your decisions, the decisions will be compared:

1. If Participant B offers you an amount that is at least as high as the minimum threshold you chose, then €20 will be divided according to Participant B's offer. You will receive the amount offered and Participant B keeps the remainder.
2. If Participant B offers less than you are prepared to accept, then no division (of the €20) will take place. Neither you nor Participant B will receive any money.

The experiment ends after this decision.

Screen 3

We ask that you fill out a brief survey. After the survey, you can choose your preferred method of payment for the experiment.

Please select your gender:
o male
o female

Please enter your age: _____

Please enter your place of birth: _____

Please enter your field of studies: _____

Please select your native language(s) (you may choose more than one):
o German
o English
o French
o Italian
o Polish
o Russian
o Turkish
o other: _____

References

Ayres, Ian, Mahzarin R. Banaji, and Christine Jolls. 2011. "Race effects on ebay." *Available at SSRN 1934432*.

Adida, Claire L., David D. Laitin, and Marie-Anne Valfort. 2010. "Identifying barriers to Muslim integration in France." *Proceedings of the National Academy of Sciences* 107 (52): 22384-22390.

Allport, Gordon Willard. 1979. *The nature of prejudice*. Basic books, 1979. (Addison-Wesley Pub. Co, Reading, Mass). Unabridged, 25th anniversary ed.

Arrow, Kenneth J. 1998. "What has economics to say about racial discrimination?." *The Journal of Economic Perspectives* 12 (2): 91-100.

Babcock, Linda, and George Loewenstein. 1997. "Explaining bargaining impasse: The role of self-serving biases." *Journal of Economic Perspectives* 11 (1): 109-26.

Becker, Gary S. 1957. "The Economics of Discrimination." *University of Chicago Press Economics Books*.

Bertrand, Marianne. 2009. "CEOs." *Annual Review of Economics* 1 (1): 121-150.

Bertrand, Marianne, and Sendhil Mullainathan. 2003. Emily, Are, and Greg More Employable Than Lakisha. 2004. "Jamal? A field experiment on labor market discrimination." *The American Economic Review* 94 (4): 991-1013.

Blanchflower, David G., Phillip B. Levine, and David J. Zimmerman. 2003."Discrimination in the small-business credit market." *Review of Economics and Statistics* 85 (4): 930-943.

Bruder, Jana, Doris Neuberger, and Solvig Räthke-Döppner. 2011. "Financial constraints of ethnic entrepreneurship: evidence from Germany." *International Journal of Entrepreneurial Behavior & Research* 17(3): 296-313.

Camerer, Colin F., and Richard H. Thaler. 1995. "Anomalies: Ultimatums, Dictators and Manners." *Journal of Economic Perspectives* 9(2): 209-219.

Carlsson, Magnus, and Dan-Olof Rooth. 2007. "Evidence of ethnic discrimination in the Swedish labor market using experimental data." *Labour Economics* 14 (4): 716-729.

Castillo, Marco, Ragan Petrie, Maximo Torero, and Lise Vesterlund. 2013. "Gender differences in bargaining outcomes: A field experiment on discrimination." *Journal of Public Economics* 99: 35-48.

Crosby, Faye, Stephanie Bromley, and Leonard Saxe. 1980. "Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review." *Psychological Bulletin* 87(3): 546.

Devine, Patricia G. 1989. "Stereotypes and prejudice: their automatic and controlled components." *Journal of personality and social psychology* 56 (1): 5-18.

Devine, Patricia G., and Andrew J. Elliot. 1995. "Are racial stereotypes really fading? The Princeton trilogy revisited." *Personality and social psychology bulletin* (21): 1139-1150.

De Groot, Olaf J., and Lutz Sager. 2010. "Migranten in Deutschland: soziale Unterschiede hemmen Integration." *DIW Wochenbericht* 77(49): 2-9.

Dovidio, John F., and Samuel L. Gaertner. 2004. "On the nature of contemporary prejudice." *Race, class, and gender in the United States: An integrated study*.

Ewens, Michael, Bryan Tomlin, and Liang Choon Wang. "Statistical discrimination or prejudice? A large sample field experiment." *Review of Economics and Statistics* 96, no. 1 (2014): 119-134.

Erat, Sanjiv, and Uri Gneezy. 2012. "White Lies." *Management Science* 58 (4): 723–733.

Fershtman, Chaim, and Uri Gneezy. 2001. "Discrimination in a Segmented Society: An Experimental Approach." *The Quarterly Journal of Economics* 116 (1): 351-377.

Fiske, Susan T. 1998. Stereotyping, prejudice, and discrimination. *The Handbook of Social Psychology, Vols. 1 and 2 (4th Ed.)*, eds Gilbert DT, Fiske ST, Lindzey G (McGraw-Hill, New York, NY, US), 357–411.

Fong, Christina M., and Erzo FP Luttmer. 2011. "Do fairness and race matter in generosity? Evidence from a nationally representative charity experiment." *Journal of Public Economics* 95 (5): 372-394.

Fryer, Roland G., and Steven D. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *The Quarterly Journal of Economics* 119 (3): 767-805.

Gelman, Andrew, Jeffrey Fagan, and Alex Kiss. 2007. "An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias." *Journal of the American Statistical Association* 102(479)813-823.

Gneezy, Uri. 2005. "Deception: The role of consequences." *American Economic Review* 95(1): 384-394.

Gneezy, Uri, John List, and Michael K. Price. 2012. *Toward an understanding of why people discriminate: Evidence from a series of natural field experiments*. No. w17855. National Bureau of Economic Research.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An experimental analysis of ultimatum bargaining." *Journal of economic behavior & organization* 3 (4): 367-388.

Kaas, Leo, and Christian Manger. 2012. "Ethnic discrimination in Germany's labour market:a field experiment." *German Economic Review* 13(1): 1-20.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. "Fairness as a constraint on profit seeking: Entitlements in the market." *The American economic review* 76(4): 728-741.

Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in Allocation Decisions." *American Economic Review* 90 (4), 1072-1091.

Konow, James. 2003. "Which Is the Fairest One of All? A Positive Analysis of Justice Theories." *Journal of Economic Literature* 41 (4): 1188-1239.

Kühnel, Steffen, Jürgen Leibold, and Anja Mays. "Die gegenseitigen Wahrnehmungen und Einstellungen von Einheimischen und MigrantInnen." In *Dabeisein und Dazugehören*, pp. 203-226. Springer Fachmedien Wiesbaden, 2013.

List, John A. 2004. "The nature and extent of discrimination in the marketplace: Evidence from the field." *The Quarterly Journal of Economics* 119 (1): 49-89.

Madon, Stephanie, Max Guyll, Kathy Aboufadel, Eulices Montiel, Alison Smith, Polly Palumbo, and Lee Jussim. 2001. "Ethnic and national stereotypes: The Princeton trilogy revisited and revised." *Personality and Social Psychology Bulletin* 27 (8): 996-1010.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The dishonesty of honest people: A theory of self-concept maintenance. " *Journal of Marketing Research* 45 (6): 633-644.

McConahay, John B. 1986. "Modern racism, ambivalence, and the Modern Racism Scale." *Prejudice, Discrimination, and Racism*, eds Dovidio JF, Gaertner SL (Academic Press, San Diego, CA, US), 91–125.

Ochs, Jack, and Alvin E. Roth. 1989. "An experimental study of sequential bargaining." *The American Economic Review* 79(3): 355-384.

Phelps, Edmund S. 1972 "The statistical theory of racism and sexism." *The American Economic Review* 62(4): 659-661.

Pinker, Steven. 2003. *The blank slate: The modern denial of human nature*. Penguin.

Ross, Catherine J. "Turkey: At the Crossroads of Secular West and Traditional East: Perennial Outsiders: The Educational Experience of Turkish Youth in Germany." 2009. *American University International Law Review 2*4: 685-943.

Schelling, Thomas. 1969. "Models of Segregation." *American Economic Review* 59(2): 488-93.

Sutter M (2009). Deception through telling the truth? Experimental evidence from individuals and teams, *The Economic Journal*, 119(534): 47-60.

Thaler, Richard H. 1988. "The Ultimatum Game." *The Journal of Economic Perspectives* 2(4): 195-206.

Wolfers, Justin. 2006. "Diagnosing discrimination: Stock returns and CEO gender." *Journal of the European Economic Association* 4 (2-3): 531-541.

5. On the Size of the Gender Difference in Competitiveness

Abstract

We design and test a new procedure for estimating the magnitude of the gender gap in competitiveness. Before working on a task, participants choose what percentage of their payoffs will be based on a piece-rate compensation scheme; the rest of their payoff is allocated to a competitive compensation scheme. This allows us to measure 101 different levels of competitiveness. We find that the size of the gender gap is larger than previous research has suggested, in particular between the most competitive participants. For example, we find that the top competitive 10 percent of our participants are all men.

5.1 Introduction

In 2013, the number of female CEOs of Fortune 500 companies reached a historical high of 23 (or 4.6 percent of all CEOs). Although the increase in female leadership is encouraging, this number shows that the gender gap in the labor market is still large.[1] All over the world, women earn, on average, less than men in similar jobs, and in all but four countries, females account for substantially less than half of the senior positions in business and government. Strikingly, in only 11 countries (2.4% of the world population) do women account for their share of senior positions in business and government (Hausmann et al. 2010). Researchers have proposed a large variety of causes to explain this difference, including discrimination and differences in work-home preferences. In this paper, we focus on the magnitudes of individuals' preferences for competition.

Over the last decade, a stream of experimental research has argued that women are less competitive than men, and that this difference could partially explain the differential success between men and women in the labor market (see Croson and Gneezy, 2009, for a survey). This research has focused on two aspects of competitiveness. One line of literature measured participants' reactions to changes in the competitive nature of the compensation schemes, and showed that, when forced

---

[1] For discussions regarding the estimated size of the gender gap in, for example, wages, and for possible economic explanations, see Altonji and Blank (1999); Bertrand and Hallock (2001); Bertrand (2010); and Goldin, Katz, and Kuziemko (2006). For recent statistics, see *USA Today*, 2011, http://www.usatoday.com/money/companies/management/story/2011-10-26/women-ceos-fortune-500-companies/50933224/1.

into a competitive setting, women perform worse than men (e.g., Gneezy, Niederle, and Rustichini, 2003; Gunther et al., 2010; Shurchov, 2012). A second line of literature has investigated individuals' self-selection into competitive environments. Typically, these papers ask participants to choose a compensation scheme for themselves (e.g., Niederle and Vesterlund, 2007). The major result of this line of literature is that women select into competitive environments less often than men.[2]

The question of how to model individual differences in competitiveness is still open. Consider the following schematic model of individuals' competitiveness. We assume each individual is characterized by some tendency to compete, denoted by c. Larger values c represent higher tendency to compete. Normalize the level of competitiveness in the population to be distributed between 0 and 1, with c=0 denoting a person who never competes, and c=1 a person who always competes. The distribution of c could take any shape.

A first hypothesis is that the distribution of c is gender specific, that is, that men and women have a different tendency to compete. In particular, according to this hypothesis, the distribution of c is more shifted toward zero for women than for men, resulting in a lower level of competitiveness for women. This is a direct preference effect that occurs if people have some inherent taste for competition distinct from, for example, risk and ambiguity preferences—a "competitive spirit."

---

[2] For further evidence on gender differences in selecting into competition, see Andersen, Ertac, Gneezy, List, and Maximiano (2012); Balafoutas and Sutter (2012); Booth and Nolen (2012); Cason, Masters, and Sheremeta (2010); Datta Gupta, Poulsen, and Villeval (2011); Dohmen and Falk (2011); Ertac and Szentes (2010); Gneezy and Rustichini (2004); Gneezy, Leonard, and List (2009); Healy and Pate (2011); Niederle, Segal, and Vesterlund (2012); Sutter and Rützler (2010); Vandegrift and Yavas (2009); and Wozniak, Harbaugh, and Mayr (2009).

An alternative hypothesis is that the distribution of c is not gender specific, but rather that the minimal level of c for which a person would choose to compete in a given task, represented by a threshold c_m is gender specific. Participants for whom $c<c_m$ choose not to compete, whereas those for whom $c>c_m$ do. This hypothesis accounts for the possibility that, even if men and women have the same competitive disposition, they may choose differently because of indirect preference. For example, different cultures and societies can affect c_m differently for men and women. Even if the initial disposition of the competitiveness of women is the same as for men, nurture can move c_m such that women will be less likely to compete (see Gneezy, Leonard, and List, 2009).

To give this schematic model an empirical basis, one needs a more refined measure of competitiveness than the one used in the current literature. The papers that study selection into competitive environments typically use the choice of incentive scheme as a binary measure of competitive behavior: participants are asked to choose whether they would like to be paid according to a piece-rate incentive scheme or a tournament incentive scheme (as presented in the left part of Figure 1 below). Note that in this line of research "competitiveness" is defined as the selection of the tournament scheme. A robust finding of this line of research is that men choose competitive incentives (the tournament scheme) more often than women, and hence, that the "average woman" is less competitive in this context than the "average man." A binary measure, which focuses on competitiveness on the extensive margin (the choice whether to enter the competition), does not allow us to test the above two hypotheses.

Further, while fewer women than men choose to compete in these experiments, on average, around one third of women do choose the competitive incentives. Hence, when considering the role of competitiveness in explaining the extent of the gender gap in elite labor market outcomes, these results present some shortcomings. In particular, they seem to understate the size of the gap and to suggest that about a third of the women are as competitive as the competitive men. Given the small gender differences in competitiveness relative to the vast gap in top competitive jobs in certain sectors of the labor market, competitiveness must only play a relatively minor part.

In contrast to the binary measure, in this paper we introduce a measure that allows us to observe 101 different levels of competitiveness: we ask participants to choose what percentage of their compensation they would prefer to be derived from the tournament scheme and what percentage to be derived from the piece-rate scheme (as presented in the right side of Figure 1). This measure is similar in spirit to Gneezy and Potters' (1997) measure of risk aversion and it allows us to measure an intensive margin of competitiveness. Importantly, this measure allows us to better understand the distribution of c.

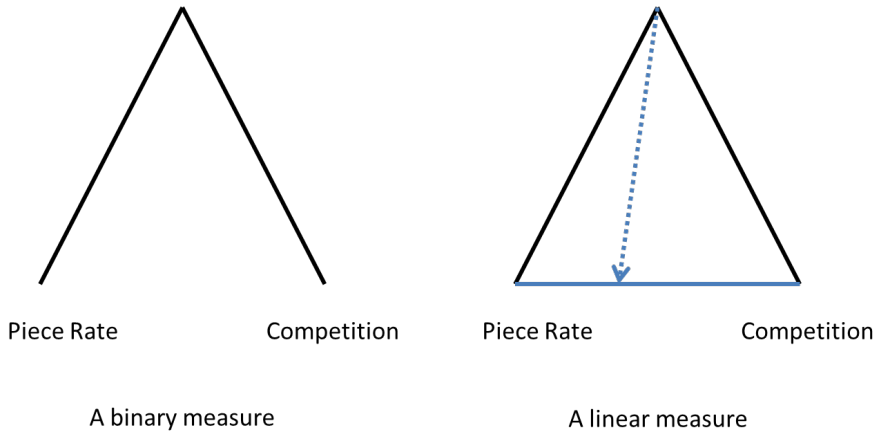

Figure 5.1 Illustration of the Binary and Linear Measure

Note: The left side represents the binary measure in which participants are asked to choose between a piece-rate compensation scheme and a competition-based compensation scheme. The right side represents the linear measure in which the participants can choose a combination of the two compensation schemes.

Using this new measure, we explore a new dimension of competitiveness—the *extent* of competitiveness. We show that the evidence for gender differences in competitiveness is much stronger than that revealed by previous experimental paradigms. Through more refined experimental procedures, we are able to collect a richer data set that allows us to perform analyses that are more informative than the examination of averages alone; we measure the *size* of the gender gap in competitiveness—the extent to which one individual is more competitive than another—and examine how the distribution of competitiveness differs between the populations of men and women. With this measure, we are able to observe the underlying distributions of both populations, which are masked when measuring competitiveness on the extensive margin, through a binary choice. The differences between men and women at the upper tail of the distribution of competitiveness cannot be observed through a binary choice—any two levels of competitiveness within the

same binary classification will appear to be the same. When looking at the population of men and women combined, we show that the ratio of women to men in the distribution of competitiveness decreases as the degree of competitiveness increases.

While in the extensive margin measure the women to men ratio reveals that of all the people who choose the tournament approximately one third are women, the intensive margin measure reveals that this proportion is much smaller in the upper tail of the distribution of competitiveness, where the women to men ratio substantially decreases. Strikingly, all the participants in the top 10 percent of the distribution of competitiveness are men. Our results suggest that the distribution of c in our sample is gender dependent.

Our refined measure of competitiveness also allows us to investigate the role of confidence, risk taking and ambiguity aversion, in determining the gender gap. We find, for example, that confidence and risk attitudes explain part of the variance of competitiveness and predict whether participants are in the top 75^{th} percentile of the competitiveness distribution; yet, they do not account for the entire gender gap.

To test the hypothesis that women and men differ in their minimum level of competitiveness required to enter a tournament, c_m, we can compare the results of the extensive and intensive measures to backup the average cutoff points or decision rules in the above model. We find that women's c_m is not larger than men's c_m. Under the assumption that participants' behavior in the two measures is directly comparable, our result suggests that the gender difference in entering the competition is due to differences in the distribution of competitiveness preferences rather than differences in the cutoffs.

If successful careers in certain segments of the labor market demand a high level of competitiveness, we can reasonably project that a weaker preference for competition will lead fewer women to commit to such career paths. Our results suggest that women with highly competitive preferences may be the exception rather than the rule. These results have implications beyond career choice and financial success. An individual's competitiveness also affects her likelihood of engaging in other competitive interactions, such as auctions and bargaining. Hence attitudes toward competition may affect, for example, entry into wage negotiations. This could have bearing on the wage gap between men and women in similar occupational positions (Babcock and Laschever, 2003; Rigdon, 2013).

5.2 Experimental Design

In our experiment all participants faced the same task and competitiveness was either measured using a binary choice (extensive margin) or a linear choice (intensive margin). For this purpose, we employed two treatments. Participants in both treatments chose how they wanted to be compensated for completing a ball-tossing task. The ball-tossing task involves tossing a tennis ball into a small basket 10 feet away. In the task, participants are given 10 opportunities to make successful tosses—tosses that land and stay in the basket. Each toss must be completed underhand. We explained the task in detail to the participants at the start of each experimental session. That is, while reading out loud the instructions about the task the research assistant

showed the tennis ball to the participants and mimicked tossing it in a basket placed 10 feet away, making clear that tosses had to be completed underhand. The experimenter did not actually perform the task to prevent a successful/unsuccessful outcome from affecting participants' beliefs about the difficulty of the task. Participants performed the task in private so that no other participant could observe their performance.

Differently from the Niederle and Vesterlund (2007) paradigm, in our experiment we did not collect a baseline measure of ability before measuring participants' willingness to compete. We made this design choice not to have a within-participant design to prevent the different stages of the experiment from affecting each other. Previous research with this task in a different participant pool (Gneezy, Leonard, and List, 2009) found no gender differences in ability. To exclude the possibility of a gender difference in ability in our sample, we conducted a separate between-participants test in which we measured ability by asking participants to engage in the task without letting them choose their compensation scheme. All instructions can be found in Appendix B. This test consisted of 84 participants (42 women) that belonged to the same participant pool as in the main experiment. These participants took part in an unrelated laboratory experiment and were asked to complete the task for no incentives. The test was conducted in the Fall 2013. While the gender differences in competitiveness we show in this paper may be task-specific (as is true for any task used), our focus is on the comparison of the measures. Future research can test whether the different measures are influenced differently by the nature of the task.

The main experiment was conducted in a university laboratory with a total of 210 participants in two treatments. The binary treatment, which tests behavior on the extensive margin, consisted of 126 participants (71 women), with six participants per session. Seven sessions of this treatment were conducted in Winter 2012, seven sessions were conducted in Fall 2013, and seven in Fall 2014.[3]

The linear treatment, testing behavior on the intensive margin, consisted of 14 experimental sessions with six participants in each session. The sessions were conducted between Winter 2012 and Spring 2013. There were 84 total participants (44 women) in the linear choice experiment. On average, participants earned $8.20 including the show-up fee.[4]

5.2.1 Measures of Competitiveness

The only difference between the binary (extensive margin) and the continuous (intensive margin) measure of competitiveness was in the decisions participants made about their compensation for the task. Each participant made this choice before the beginning of the task.

Competitiveness on the Extensive Margin

---

[3] We originally had 84 participants per treatment but did not have data on confidence, risk, and ambiguity for half of the subjects who performed the binary task. In order to make use of that data and make sensible comparisons between treatments, we collected an additional 42 observations in the binary measure. We thank an anonymous referee for this suggestion.

[4] There was no difference in earnings across the two measures ($8.46 in the binary elicitation vs. $7.90 (Sd=4.04) in the linear elicitation, Mann-Whitney test, p=.38).

The typical elicitation of competitiveness focuses on choices on the extensive margin. This measure entails a binary choice between two compensation schemes: a tournament compensation scheme (*T*) and a piece-rate compensation scheme (*PR*). The piece-rate scheme is based on individual performance alone: participants are paid $1 per successful toss independent of others' performances. The tournament compensation scheme pays $3 per successful toss if a participant wins against a randomly chosen opponent. The opponent was chosen ex post from the entire pool of participants from the same session, men and women, not just those who chose to compete. The participant is paid nothing if she loses in the competition, and $1 per successful toss in case of a tie.

Competitiveness on the Intensive Margin

To measure competitiveness on the intensive margin we introduce a continuous measure of competitiveness. This measure asks participants to choose a linear combination of tournament compensation and piece-rate compensation to compose her overall payoff for the given task. The procedure for this experiment is similar to the binary choice experiment except that participants choose how to allocate an endowment between the piece-rate scheme and the tournament scheme. Participants were endowed with 100 points and were asked to allocate a portion of these points, from 0 to 100 inclusive, to the tournament scheme and the rest to the piece-rate scheme. At the end of the experiment, we paid participants $1 for every 100 points earned.

That is, the decision maker receives 100 points and is asked to choose how much of it, *t*, she wishes to invest in the tournament option, *T*, and how much to invest in the piece-rate option, *PR*. The payoffs are then (100-*t*+3*t*)*(the number of successful tosses) if the participant scores higher than her opponent, and (100-*t*)*(the number of successful tosses) if the participant scores lower than her opponent. In case of a tie, the participants simply get 100 times the number of successful tosses, or (100-*t*+*t*)*(number of successful tosses).

This allocation t represents the percentage of the resulting tournament payoff that will be included in the realized payoff. The remainder of the realized payoff is comprised of the complementary percentage, (100-t)%, of the resulting piece-rate payoff. Thus you might calculate what she would have received from the tournament scheme, tournament payoff, and what she would have received from the piece-rate scheme, piece-rate payoff, according to the definitions above. Then her total compensation for the task would be calculated according to $\Pi = \left(\frac{t}{100}\right)\pi^{T} + \left(1 - \frac{t}{100}\right)\pi^{PR}$, with π^{T}, her tournament payoff and π^{PT}, her piece-rate payoff. For example, if t=50, we can say the agent receives 50 percent of her resulting tournament payoff and 50 percent of her resulting piece-rate payoff. This representation is equivalent to the point-based representation.

This allocation to the tournament, *t*, is our measure of competitiveness. An individual is deemed more competitive than another if she chooses to include a greater amount of the tournament payoff in her chosen payoff combination than another individual.

Additional Measures

In addition to eliciting levels of competitiveness, in all sessions but the Fall 2013 ones we measured other factors that may affect competitiveness. With these measures we can observe the effect of hedging uncertainty and beliefs. There may of course be other factors contributing to gender differences in competitiveness, such as social preferences (see Balafoutas et al., 2012, for such a relation between social preferences and competitiveness). The instructions are reported in Appendix C.

Confidence

Participant's decision to compete might be affected by their confidence about their expected relative performance as compared to a random opponent from their same session. To explore the role of confidence, we used two measures that were administered before the start of the ball-tossing task but after the choice of an incentive scheme. First, we asked participants to guess their expected number of successful tosses on a scale from 0-10 ("How many successful tosses do you think you will make?"). We label this measure "Expected performance." Second, we asked participants to state the expected likelihood of winning against a random opponent, as a percentage from 0-100 ("What do you believe is the probability that you will make more successful tosses than a randomly selected opponent?"), which we refer to as "Confidence of Winning."

Measuring beliefs is always a tricky task. We decided not incentivize this belief elicitation in order to keep the instructions simple and avoid cross influence

between beliefs and effort on the task. There is no strategic reason for participants to misreport their beliefs in our experiment. As for the timing of the question, we ask about expected performance after the entry decision has been made. This could produce a bias, coming from the fact that the choice of tournament by itself could cause the participants to report higher confidence (just to reaffirm the decision). Since more men selected into the tournament, or selected to participate more heavily, this in itself could cause us to observe more "confidence" among men than women. However, eliciting confidence before the choice of a payment scheme could have biased the entry decision.

Risk attitudes

Since the choice of being compensated according to a tournament scheme (or of allocating more points to the tournament scheme) can depend on the participant's risk attitudes, we elicited risk attitudes using two different measures. First, after making the choice but before performing the task, we elicited risk attitudes through multiple price list (MPL; see Holt and Laury, 2002) measure of risk aversion. The measure was incentive-compatible. Participants were presented with a series of 10 decisions between pairs of gambles (A and B). In all 10 decisions the payoff for each gamble, A and B, remained constant but the probability of getting the higher payoffs (B) increased moving from decision 1 to decision 10. We asked participants to indicate their "switch point", the point at which they decided to switch from choosing to be paid according gamble A to choosing to be paid according to gamble B. The "switch point" serves as a measure of Risk aversion, with more risk-averse

participants indicating higher the switch points. We denote this measure as "Risk Aversion." Decisions made using this measure were compensated at the end of the experiment (see payment procedure below).

We also elicited self-assessed risk taking on a scale from 0-10 using the following question: "Please answer the following question using a 1-10 scale, where *1=completely unwilling* and *10=completely willing*: Rate your willingness to take risks in general." This measure is adapted from Dohemen et al. (2011), which find it to be predictive of risky behaviors and of participants' choices in an incentivized risk task.

Ambiguity aversion

We also assessed ambiguity preferences with an MPL over known and unknown lotteries. As is typical of MPL, participants are presented with a series of 20 decisions. Each decision entails a choice over a known and an unknown lottery. Similar to the risk measure, participants must indicate a "switch point"—the point at which they are willing to switch from entry into the known lottery to entry into the unknown lottery. This switch point serves as a measure of aversion to ambiguity and represents the premium the agent is willing to pay to avoid the ambiguous outcome. Responses for the two MPLs (risk and ambiguity measures) were compensated at the end of the experiment. In particular, participants were paid for one of the two MPLs, determined at random by a coin-flip.

5.2.2 Procedure

Participants were invited to the lab using standard recruiting procedures. Each session had six participants. We invited more participants to make sure to have six people session and in case more than 6 people showed up, we dismissed the exceeding participants. Our goal was to have gender-balanced sessions, with 3 women and 3 men. However, since we could not recruit participants by gender, this was not possible at all times. After being seated at their computer station, participants started to read the instructions. The instructions explained that participants would participate in a ball-tossing task and that they had to decide how to be compensated for it. To make sure participants understood the instructions, before they made their decision, an experimenter also read the instructions out loud. The experimenter also gave a demonstration of the task without actually tossing the ball in the basket. Participants were invited to ask any clarification questions about the task and the payment scheme. Questions were taken by the research assistant in private. In both treatments, participants did not practice the task before making their compensation choice.

In the "Extensive Margin" treatment, participants had to make a binary choice between being compensated according to a piece-rate ($1 for each successful toss) or to a tournament payment scheme ($3 for each successful toss if the total number of tosses was greater than the total tosses of a random opponent from the same session). In the "Intensive Margin" treatment, participants were informed that they had to divide 100 points between a piece rate and a tournament compensation scheme. Participants were explained that the tournament option paid 3 cents per point for each successful toss if their total number of tosses was greater than the total tosses of a random

participant in the room. The piece rate option paid 1 cent per point for each successful toss. Participants knew that all 100 points had to be allocated. We did not ask check-up questions in order not to provide participants with an anchor for their decisions. However, we instructed participants to raise their hand and ask questions in case they did not understand how the payment worked. For all the 84 participants in the "Intensive Margin" measure, the number of points allocated to tournament and piece rate totaled to 100, suggesting that participants did understand the instructions.

Throughout the experiment gender was not made salient. Participants only knew that if they selected the tournament their performance would be matched with the performance of a random opponent in the room. Most of the sessions were gender balanced. The computer stations in the lab faced the walls of the laboratory and were separated by dividers, preventing participants from looking at each other without completely turning. This does not prevent participants from knowing the gender composition of the experimental session before the experiment started, but we did not make gender salient.

In all sessions but the Fall 2013 ones, after making their choices participants were distributed a short survey aimed at eliciting their confidence. After that, they were handed out two separate envelopes. The first envelope contained the instructions and decision sheet for the risk attitudes measure, whereas the second envelope contained the ones for the ambiguity preferences measure.[5] Participants were informed that in this portion of the experiment they would be paid according to the realization of

[5] Participants in the Fall 2013 sessions did not complete such measures. We collected additional data that includes these measures in Fall 2014. We thank an anonymous referee for the suggestion.

one decision across both tasks, and that the task determining their payment would be determined by a coin toss at the end of the experiment. If the coin landed on heads, participants would be paid according to the risk measure. If it landed on tails, they would be paid according to the ambiguity measure.

After that, each participant was directed to a separate room to perform the ball-tossing task, while the rest of the participants waited at their computer station. No communication was allowed between participants at any moment throughout the experiment. At the end of the experiment, participants filled out a short survey of basic demographic information, asking them for their age, ethnicity, spoken language, major, and GPA. The questionnaire also contained the self-reported measure of risk. At the end of the experiment, we paired each participant anonymously with a random opponent from the same session. We then paid them according to their choice of compensation scheme offerings and, in the relevant cases, according to the outcome of the tournament. After that, we also paid each participant for either the risk or the ambiguity measure, depending on the outcome of the coin toss. We used a 10 or 20-sided dice to determine the decision row according to which participants were paid.

In the next section we start by presenting the results of the pre-test aimed at measuring participants' ability, then the results of the extensive margin measure, followed by the results of the intensive margin measure.

5.3 Results

5.3.1 Ability

We first present the data of the separate test (N=84) in which participants performed the ball-tossing task with no monetary incentives. We do not find significant gender differences in performance in this task in our sample. This result is in line with previous research using the same task (Gneezy, Leonard, and List, 2009). On average, women completed 2.12 tosses successfully (SE=0.19), while men completed 2.40 tosses successfully (SE=0.25). Because in competition the distribution might also be important, the cumulative distribution of the number of successful tosses for each gender is displayed in Figure 2.

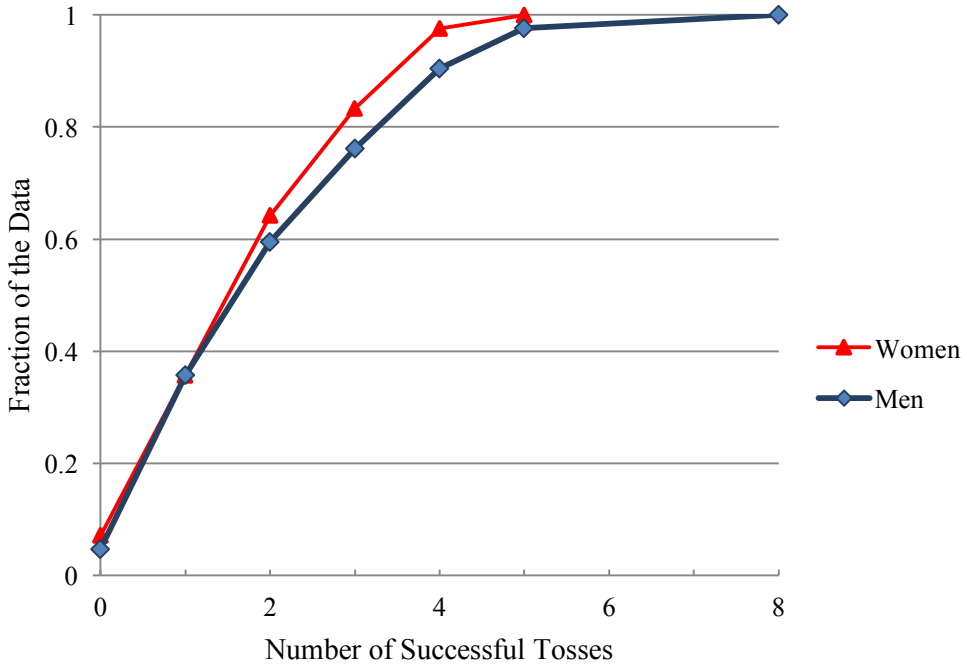

Figure 5.2 Empirical CDF of Number of Tosses by Gender

For every possible number of tosses, the graph shows the proportion of women or men who successfully scored up to that many tosses. The performance distributions of men and women are not statistically different (z= -0.547, p=.584 Mann-Whitney). This suggests that differences in choice of competitive scheme are not driven by gender differences in ability to perform this particular task.

5.3.2 Competitiveness on the Extensive Margin

The results of the extensive margin measure treatment provide a benchmark to which we can compare the results of the intensive margin measure. A total of 126 participants (71 females) took part in three waves of this treatment. The average age

was 21 years old. Additional descriptive statistics about the sample are reported in Appendix A.

The tournament entry results for the three rounds of experimental sessions are summarized in Table 1. We observe no statistical difference in the proportion of women and men who selected the tournament across the three waves of sessions (Men: $\chi^2(2)=3.02$, p=.22; Women: ($\chi^2(2)=0.35$, p=.84). Therefore, we pool all the binary task data together for the analyses. Of 126 individuals in the sample, 52.4 percent chose to participate in the tournament. Of the women, 32.4 percent chose to participate in the tournament, whereas 78.2 percent of men chose the tournament. This difference is statistically significant ($\chi^2(1) = 26.05$, p<.001).

Table 5.1 Fraction of Participants who selected the Tournament

	2012 Sessions (N=42)		2013 Sessions (N=42)		2014 sessions (N=42)		Pooled data (N=126)				
	Fraction	*N*	*Fraction*	*N*	*Fraction*	*N*	*Fraction*	*N*	*Min*	*p50*	*Max*
Men	0.71	14	0.70	20	.90	21	0.78	55	0	1	1
Women	0.29	28	0.36	22	.33	21	0.32	71	0	0	1
All	0.43	42	0.52	42	61.9	42	0.52	126	0	1	1

This gender gap in competitiveness replicates the results of previous research in selection into competitive environments using the same or a different task. In a task requiring participants to add up some numbers, Niederle and Vesterlund (2007) find a similar gap in tournament entry, with 73% of the men and 35% of the women choosing the competitive scheme. In a maze-solving task, Gupta et al. (2005) find that

60% of men, but only 34% of women choose the tournament option. Further, using the number addition task, Dargnies (2009) find that 51.3% of the women and 84.6% of the men chose the tournament option, Balafoutas and Sutter (2012) find that the tournament is chosen by 30.4% of the women and about twice as many men, Niederle et al. (2010) find that it is chosen by 31% of the women and 73% of the men, whereas Healy and Pate (2011) that it chosen by 28% of the women and 81% of the men. In a sample of children, Sutter and Rutzler (2010) also find a gender gap, with 19% of girls and 40% of boys choosing to compete on the NV task. In a similar math task, Dohmen and Falk (2011) found that 37.3% of the women and 62.3% of the men self-selected into a tournament compensation scheme.

Using the same task we adopted in our experiment, the ball-tossing task, Gneezy et al. (2009) find a comparable gender gap in an African patriarchal society, where 50% of the men and 26% of the women compete. With the same task, Andersen et al. (2013) find a slightly larger gap in a sample of adolescents from an Indian patriarchal society (19% vs. 67%), while no gap is found in children from the same society, and in children or adolescents from a matriarchal society. Other research has showed that when using more female-oriented tasks (e.g., verbal tasks) men are not more likely to select into the tournament than women, possibly because women are more confident about their performance in such tasks (Wozniak et al., 2010; Gosse and Reiner, 2010; Shurchkov, 2011). Thus, the gender gap we observe in our sample is in line with the overall results observed in the literature in which the tournament in chosen by about twice as many men than women (see Niederle and Vesterlund, 2010 for a review).

To confirm the robustness of our result, we further investigate it using a regression framework. Table 2 reports the results of different specifications of a probit model in which we regress a tournament entry dummy variable on a female dummy. As shown in the table, women are significantly less likely to choose the competitive scheme. The estimated marginal effect reported in column 1 suggests that women are 36.4 percentage points less likely to enter the tournament than men.

Since the women to men ratio was not constant across sessions in column 2 we control for the gender composition of the sessions. In particular, whereas most of the observations (66.7%) come from gender-balanced sessions (3 men and 3 women), some of the sessions (28.6% of the observations) were characterized by a majority of women (3 sessions had 4 women and 2 men, and 3 sessions had 5 women and 1 men), and one session (4.8% of the observations) had a majority of men. Since the gender of a potential competitor may affect participants' willingness to compete, it is possible that participants' choices differed in the unbalanced sessions. We control for this heterogeneity in the gender composition of the sessions by adding to the model a variable indicating the women to men ratio in each session[6]. When adding this variable to the model we find women to be 39.9 percentage points less likely to select into the tournament. Further, running a regression only on the 66.7% of the participants (N=84) who took part in gender-balanced sessions leads to a similar

[6] Alternatively, considering a) the total fraction of women in the session, b) adding to the model dummy variables for sessions with more women than men and for sessions with more men than women, or c) adding session dummies to the regression results in a similar gender gap.

result, with an estimated gender gap of 36.7 percentage points (p<.001). The results are reported in Appendix A.

Table 5.1 Probit of Tournament Entry Decisions

Choice of Tournament	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-.364***	-.399***	-.423***	-.334***	-.318***	-.313 ***	-340***
	(.046)	(.086)	(.103)	(.112)	(.114)	(.116)	(.123)
Gender composition		.001	-.001	-.016	-.021	-.016	-.004
		(.036)	(.037)	(.048)	(.047)	(.049)	(.049)
Expected Performance			.005				
			(.027)				
Confidence Winning				.011**	.011**	.011**	.011**
				(.004)	(.005)	(.005)	(.005)
Self-reported Risk					-.001	-.001	-.000
					(.034)	(.033)	(.034)
Risk Aversion						046	.049
						(.030)	(.031)
Ambiguity aversion							-.019
							(.012)
Year dummy	N	Y	Y	Y	Y	Y	Y
Observations	126	126	84	84	83	80	80
Pseudo R^2	.156	.166	.217	.307	.304	.312	.329

Note: The table presents marginal effects estimated from probit regression. Dependent variable: Choice of tournament (1 tournament and 0 piece-rate). Gender composition refers to the women to men ratio in each session. Expected performance refers to the estimated number of successful tosses. Confidence of winning refers to subjects' expected likelihood of winning against a random opponent from the same session. Self-reported Risk refers to the self-reported willingness to take risks. Risk aversion refers to the incentivized measure of risk. Ambiguity aversion refers to the incentivized measure of ambiguity. Marginal effects are estimated at a man in a gender-balanced session, and at the mean for all the other variables. Robust standard errors are in parenthesis.

When considering participants' performance conditional on tournament choice, we find that subjects completed an average of 2.48 tosses (SD=1.68). Further, we find that men performed better than women under the tournament (MW test, z = 2.277, p=.02), which is in line with previous literature (Gneezy, Niederle, and Rustichini,

2003), and marginally better than women under the piece rate (z = 1.717, p=.09).[7] Both results become insignificant if we exclude the top participants who perform 5 or more successful tosses (11.66% of the participants; MW test, z =1.438, p=.15 for the tournament; MW test, z = 1.289, p=.20 for the piece rate). If we exclude these observations and run a regression of tournament on female we find that females are still 35.5 percentage points less likely to select the tournament than men (p<.001, N=106, see Appendix A)."

Determinants of Competitiveness on the Extensive Margin

In this section, we investigate whether the gender gap in tournament entry is driven by gender differences in participants' confidence of winning the competition, in risk preferences and ambiguity aversion.

Confidence. To measure confidence, we elicited participant's belief about their performance as well as their likelihood of winning against a random opponent. We find that both men and women in our sample were overconfident regarding the number of successful tosses they would perform in the task. On average, participants expected to successfully make 5.04 tosses (SD=1.93, N=84), which is bigger than the actual average number of tosses completed by these participants (z=7.62, p<.001, signrank test). In line with previous literature we find a gender gap in overconfidence. In

---

[7] The data regarding the number of successful tosses is missing for one of the session (N=6) as it was not collected due to a mistake in the experimental procedures.

particular, men expected an average of 5.79 tosses (SD=1.75, N=35), while women expected to successfully complete an average of 4.51 tosses (SD=1.90, N=49), with the two distributions being significantly different (z=2.550, p=0.01, Mann-Whitney). When we look at participants' confidence of winning against a random opponent, we find a similar gender difference. On average, men reported expected likelihood of winning is 63.13%, while women's is 43.2% (z=4.731, p<.001, Mann-Whitney). The two measures of confidence are strongly correlated (r=0.62, p=.001).

To explore the extent at which participants' beliefs about their expected performance in the task and their likelihood of winning against a random opponent affect their choice to compete, we add these variables to the regression model reported in Table 2. Since the two variables are highly correlated, we do not add both of them to the same model. Column (3) shows that participants' expected performance is not correlated with tournament entry. Adding this variable to the model does not reduce the gender gap in competitiveness. Column (4) instead shows that participant's confidence of winning against a random opponent is significantly correlated with the tournament entry decision. Adding this variable to the model reported in Column (2) reduces the gender gap from 39.9 to 33.4 percentage points. This is consistent with the results observed in the literature (e.g. Niederle and Vesterlund, 2007). While part of the gender gap in tournament entry can be attributed to confidence about the likelihood of winning, a substantial gap between men and women's choices to compete remains.

Risk. To investigate whether participants' risk preferences affected their decision to compete we look at the incentivized risk aversion measure (MPL, Holt & Laury, 2002) as well as the self-reported risk measure. The two measures are not significantly correlated (r=-0.134, p=.238). We do not observe gender differences in the incentivized Risk Aversion measure (Switch point $Mean_{Men}$=6.34, SD=1.81 vs. $Mean_{Women}$ =6.78, SD=1.99, z=-1.022, p=.307, Mann-Whitney). However, we do find a gender difference in the self-reported Risk measure, with men being on average more likely to take risk ($Mean_{Men}$=6.77, SD=1.80 vs. $Mean_{Women}$ =5.48, SD=1.87, z=-3.27, p=.001, Mann-Whitney). When adding the risk measures to the model (columns 5 and 6), we find no significant effect of the risk measures on the tournament dummy and the gender gap does not substantially change.

Ambiguity. Finally, we investigate whether ambiguity attitudes affect participants' tournament entry decision. We do not observe gender differences in ambiguity aversion. The average switching point of men was 11.86 (SD=4.83) while the average switching point of women was 11.62 (SD=4.73, z=.26, p=0.80, Mann-Whitney). Adding this variable to the regression model shows no correlation with the tournament entry decision. Importantly, controlling for ambiguity preferences in addition to risk preferences and confidence leaves the gender gap at 32.7 percentage points.

In Appendix A we report additional regression analyses in which we include demographics controls such as age, ethnicity, first language spoken, academic major, and GPA, and show that the gender gap remains.

Taken together, these results show that competitiveness on the extensive margin is partly explained by confidence, whereas risk and ambiguity preferences do not correlate with tournament entry decisions. Importantly, accounting for such measures leaves a substantial gender gap in tournament entry. In the next section we will present the results of the measure of competitiveness on the intensive margin, and investigate whether this measure provides greater insights on the relationship between competitiveness, confidence and other preferences.

5.3.3 Competitiveness on the Intensive Margin

A total of 84 participants (44 females) took part in the two rounds of data collection for this experiment. Table 3 presents the summary of tournament allocations by gender. We find no difference between the distributions of points allocated in the two rounds of sessions ($z=-0.987$, $p=.32$, Mann-Whitney). Hence, we pool the data for the analyses. Overall, participants allocated an average of 50.11 (SD=28.09 points) points to the tournament. The median allocation was 50 points. The average number of points allocated to the tournament is markedly different for men and women: on average, women allocated 35.27 points to the tournament (SD=21.19) whereas men allocated 66.43 points to the tournament (SD= 25.74). The distribution of points allocated to the tournament differs by gender ($z=-4.99$, $p<.001$, Mann-Whitney).

Table 5.2 Summary Statistics for Competitiveness on the Intensive Margin

Tournament Allocations

	2012	*2013*	*Pooled*	*Min*	*25th Percentile*	*Median*	*75th Percentile*	*Max*
Men	63.8 (27.86)	69.05 (23.86)	66.43 (25.74)	10	50	70	90	100
Women	31.64 (20.39)	38.91 (21.81)	35.27 (21.19)	0	20	35.5	50	80
All	46.95 (28.93)	53.26 (28.81)	50.11 (28.09)	0	30	50	70	100

Importantly, the shapes of the distributions are also different. This fact is evident in the smoothed PDF in Figure 3 and the empirical CDF in Figure 4. The distribution for men is visibly shifted to the right, along the axis of competitiveness, with respect to the distribution for women. The summary statistics provided in Table 3 also depict this difference in distributions; the quartiles calculated for each population are strikingly different. For example, only the most competitive 25 percent of women allocated 50 points or more to the tournament, whereas only the 25 percent least competitive men allocated fewer than 50 points. Only the most competitive woman, who allocated 80 points, allocated more points than the median man (who allocated 70 points).

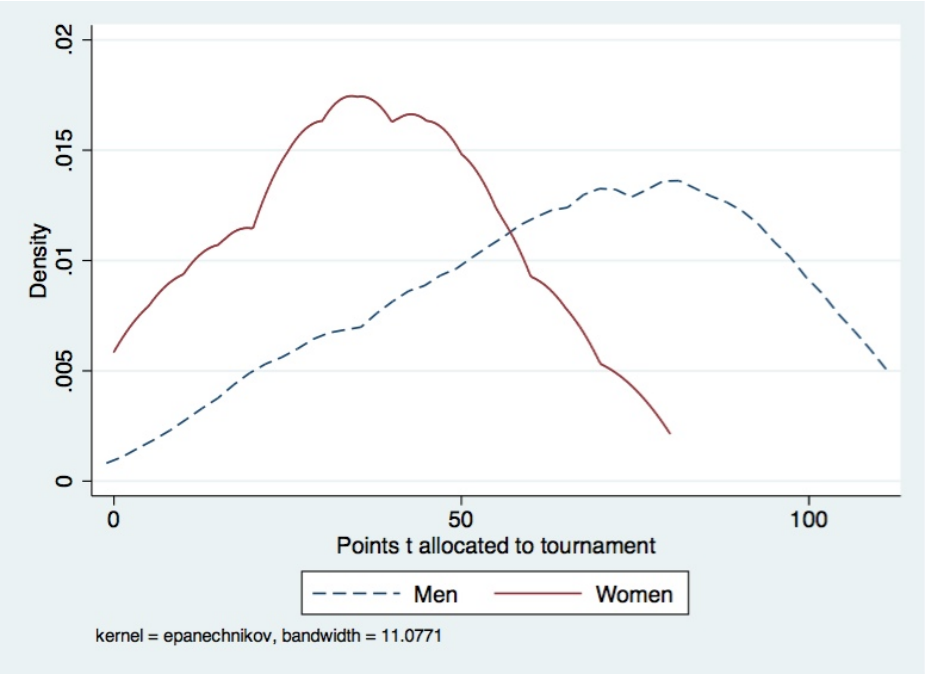

Figure 5.3 Smoothed PDF of Tournament Allocations by Gender

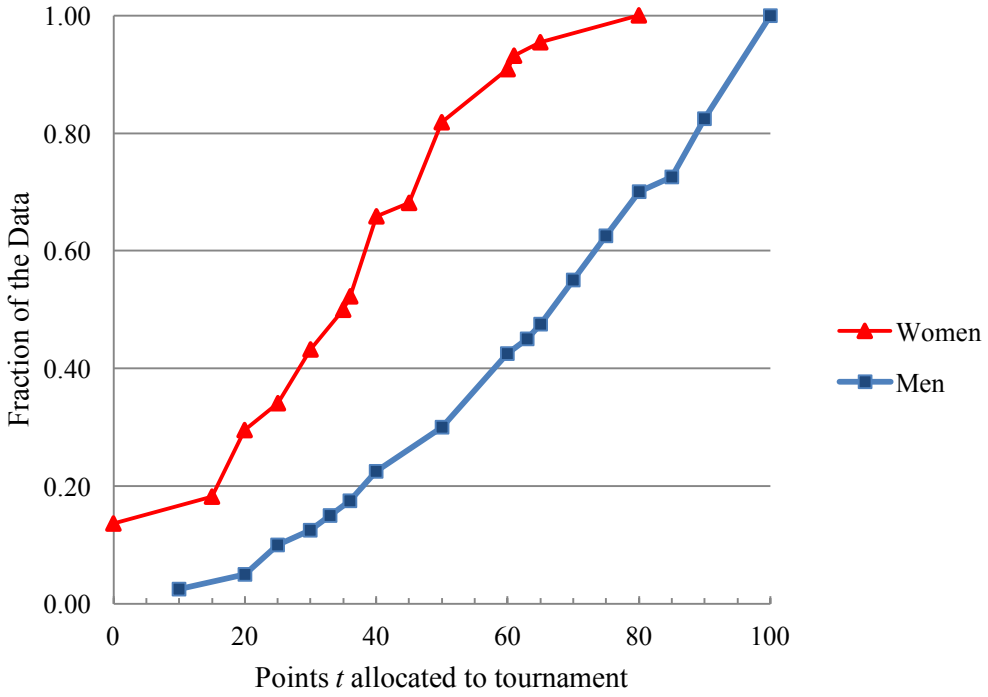

Figure 5.4 Empirical CDF of Tournament Allocations by Gender

Table 4 reports the empirical results from different specifications of an OLS regression where the number of points *t* allocated to the tournament option is regressed on a gender dummy. The first specification of the model reported in column (1) shows that women allocate significantly fewer points to the tournament scheme than men (β=-31.2, $p<.001$). In the experiment, all but two sessions were gender balanced (3 men and 3 women). In the two unbalanced sessions, the fraction of women to men was 4 to 2. When controlling for the gender composition of the session by adding to the regression model a variable indicating the women to men ratio in the sessions (column 2), we find that women still allocate significantly fewer points to the tournament option. The coefficient of the gender composition variable indicates that participants

allocate a larger number of points to the tournament in the sessions with higher fraction of women, though caution is needed in interpreting this result as it is based on only two sessions with more women than men. In addition, we find a non-significant effect of the interaction between gender and gender composition, suggesting that gender composition does not affect women and men differently (β=2.64, p=.852), and hence we do not include the interaction term in the models reported in Table 4.

Table 5.3 OLS Regression of Tournament Allocations

Tournament Allocations	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-31.5***	-32.3***	-28.8***	-17.6***	-14.7**	-12.6**	-16.32***
	(5.17)	(5.17)	(5.42)	(5.86)	(6.00)	(5.79)	(5.82)
Gender composition		14.0**	12.85*	8.79**	4.72	4.84	4.85
		(6.35)	(7.01)	(4.28)	(4.60)	(4.40)	(4.24)
Expected Performance			2.64*				.
			(1.44)				
Confidence Winning				.664***	.533***	.594***	.539***
				(.118)	(.137)	(.136)	(.135)
Self-reported Risk					3.22**	3.16**	3.03*
					(1.51)	(1.52)	(1.64)
Risk Aversion						.533	.567
						(1.09)	(1.08)
Ambiguity aversion							-.809
							(.549)
Constant	77.5***	47.86***	33.91***	1.98	2.32	-6.21	8.40
	(12.5)	(8.66)	(12.69)	(12.21)	(12.63)	(16.93)	(19.85)
Year dummy	N	Y	Y	Y	Y	Y	Y
Observations	84	84	84	84	82	78	77
R^2	.31	.350	.382	.507	.538	.573	.587

*** p<.01, ** p<.05, * p<.10

Note: The table reports OLS estimates. Dependent variable: Points allocated to tournament. Gender composition refers to women to men ratio in the session. Expected performance refers to the estimated number of successful tosses. Confidence of winning refers to subjects' expected likelihood of winning against a random opponent from the same session. Self-reported Risk refers to the self-reported willingness to take risks. Risk aversion refers to the incentivized measure of risk. Ambiguity aversion refers to the incentivized measure of ambiguity. Robust standard errors are reported in parenthesis.

When looking at participants' performance after their allocation decision, we find that on average participants successfully complete 1.90 tosses (SD= 1.65). We find that men perform marginally better than women (z = 1.85, p=.064), though this difference becomes insignificant if we exclude participants who perform more than 5 tosses (5.95% of the participants; z = 1.335, p=.182). Further, if we exclude these participants we still find that women allocate 29.8 fewer points to the tournament than men (p<.001, N=79, see Appendix A).

Determinants of Gender Differences in Competitiveness on the Intensive Margin

Next, we investigate whether confidence, and preferences for risk and ambiguity affect participants' allocation decisions.

Confidence. As in the sample of participants who participated in the extensive margin measure sessions, participants in this sample were overconfident about their task-ability. On average, participants expected to successfully complete 4.90 tosses (SD=1.90), which is bigger than the average number of tosses actually completed by these participants (z= 7.94, p<.001, signrank test). However, men were significantly more confident about their performance than women (Men: average =5.75, SD=1.71, N=40 vs. Women: average = 4.28, SD=1.89, N=44), with the two distributions being statistically different (z=3.27, p<.001, Mann-Whitney). Similarly, men expected chance of winning was 65.15% (SD=16.19), while women's was only 43.70% (SD=17.55, p<.001, Mann-Whitney). The two confidence measures are strongly correlated (r=.52, p<.001).

To investigate whether participants' tournament allocations depend on participants' beliefs about their ability and likelihood of winning, we add these variables to the regression model reported in Table 2. Column (3) shows that the number of points allocated to the tournament is marginally affected by individuals' expected performance. However, adding expected performance to the model does not reduce the gap in female average tournament allocations. Column (4) shows that confidence of winning is positively correlated with tournament allocations. That is, more confident individuals allocate more points to the tournament, regardless of their gender. However, considering a men and a woman with the same level of confidence, a woman allocates less points to the tournament. This result shows that although accounting for differences in confidence between men and women substantially reduces the gap in tournament allocations, confidence of winning cannot explain it entirely.

Risk. Similarly to the binary measure sessions, we find a significant gender difference in the self-reported risk measure. On average, men rated themselves as more willing to take risks than women ($Mean_{Men}$=6.79, SD=1.70 vs. $Mean_{Women}$ =4.78, SD=1.79, MW test, p<.001). We do not observe a gender difference in the incentivized measure of risk (Men's average switch point=7.24, SD=1.91; Women's average switch point=7.21, SD=2.04, p=.964, Mann-Whitney). The two risk measures are not correlated (r=-.023, p=.84).

In column (5) and (6) of Table 4 we explore whether risk preferences affect participants' tournament allocation decisions by adding these variables to the

regression reported in Column (4). Column (5) shows that participants' tournament allocation is partly explained by the self-reported measure of risk. In particular, participants' who describe themselves as more likely to take risks allocate more points to the tournament. However, the gender differences in tournament allocation remains. In column (6) we show that the result is robust to adding to the model the incentivized measure of risk.

Ambiguity. We find a significant difference in ambiguity aversion in this sample, with men being more ambiguity averse than women (Men: average switching point=12.63 (SD=4.99); Women: 9.65 (SD=3.74), p=.007, Mann-Whitney). When adding the ambiguity measure to the regression model, we find that it does not contribute to explaining tournament allocations decisions. Column (7) shows that when controlling for this variable, women still allocate significantly fewer points to the tournament.

Overall, these results show that confidence about winning as well as risk preferences can partly explain participants' allocation decisions. However, when accounting for beliefs and risk preferences, women still allocate significantly less points to the tournament. This result provides additional evidence that participant's taste for competition seems to arise from something different from such preferences—a "competitive spirit". In Appendix A, we report results of regressions in which on top of the variables reported in Table 4, we control for additional demographics, showing that the gender gap in points allocated to the tournament option remains.

Confidence and Risk in the two measures of Competitiveness. Our results also show that, as compared to the extensive margin measure of competitiveness, measuring competitiveness through a linear allocation task provides a finer characterization of the relationship between competitiveness and other economic preferences. In the binary elicitation, our analyses show that participants' confidence predicts tournament entry, even when controlling for gender. With respect to risk, whereas the self-reported measure of willingness to take risk predicts tournament entry decisions in a probit regression of risk on tournament entry and no other control variables by increasing subjects likelihood of entering the tournament by 8 percentage points ($p=.006$), the effect becomes non-significant when controlling for confidence ($p=.363$) or gender $p<.175$). In our sample, measuring competitiveness with the binary choice does not allow us to pick up a relationship between risk preferences and competitiveness. In contrast, the continuous elicitation of competitiveness not only captures the relationship between competitiveness, confidence, and gender, but it also allows us to identify the relationship between (self-reported) risk preferences and competitiveness.

5.3.4 The Gender gap in Competitiveness in the Two Measures

In line with research examining the size of the gender gap in science and in standardized tests (e.g. Fryer and Levitt, 2010; Pope and Sydnor, 2010; Ellison and Swanson, 2010), in this section we examine the women to men ratio across the different percentiles of the distribution of competitiveness of the linear measure and

compare it to the one captured by the binary measure. This allows us to observe whether women are especially underrepresented in certain percentiles of the distribution of competitiveness. To compute the women to men ratio we consider the fraction of women who selected into the tournament out of all the women in the treatment and divide it by the fraction of men who choose the tournament out of all the men in the treatment. This allows us to correct for the unbalance in the number of men and women in the two treatments of our experiment (71 women and 55 men in the treatment with binary measure, and 40 women and 44 men in the treatment with the linear measure) and to directly compare the size of the gender gap between treatments.

In particular, for the binary measure, we compute this ratio by dividing the fraction of women who choose the tournament, $w(T)$ by the fraction of men who choose the tournament, $m(T)$. For the linear measure we calculate this ratio at various percentiles of the distribution of competitiveness, represented by the points t allocated to the tournament. At the different percentiles in the distribution, we consider $w(t)$ the fraction of women who allocate t points or higher to the tournament option and $m(t)$ the fraction of men who allocated t points or higher to the tournament option, and calculate the women-men ratio $w(t) / m(t)$. These ratios are displayed in Table 5. Note that for the linear measure, Table 5 reports the empirical CDF, the fraction of men, women and their ratio in the top x percentile of the distribution of competitiveness. In Appendix A, we report these fractions and ratios across the 4 quartiles of the distribution.

Table 5.4 Women to Men Ratio

	Binary	**Linear**						
	Tournament choice	*Percentile*						
		1th	*10th*	*25th*	*50th*	*75th*	*90th*	*95th*
t to tournament	-	$t \geq 0$	$t \geq 15$	$t \geq 30$	$t \geq 50$	$t \geq 70$	$t \geq 90$	$t = 100$
Fraction of Men	0.78	1	0.98	0.90	0.78	0.53	0.28	0.18
Fraction of Women	0.32	1	0.86	0.66	0.32	0.05	0	0
Women to Men Ratio	**0.41**	**1**	**0.88**	**0.73**	**0.41**	**0.09**	**0**	**0**

Note: The table displays the fraction of Women, the fraction of Men, and the Women to Men ratio in the binary and linear measure.

As shown by Table 5, the binary measure captures a women-to-men ratio of 0.41 to 1, with about a third of the women and a little over two thirds of the men choosing to compete. Is this ratio constant in all the percentiles above a certain cutoff of the distribution of competitiveness in the linear measure? Our analysis reveals that this is not the case; we find that this ratio decreases for increasing degree of competitiveness. Figure 5 provides a graphical representation of the decay in the women-men ratio at the higher percentiles of the distribution of competitiveness. This ratio is plotted as a function of the number of points *t* allocated to the tournament, which are represented on the x-axis by percentile ranks of all the sample of men and women. The 1st percentile of the distribution serves as a benchmark where the women-men ratio is 1, as it is computed on all participants who allocated zero or more points to the tournament. If women and women were equally distributed across the distribution of competitiveness, we should observe this ratio at all percentiles of the

distribution. When focusing on the top 50 percent of the distribution (by looking at this ratio among the participants who allocated a number of points that is greater of equal to the 50^{th} percentile of the distribution of competitiveness), the women to men ratio is 0.41 to 1. However, the gap between the fractions of men and women widens substantially when moving towards the upper tail of the distribution. The women-to-men ratio becomes 0.09 on the 75^{th} percentile and above; it becomes zero at the 90^{th} percentile, as all the participants in the top 10 percent of the distribution are men.

We investigate whether the estimated gender-gap at the median and on the top percentiles of the distribution of competitiveness differ from the estimated gender-gap observed with the binary measure. To compare the observations from the two elicitation of competitiveness, we consider the top-x percentile of all players (men and women) ranked in terms of the linear measure of competitiveness. We classify as "competitive" the participants in the top-x percentile and "non-competitive" the participants who are below the top-x percentile. In the binary measure, we code as "competitive" the participants who chose the tournament and as "non-competitive" the participants who chose the piece-rate. We compare whether the proportion of men and women among the competitive participants are the same across the two measures.

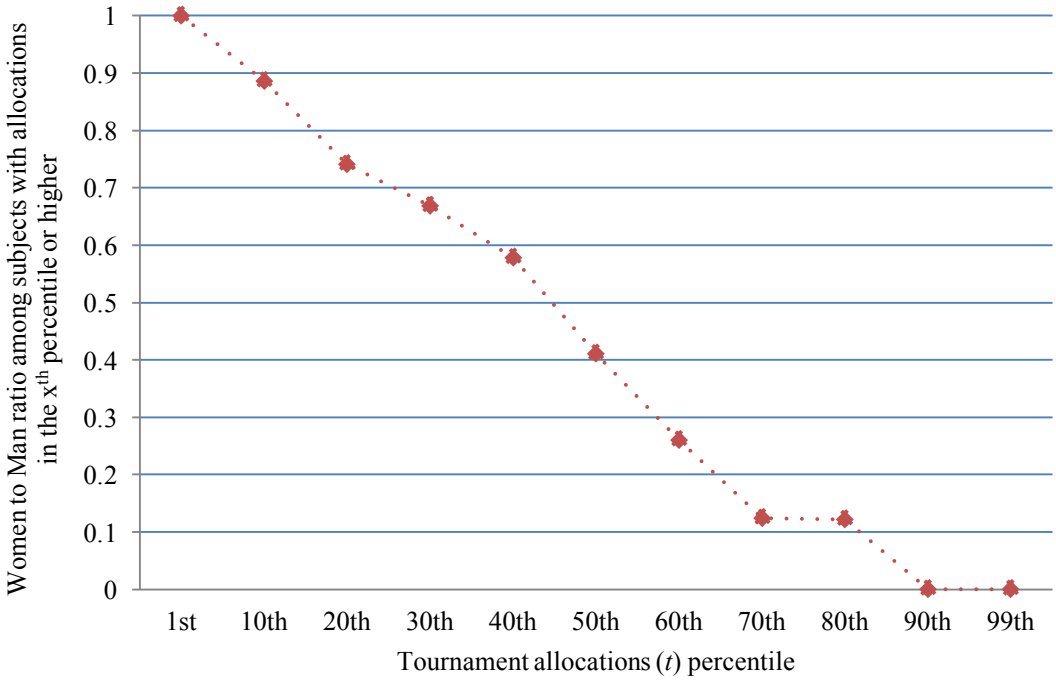

Figure 5.5 Women to Men ratio along the Distribution of Competitiveness

A Fisher exact test reveals that in the top 50th percentile of the distribution of competitiveness, the proportion of men and women does not differ across the two elicitations (p=.83). Indeed, the women to men ratio are similar in the two cases. This result confirms that the gender gap observed at the median of the distribution of competitiveness in the continuous measure is similar to the gap we observe in the binary measure. However, when we compare the fraction of men and women at the 75th percentile of the distribution of competitiveness and above to the proportion of people who choose the tournament when the choice to compete is binary, the two proportions differ (p=.016, Fisher exact). Similarly, the proportion of men and women in the top 10th percentile of the distribution is significantly different than this

proportion in the binary measure (p=.028, Fisher exact). As shown by the smaller women to men ratio displayed in table 5, the gender gap on the right tail of the competitiveness distribution is larger than the gap of the binary measure. It is important to mention, however, that the fractions of men and women are not always balanced in these comparisons. If we run the same tests only for the observations from gender balanced sessions (N=84 in the extensive margin measure and N=72 in the intensive margin one) we find a similar result.

When we compare the fraction of women and men above the 50th percentile of the distribution of tournament allocation with the fraction of men and women who compete in the extensive margin elicitation, we find no difference in the proportions (p=1.0, Fisher exact). However, we find a statistical difference in the proportions if we compare subjects above the 75th percentile of the tournament allocation distribution to those who selected into the tournament in the extensive margin measure (p=.05, Fisher exact). If we repeat the analysis on subjects who are on the top 10th percentile we do not find a statistical difference (p=.10, Fisher exact) though this may be driven by limited power due to the low number of observations; there are no women among these subjects.

Taken together, these results show that while about a third of the women show some level of competitiveness, the most competitive people in our sample are almost all men. The fraction of women among the most competitive participants is smaller than what is captured by a measure that relies on a binary choice. We will use the observation that the proportion of men and women who compete in the binary measure is about the same at the median of our new measure to back up the cutoff c_m below. If

those who are hired for very competitive jobs are drawn from the pool of individuals in the very upper tail of the distribution of competitiveness, then the large gender gap we observe in the real world could also be partly due to the fact that women are largely underrepresented at the top of the distribution. This gap does not need to appear when a person is considered for the job of a CEO; it could start at a much earlier stage in which the person is considering a future career path.

In the next section we explore whether confidence and risk preferences, which are predictive of the number of points allocated to the tournament, can predict who ends up among the most and least competitive participants in the distribution.

Determinants of the top and bottom 25th percentiles of the competitiveness distribution.

In this section, we explore whether confidence and risk preferences can predict whether participants end up in the top (bottom) tail of the distribution of competitiveness based on the number of points allocated to the tournament in the intensive margin measure. For this purpose, we regress these measures on a dummy variable coded as 1 when a participant is in the top (bottom) 25th percentile of the competitiveness distribution. Table 5 reports the results of the probit regressions. Column 1 (of Panel A) shows that participants who were more confident about their likelihood of winning are also more likely to be in the top 25th percentile of the competitiveness distribution. Considering participants' gender in addition to confidence improves the goodness of fit of the model. The estimated marginal effects reported in column 2 shows that both confidence and gender significantly predict

which participants are the most competitive. When controlling for confidence, women are 50 percentage points less likely to be in the top 25% of the distribution. Column 3 and 4 show that on top of confidence and gender, risk preferences are also predictive of the most competitive participants.

Table 5 (Panel B) reports the analyses on the participants who are in the bottom 25th percentile of the competitiveness distribution. Column 1 shows that confident participants are less likely to be among the least competitive participants. Considering participants' gender (column 2) does not improve the fit of the model. While confidence remains predictive of whether participants' are in the bottom 25% of the competitiveness distribution, gender is not. Risk preferences are also not predictive of whether participants are at the bottom of the distribution.

These results show that the gender composition of the top 25th percentile of the competitiveness distribution is not driven only by differences in confidence and attitude toward risk between men and women. In other words, confidence and propensity to risk explain only part of the gender gap among the most competitive individual. On the other hand, the gender composition of the bottom 25th percentile of the competitiveness distribution is exclusively determined by differences in confidence between men and women.

Table 5.5 Determinants of the Most and Least Competitive Participants

	Panel A DV: Top 25th percentile				**Panel B** DV: Bottom 25th percentile			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Female		-.504*** (.188)		-.451** (.190)		-.012 (.091)		-.039 (.106)
Confidence	.015*** (.003)	.016*** (.004)	.011*** (.003)	.013*** (.004)	-.012*** (.003)	-.013** (.005)	-.011*** (.003)	-.013** (.005)
Self reported Risk			.053** (.025)	.066* (.038)			-.022 (.024)	-.028 (.030)
N	84	84	82	82	84		82	82
Pseudo R^2	.32	.42	.35	0.43	.37	.37	.38	.38

*** p<.01, ** p<.05, * p<.10
Note: The table presents marginal effects estimated from probit regression. The dependent variable in Panel A is a dummy variable coded as 1 if participants are in the top 25% of the distribution of tournament allocations, and zero otherwise. The dependent variable in Panel B is a dummy variable coded as 1 if participants are in the bottom 25% of the distribution of tournament allocations, and zero otherwise. Marginal effects are estimated at a man, and at the mean for all the other variables. Robust standard errors are in parenthesis.

The intensive margin measure is more useful in estimating the size of the gender gap in preferences for competition than the extensive margin measure. Since the linear measure provides a finer measure, it allows for a deeper investigation of the relationship between competitiveness and economic preferences. The regression model investigating how the binary choice to compete depends on confidence, risk and ambiguity preferences do not explain much of the decision to enter the tournament. The same model accounts for nearly 60% of the variation in competitiveness in the linear measure, providing a much better fit. Further, since it allows for a finer representation, the finer measure has the methodological advantage that for a given power level fewer observations are needed to correctly reject a null hypothesis.

5.4 Discussion

Competiveness is a personal trait with important economic implications; hence, understanding what affects the tendency to compete is useful for economic analysis. We propose a new method that allows us to obtain a refined measure of individual competitiveness. Previous measures did not allow for variation in levels of competitiveness, and therefore masked the real size of the gender gap. As we argued in the introduction, competitiveness is a multi-dimensional personal trait that cannot be captured by a single design. Elements that could go into it include the reaction to competitive incentives as well as the selection into competitive environments. Future research can use our findings in addition to other findings in the literature in a search for a more comprehensive measure.

With our proposed measure we find the size of the gender gap in competitiveness to be much larger than what previous research has found. These large gender differences in the size of the tendency to compete may have important implications for labor market outcomes. While women may choose to apply for jobs characterized by a moderate degree of competitiveness, they might opt out from highly competitive environments.

Our results provide insights into how to model competitiveness. We find that the distribution of competitiveness *c* is gender dependent. The combination of the model presented in the introduction and the results allows us to calculate the average cutoff points used by men and women when choosing to compete and explore whether c_m also differs by gender. Of the women in the binary treatment, approximately 32

percent chose to enter the tournament. Of the women in the intensive margin treatment, approximately 68 percent allocated to the tournament 45 points or less. If we were to extend the results of the extensive margin measure into the context of the intensive margin measure, around 45 points would be the cutoff point, or decision rule, for women: women who allocate 45 points or more in the linear measure are the type we assume chose to enter the tournament in the extensive margin treatment.

In contrast, about 78 percent of men chose to enter the tournament in our extensive margin measure sample, where 77.5 percent of men allocated more than 40 points to the tournament. Thus, following the same reasoning as for women, the average cutoff point of men is about the same as that of the women: a man allocating 40 points or less to the tournament would likely not choose to enter the tournament.

This suggests that the distribution of c in our sample is gender dependent. Furthermore, women's c_m does not appear to be larger than men's c_m. Women in this comparisons use a similar cutoff when deciding whether to enter the tournament. While this comparison is based on strong assumptions that behavior is directly comparable between the two measures, the conclusion that the gender difference in entering the competition is due to differences in the distribution of competitiveness preferences rather than differences in cutoffs is important.

The difference in the distribution of competitiveness, and in particular the smaller proportion of women with very high c, can help us understand differences in the labor market. We show that the women-men ratio decreases as a function of the degree of competitiveness. The lack of women in many top positions in the labor market is not surprising as among the most competitive individuals, very few are

women. While we acknowledge that several other factors, such as discrimination, contribute to the lack of women in top corporate positions, our results suggests an additional explanation for the extreme lack of women in highly competitive positions.

The results we present in this paper show that men are not simply more likely to opt into competitive environments. The refined measure we propose shows that the magnitude of the gender difference is large. This new measure can also help improving our understanding of the effect of competitiveness on real life outcomes. For example, Buser, Niederle and Oosterbeek (2012) validate the binary measure by showing that it correlates with career choices made by high school students. In their experiment, 51 percent of the men and 77 percent of the women chose not to enter the competition. Hence, the rest of the analysis they perform had to be done based on a minority of participants. A more refined measure could have helped in getting more insight from a larger portion of the participants. Another example is Flory, Leibbrant, and List (2015) who used a field experiment to show that women are less likely to select into competitive work settings relative to men. By varying the role of competition in determining a job wage, they find that women are less likely to apply for jobs in which competitiveness plays a larger role in determining the compensation package.

Future research can use this setup to further our understanding of the underlying causes of the gender gap in labor markets, and the ways to model it, including by using a design in which the same person is performing the same task under different incentives as in Niederle and Vesterlund (2007); see Charness, Gneezy and Kuhn (2012) for a discussion of advantages and disadvantages of such an

approach. Applications of this method could investigate the sources of gender differences in competitiveness observed in our experiment. Differences in competitiveness could arise from differences in these preferences, either directly or indirectly.

Chapter 5, in full, has been submitted for publication of the material as it may appear in Management Science, 2015, Gneezy, Uri, Aniela Pietraz, and Silvia Saccardo. “On the Size of the Gender Difference in Competitiveness.” The dissertation author was the co-primary investigator and author of this paper.

Appendix A. Additional Analyses

A1. Demographics

Demographic Variables and Competitiveness on the Extensive Margin

In this section we investigate whether demographic variables such as age, ethnicity, english native language, GPA, and major affect the gender gap in tournament entry. The descriptive statistics reported in the table reveal that men and women in our sample do not differ in most of the demographic variables. Some differences are that women in our sample were on average, about a year younger than men, were less likely to major in technical fields, and were more likely to be of Asian ethnicity than men, though not significantly so.

Age. Subject's average age was 21.05. The age distribution in this treatment differs across genders (MW, z=1.96, p=.05), though this difference is entirely driven by 4 outliers (all men) who were between 28 and 32 years old. When excluding these outliers, we observe no gender difference in the participants' age distribution (MW, z=1.32, p=0.188). In Table A1 we add demographic controls to the regression results that are reported in Table 2 in the main text. As shown in the table, when we include age as a control variable, we see that the gender gap in tournament entry does not change. Importantly, age does not correlate with tournament entry in any of the regression specifications. Alternatively, if we limit our regression analysis only to participants that are younger than the median age (21), we observe a gender gap of 37 percentage points; if we limit the analysis to individuals who are 21 or older, we observe a gender gap of 34.5 percentage points (analysis available upon request). These results suggest that the gender gap in tournament entry we observed in our data is not driven by a difference in men and women's age.

Ethnicity and language. Out of the 126 participants who participated in the Extensive Margin measure treatment, 53.97% were Asian, 23.02% were White, 11.11% were Hispanic or Latino while the remaining subjects had either a mixed ethnicity (6.35%) or did not indicate their ethnicity in the survey (5.36%). This sample is representative of student population were the experiment was conducted. Further, 79.37% of the participants indicated that English was their native language while the remaining participants (20.63%) indicated another language. Since Asians are the largest ethnic group in our sample, we add to the model a dummy variable coded as 1 if subjects indicated to be of an Asian ethnicity, and zero otherwise. We also control for whether subjects were native English speakers. As shown in Table A1, controlling for these variables in the regression model leaves a substantial gender gap in competitiveness. We also observe that subjects of Asian ethnicity were less likely to select into the tournament (columns 1-2), though this result becomes insignificant when we control for GPA and the other variables such as confidence, risk, and ambiguity aversion (columns 3-9).

Major. Participants came from a variety of departments at UCSD: 23% were students in technical fields such as engineering, computer science or mathematics,

29% majored in biology, chemistry or other sciences, 32.5% majored in social sciences like economics, psychology, sociology or political sciences, and 10% were students in the humanities and arts. A smaller proportion of women came from technical fields (p=.03, Fisher exact). In Table A1 (column 3), we add dummy variables to the regression model in order to control for participants' major. The result shows that including such variables does not substantially affect the size of the gender gap in competitiveness. None of these dummy variables are predictive of tournament entry.

GPA. We find no gender differences in GPA between subjects in our sample (z=-.36, p=.718)[8]. Further, GPA is not correlated with tournament entry nor it affects the gender gap if included to the regression model (column 3-8).

[8] Note that GPA was only collected for the 84 individuals from whom we collected the measures of confidence, risk and ambiguity.

Table 5.A1 Probit of Tournament Entry

Choice of Tournament	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-.351*** (.105)	-.449*** (.113)	-.432*** (.153)	-.432*** (.1154)	-.373*** (.124)	-.363*** (.128)	-.373*** (.142)	-.404***
Gender composition	.005 (.030)	.002 (.036)	.001 (.034)	.001 (.034)	-.014 (.046)	-.019 (.048)	-.027 (.051)	-.016 (.050
Expected Performance				-.002 (.026)				
Confidence Winning					.010** (.005)	.010* (.005)	.011** (.005)	.011**
Self reported Risk						.015 (.038)	.016 (.038)	.016 (.038
Risk Aversion							.057 (.036)	.063 (.039
Ambiguity Aversion								-.022*
Age	-.000 (.017)	.001 (.021)	.009 (.025)	-.008 (.023)	-.001 (.030)	-.001 (.030)	-.001 (.031)	-.005 (.029
Asian	-.139** (.068)	-.148* (.083)	-.130 (.086)	-.128 (.085)	-.162 (.115)	-.166 (.115)	-.153 (.121)	-.168 (.118
Non-native speaker	.037 (.095)	.020 (.115)	-.045 (.117)	-.047 (.112)	.018 (.174)	.054 (.188)	.076 (.196)	.030 (.192
GPA			-.053 (.105)	-.054 (.106)	-.065 (.122)	-.049 (.125)	-.100 (.133)	-.103 (.137
Major	N	Y	Y	Y	Y	Y	Y	Y
Year dummy	Y	Y	Y	Y	Y	Y	Y	Y
Observations	119	119	81	81	81	80	77	77

*** p<.01, ** p<.05, * p<.10

Note: The table presents marginal effects estimated from probit regression. Dependent variable: Choice of tournament (1 tournament and 0 piece-rate). Gender composition refers to the women to men ratio in each session. Expected performance refers to the estimated number of successful tosses. Confidence of winning refers to subjects' expected likelihood of winning against a random opponent from the same session. Self-reported Risk refers to the self-reported willingness to take risks. Risk aversion refers to the incentivized measure of risk. Ambiguity aversion refers to the incentivized measure of ambiguity. Asian is a dummy variable coded as zero if subjects were of Asian ethnicity and zero otherwise. Non-native speaker is dummy variable coded as 1 if subjects were not English native, and zero otherwise. Major-dummies are dummy variables for the following majors Engineering and Math, Social Science, Literature and Art, with the Science major as a baseline. Marginal effects are estimated at a white English native man from a gender-balanced session, majoring in science. Robust standard errors are in parenthesis.

Demographic Variables and Competitiveness on the Intensive Margin

Next, we investigate whether controlling for demographic variables reported in Panel B of Table A2 affects the allocations of points to the tournament in the Intensive Margin treatment. The descriptive statistics reported in the table reveal that men and women in our sample do not differ in most of the demographics However, women in this sample were more likely not to be English native speakers, less likely to major in technical fields and more likely to major in the social sciences. Further, their average GPA that was higher than that of men.

Age. We find no significant differences in men and women's age distributions in this sample. In Table A2, we report the regression results from the main text and add the demographic controls to the model. We find that age does not correlate with tournament allocation in any of the regression specifications.

Ethnicity and language. As shown in Table A1, out of the 84 participants who participated in binary version of the experiment, 69.05% were Asian, 11.90% were White, 11.90% were Hispanic or Latino while the remaining subjects had either a mixed ethnicity (3.6 %) or did not indicate their ethnicity in the survey (3.6%). The mix of ethnicities is similar to the one detected in the binary measure. We do not find differences in the proportion of men and women across ethnicities. Further, 46.43% of our subjects were not English natives. Of those, the majority were women (p=.02, Fisher Exact). In table A3 we report the results of OLS regressions in which we include a dummy variable indicating whether subjects were of Asian ethnicity, as well as a control dummy variable for whether subjects were English natives. None of the variables correlate with tournament allocation. Importantly, controlling for these variables does not substantially affect the gender gap in tournament allocation.

Major. Participants came from a variety of departments at UCSD: 21% were students in technical fields such as engineering, computer science or mathematics, 27% majored in biology, chemistry or other sciences, 45% majored in social sciences, and about 6% were students in the humanities and arts. In Table A3 (column 2), we add dummy variables to the regression model in order to control for participants' major. The result shows that including such variables does not substantially affect the size of the gender gap in competitiveness. None of these variables correlates with tournament allocation.

GPA. We find a significant gender differences in GPA between subjects in our sample (z=-2.159, p=.031, Mann-Whitney) with women having a higher GPA than men. In Table A3 we explore the relationship between GPA and tournament allocation (columns 3-8). We find that GPA is negatively correlated with tournament entry. That is, when controlling for gender and other demographics, people with lower GPA allocated less points to the tournament. However, the female coefficient remains large and significant in all regressions specifications.

Table 5.A2 OLS Regression of Tournament Allocation

Points allocated to tournament	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fmale	-33.57***	-36.34***	-31.70***	-28.48***	-176.96**	-15.72**	-12.95*	-16.04**
	(5.69)	(6.16)	(6.47)	(6.54)	(6.97)	(7.21)	(7.08)	(7.31)
Gender composition	13.92**	13.50**	13.44**	11.98**	8.12**	4.99	4.74	4.20
	(6.03)	(5.70)	(5.45)	(5.82)	(3.91)	(4.76)	(4.74)	(4.76)
Expected Performance				2.76**		.		
				(1.23)				
Confidence Winning					.614***	.518***	.564***	.511***
					(.122)	(.142)	(.145)	(.146)
Self reported Risk						2.41	2.56	2.79
						(1.79)	(1.83)	(1.97)
Risk Aversion							-.130	-.123
							(1.21)	(1.22)
Ambiguity aversion								-.702
								(.613)
Age	1.35	1.65	1.73	1.86	1.43	1.19	1.31	1.70
	(1.40)	(1.47)	(1.55)	(1.50)	(1.24)	(1.26)	(1.26)	(1.36)
Asian	-7.31	-7.60	-8.49	-8.05	-5.50	-5.81	-5.73	-5.51
	(5.73)	(5.87)	(5.80)	(5.50)	(4.76)	(4.79)	(4.96)	(4.83)
Non-native speaker	-.191	1.53	-.251	.617	.722	5.76	1.64	1.22
	(5.77)	(5.62)	(5.69)	(5.46)	(4.77)	(4.71)	(4.84)	(5.13)
GPA			-16.62**	-16.95**	-14.64**	-14.07*	-12.54**	-11.65*
			(6.95)	(6.59)	(5.85)	(5.72)		(6.31)
Constant	28.97	28.17	78.03*	63.50	41.05	36.19	20.55	21.22
	(29.63)	(30.14)	(40.13)	(38.41)	(32.59)	(34.18)	(34.87)	(39.80)
Major controls	Y	Y	Y	Y	Y	Y	Y	Y
Year dummy	N	Y	Y	Y	Y	Y	Y	Y
Observations	81	81	78	78	78	78	74	73
R^2	.372	.394	.430	.461	0.561	0.573	0.590	.602

*** p<.01, ** p<.05, * p<.10

Note: The table presents OLS estimates. Dependent variable: points allocated to the tournament. Gender composition refers to the women to men ratio in each session. Expected performance refers to the estimated number of successful tosses. Confidence of winning refers to subjects' expected likelihood of winning against a random opponent from the same session. Self-reported risk refers to the self-reported willingness to take risks. Risk aversion refers to the incentivized measure of risk. Ambiguity aversion refers to the incentivized measure of ambiguity. Asian is a dummy variable coded as zero if subjects were of Asian ethnicity and zero otherwise. Non-native speaker is dummy variable coded as 1 if subjects were not English native, and zero otherwise. Major-dummies are dummy variables for the following majors Engineering and Math, Social Science, Literature and Art, with Science as the baseline. Robust standard errors are in parenthesis.

A2. Analysis on the restricted sample of subjects from gender balanced sessions.

Table 5.A3 and Table 5.A4 report the regression results for the sample of participants from gender balanced sessions for the extensive (N=84) and intensive margin (N=72) of competitiveness respectively. The regressions show that the results

for these participants are in line with the analyses we report in the main text, where we control for the women to men ratio.

Table 5.A3. Probit of Tournament Entry

Choice of Tournament	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-.366*** ((.056)	-.365*** ((.056)	-.351*** (.073)	-.304*** .(072)	-.300*** (.073)	-.296*** (.072)	-.294*** (.071)
Expected Performance			-.003* (.021)			.	
Confidence Winning				.006** (.003)	.008** (.003)	008*** (.003)	.008*** .003
Self reported Risk					-.021 (.035)	.025 (.034)	-.018 (.035)
Risk Aversion						.051** (.024)	.049** (.024)
Ambiguity Aversion							-.010 (.010)
Year dummy	N	Y	Y	Y	Y	Y	Y
Observations	84	84	60	60	59	59	59
Pseudo R2	.217	.228	.275	.344	.343	.391	.398

*** p<.01, ** p<.05, * p<.10

Note: The table presents marginal effects estimated from probit regression for the restricted sample of subjects from gender balanced sessions. Dependent variable: Choice of tournament (1 tournament and 0 piece-rate). Confidence of winning refers to subjects' expected likelihood of winning against a random opponent from the same session. Self-reported Risk refers to the self-reported willingness to take risks. Risk aversion refers to the incentivized measure of risk. Ambiguity aversion refers to the incentivized measure of ambiguity. Marginal effects are estimated at a man, and at the average for all the other variables. Robust standard errors are in parenthesis.

Table 5.A4 OLS Regression of Tournament Allocation

Points allocated to tournament	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-32.64***	-32.63***	-28.46***	-18.8***	-15.39**	-13.01*	-16.12**
	(5.57)	(5.53)	(5.53)	(6.70)	(6.99)	(6.68)	(6.93)
Expected Performance			4.01***				
			(1.33)				
Confidence Winning				.612***	.489***	.555***	.514***
				(.135)	(.151)	(.149)	(.149)
Self-reported Risk					3.24**	3.20*	3.12*
					(1.61)	(1.61)	(1.75)
Risk Aversion						.353	.419
						(1.26)	(1.24)
Ambiguity Aversion							-.649
							(.597)
Constant	65.19***	61.24***	38.88***	22.67**	9.11	1.76	13.09
	(4.29)	(5.17)	(8.81)	(10.86)	(13.74)	(18.29)	(22.04)
Year dummy	N	Y	Y	Y	Y	Y	Y
Observations	72	72	72	72	70	66	65

*** p<.01, ** p<.05, * p<.10

Note: The table presents OLS estimates for the restricted sample of subjects from gender balanced sessions. Dependent variable: points allocated to the tournament. Expected performance refers to the estimated number of successful tosses. Confidence of winning refers to subjects' expected likelihood of winning against a random opponent from the same session. Self-reported risk refers to the self-reported willingness to take risks. Risk aversion refers to the incentivized measure of risk. Ambiguity aversion refers to the incentivized measure of ambiguity. Robust standard errors are in parenthesis.

A3. Analysis on the restricted sample of participants with no difference in performance

As we report in the main text, in the Extensive Margin treatment we find that men perform better than women under the tournament (MW test, z=2.277, p=.02) and marginally better than women under the piece rate (z=1.717, p=.09). Both results become insignificant if we exclude the top 11.6% of the observations who perform 5 or more successful tosses (MW test, z =1.438, p=.15 for the tournament; MW test, z = 1.289, p=.20 for the piece rate). Importantly, if we exclude these observations and regress tournament on female we find that females are still 35.5 percentage points less likely to select the tournament than men. The regression results are illustrated in Table A6 below. Similarly, in the Intensive Margin treatment we find that men perform marginally better than women (z=1.85, p=.064), though this difference becomes

insignificant if we exclude subjects (5.95%) who perform more than 5 tosses (z = 1.335, p=.182). Further, if we exclude these subjects we still find that women allocate 29.8 fewer points to the tournament than men. The regression results are reported in the table below.

Table 5.A5 Probit of Tournament Entry

Choice of Tournament	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-.355*** (.051)	-.351*** (.052)	-.344*** (.056)	-.301** (.069)	-.297*** (.068)	-.288*** (.064)	-.303*** (.052)
Gender composition		-.002 (.039)	-.005 (.035)	-.012 (.044)	-.018 (.045)	-.018 (.044)	-.002 (.037)
Expected Performance			.007 (.023)				
Confidence Winning				.007** (.003)	.007*** (.003)	.008*** .003	.007*** (.002)
Self reported Risk					-.017 (.026)	-.017 (.025)	.022 (.022)
Risk Aversion						.030 (.022)	.030 (.020)
Ambiguity Aversion							-.020** (.008)
Year dummy	N	Y	Y	Y	Y	Y	Y
Observations	106	106	72	72	71	69	69
Pseudo R2	.136	.150	.229	.300	.301	.310	.355

*** p<.01, ** p<.05, * p<.10

Note: The table presents marginal effects estimated from probit regression for the restricted sample of subjects with no differences in performance. Dependent variable: Choice of tournament (1 tournament and 0 piece-rate). Gender composition refers to the women to men ratio in each session. Expected performance refers to the estimated number of successful tosses. Confidence of winning refers to subjects' expected likelihood of winning against a random opponent from the same session. Self-reported Risk refers to the self-reported willingness to take risks. Risk aversion refers to the incentivized measure of risk. Ambiguity aversion refers to the incentivized measure of ambiguity. Marginal effects are estimated at a man from a gender-balanced session, and at the average for all the other variables. Robust standard errors are in parenthesis.

Table 5.A6 OLS Regression of Tournament Allocation

Points allocated to [illegible]	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-29.78***	-31.09***	-27.89***	-17.03***	-13.44**	-11.25*	-14.81**
	(5.51)	(5.51)	(5.83)	(6.13)	(6.21)	(5.88)	(6.24)
Gender composition		14.50**	13.33*	8.99	4.62	4.60	4.63
		(6.35)	(6.93)	(4.35)	(4.67)	(4.43)	(4.35)
Expected Performance			2.45		.		
			(1.51)				
Confidence Winning				.664***	.528***	.590***	.537***
				(.123)	(.140)	(.137)	(.138)
Self-reported Risk					3.48**	3.44**	3.27*
					(1.51)	(1.51)	(1.66)
Risk Aversion						.424	.506
						(1.09)	(1.08)
Ambiguity aversion							-.722
							(.582)
Constant	65.06***	46.67***	33.84**	10.43	.295	-7.26	6.03
	(4.43)	(8.78)	(12.79)	(11.09)	(12.57)	(17.09)	(20.80)
Year dummy	N	Y	Y	Y	Y	Y	Y
Observations	79	79	79	79	78	74	73
R^2	.282	.326	.350	.479	.519	.555	.566

*** p<.01, ** p<.05, * p<.10

Note: The table presents OLS estimates for the restricted sample of subjects with no difference in performance. Dependent variable: points allocated to the tournament. Gender composition refers to the women to men ratio in each session. Expected performance refers to the estimated number of successful tosses. Confidence of winning refers to subjects' expected likelihood of winning against a random opponent from the same session. Self-reported risk refers to the self-reported willingness to take risks. Risk aversion refers to the incentivized measure of risk. Ambiguity aversion refers to the incentivized measure of ambiguity. Robust standard errors are in parenthesis.

A4. The Gender Gap across the quartiles of the distribution of tournament allocation

Table 5.A7 Women to Men ratio across the quartiles of the distribution of tournament allocation

	Extensive Margin	***Intensive Margin***			
	Tournament choice	*1^{st}-25^{th}*	*25^{th}-49^{th}*	*50^{th}-75^{th}*	*75^{th}-99^{th}*
t to tournament	-	$t \leq 25$	$25 < t < 50$	$50 \leq t < 70$	$t \geq 70$
Fraction of Men	0.78	.10	0.13	0.25	0.53
Fraction of Women	0.32	.34	0.34	0.27	0.05
Women to Men Ratio	**0.41**	**3.4**	**2.62**	**1.08**	**0.09**

The Table illustrates the fraction of men and women across the four quartiles of the distribution of competitiveness. As shown in the table, we see high women to men ratios in the first two quartiles of the distribution. Only 23% of the men are below the median of the distribution. In the third quartiles of the distribution we see that the fraction of men and the fraction of women are about the same, with a women to men ratio of 1.08 to 1. Instead, we see that more than half of the men (53%) are in the top quartile of the distribution, whereas only 5% of the women are in this quartile. On this quartile of the distribution the women to men ratio is 0.09. Overall, we observe a higher proportion of women in the first and second quartiles of the distribution ($p=.01$ and $p=.02$, Fisher exact), while a higher proportion of men in the top quartile ($p<.001$, Fisher exact).

Appendix B. Instructions

Ability

Welcome to the experiment. The experiment is simple. We will keep anonymous any and all information that we receive from you during this session. Please read the following instructions carefully.

Thank you for your participation. In this study, you will be asked to participate in a ball-tossing task.

In this task, you will toss a tennis ball into a small bin 10 feet away. You will have 10 opportunities to toss the ball. The toss must be underhand. A successful toss is a toss that lands in the bin and stays in the bin. The task itself will be completed at the end of the session, and we will record your results individually; that is, no other participants will observe your performance in this task.

Competitiveness on the Extensive Margin Measure

Welcome to the experiment. The experiment is simple and if you will follow the instructions you may earn a considerable amount of money that will be paid to you, privately and in cash, at the end of the experiment. We will keep anonymous any and all information that we receive from you during this session. Please read the following instructions carefully.

Thank you for your participation. In this study, you will be asked to participate in a ball-tossing task, and to choose how you would like to be compensated.

In this task, you will toss a tennis ball into a small bin 10 feet away. You will have 10 opportunities to toss the ball and you will be paid according to your success. The toss must be underhand. A successful toss is a toss that lands in the bin and stays in the bin. The task itself will be completed at the end of the session, and we will record your results individually; that is, no other participants will observe your performance in this task.

You are now asked to select how you would like to be paid for the completion of this task by filling out the bottom of this sheet. This sheet will be collected before the start of the ball-tossing task.

You may choose to be paid by piece-rate or by participating in a tournament. You may choose only one.

A. The piece-rate option pays $1.00 for each successful ball toss

B. The tournament option pays $3.00 for each successful ball toss if you have more successful tosses than a randomly chosen participant in the room. If you have fewer successful tosses than that participant, you will be paid zero for this part of the experiment. In case the two of you tie, you will be paid $1.00 for each successful toss.

Please select how you would like to be paid now, by circling the desired option below.

I would like to be paid by:

Piece-Rate Tournament

Competitiveness on the Intensive Margin Measure

Welcome to the experiment. This experiment is simple and if you follow the instructions you may earn a considerable amount of money that will be paid to you, privately and in cash, at the end of the experiment. We will keep anonymous any and all information that we receive from you during this session. Please read the following instructions carefully.

Thank you for your participation. In this study, you will be asked to participate in a ball-tossing task, and to choose how you would like to be compensated.

In this task, you will toss a tennis ball into a small bin 10 feet away. The toss must be underhand. You will have 10 opportunities to toss the ball and you will be paid according to your success. A successful toss is a toss that lands in the bin and stays in the bin. The task itself will be completed at the end of the session, and we will record your results individually; that is, no other participants will observe your performance in this task.

The decision you are asked to make is how you would like to be paid for the completion of this task. You will make this decision before the start of the ball-tossing task.

You are endowed with 100 points and asked to choose the portion of this amount (between 0 and 100 points, inclusive) that you wish to invest in a tournament. The rest of the 100 points will be invested in a piece-rate compensation scheme.

The payments for each point invested in the options are as follows:

A. The piece-rate option pays 1 cent per point for each successful toss

B. The tournament option pays 3 cents per point for each successful toss if you have more successful tosses than a randomly chosen participant in the room. If you have fewer successful tosses than that participant, you will be paid zero for this part of the experiment. In case the two of you tie, you will be paid 1cent per point for each successful ball toss.

We now ask you to choose how many of the 100 points you would like to invest in option A (the piece-rate) and how many in option B (the tournament). Please remember that the two numbers should add up to 100 points.

I would like to invest:

_____ points in option A _____points in option B

Additional measures –Instructions
Confidence Questionnaire

Please complete the following questions. Raise your hand when you are finished. This form will be collected prior to the start of the timed task.

How many successful tosses do you think you will make?

_____

What do you believe is the probability that you will have more successful tosses than a randomly selected opponent? Please give a percentage from 0-100.

_____

Risk Assessment

Please answer the following question using a 1-10 scale, where *1=completely unwilling* and *10=completely willing*:

Rate your willingness to take risks in general: _____

Incentivized Risk Aversion Elicitation

Instructions for Task H

In addition to the Instructions, this envelope contains a Decision Sheet. Please look on to your Decision Sheet as you read these Instructions to ensure that you understand the procedure of the experiment. If you have a question at any point, please raise your hand.

The Decision Sheet contains 10 separate Decisions numbering 1 through 10. Each of these Decisions is a choice between "Option A" and "Option B". One of these decisions will be randomly selected to determine your earnings. A ten-sided die will be used to determine the payoffs. After you have made your choice, this die will be rolled twice: once to select one of the 10 Decisions to be used, and then again to determine your payoff for the Option associated with that decision, either A or B, given your choice at that decision.

To choose an Option for each decision, you will make one choice in the "Switch" column on the right. This choice indicates that you would like to switch from Option A to Option B, and will signify whether Option A or Option B is used to determine your earnings for each of the 10 decisions. For each decision **before** your choice, Option A will be used for payment. For each decision **after** your choice, including the decision where the choice was made, Option B will be used.

For example, if the die roll outcome is 6, Decision No. 6 will determine payment.

1. If your "Switch" number is **after** Decision No. 6, then Option A be used to determine your payoff. You will have a 6/10 chance of earning 200 tokens, and a 4/10 chance of earning 160 tokens.
2. If your "Switch" number is **before or at** Decision No. 6, then Option B will be used to determine your payoff. You will have a 6/10 chance of earning 385 tokens, and a 4/10 chance of earning 10 tokens.

Namely, once we select a decision to determine your earnings, if that decision came **before** your choice to switch, Option A will be used. If that decision came **after** or **at** your choice, Option B will be used.

Please look at Decision 3 at the top of the Decision Sheet. Option A pays 200 tokens with a chance of 3/10, and 160 tokens with a chance of 7/10. Since each side of a ten-sided die has an equal chance of being the outcome in a throw, this corresponds to Option A paying 200 tokens if the throw of the die is 1, 2 or 3, and 160 tokens if the throw of the die is any other number (4 through 10). Option B pays 385 tokens if throw of the die is 1, 2 or 3, and 10 tokens if the throw of the die is any other number (4 through 10). The other Decisions are similar, except that as you go down the table, the chances of the higher payoff for each Option increase. For Decision 10 in the bottom row, no die will be needed since each Option pays the highest payoff for sure. Your choice there is between 200 tokens and 385 tokens.

Once you are done with both tasks H and T, you will proceed to another room where an experimenter will flip a coin. If the outcome is Heads, the experimenter will throw a ten-sided die to select which of the ten Decisions will be used. The die will then be thrown again to determine your earnings for the Option you chose for the selected Decision. Earnings in tokens will be converted to dollars such that 20 tokens = $1, so if your payoff was 200 tokens you would earn $10. This will be added to your previous earnings, and you will be paid in cash when finished.

Please raise your hand if you have any questions. If you do not have any questions, please proceed to the Decision Sheet and mark your choices.

Decision Sheet

Please indicate at which decision you would like to switch from Option A to Option B by putting a check mark in the box of the Switch column. You should have 1 check mark. For any decisions before this check mark, Option A will be used to determine payment. For any decisions after and including the check mark, Option B will be used.

NO.	Option A	Option B	Switch
1	1/10 chance of 200 tokens 9/10 chance of 160 tokens	1/10 chance of 385 tokens 9/10 chance of 10 tokens	
2	2/10 chance of 200 tokens 8/10 chance of 160 tokens	2/10 chance of 385 tokens 8/10 chance of 10 tokens	
3	3/10 chance of 200 tokens 7/10 chance of 160 tokens	3/10 chance of 385 tokens 7/10 chance of 10 tokens	
4	4/10 chance of 200 tokens 6/10 chance of 160 tokens	4/10 chance of 385 tokens 6/10 chance of 10 tokens	
5	5/10 chance of 200 tokens 5/10 chance of 160 tokens	5/10 chance of 385 tokens 5/10 chance of 10 tokens	
6	6/10 chance of 200 tokens 4/10 chance of 160 tokens	6/10 chance of 385 tokens 4/10 chance of 10 tokens	
7	7/10 chance of 200 tokens 3/10 chance of 160 tokens	7/10 chance of 385 tokens 3/10 chance of 10 tokens	
8	8/10 chance of 200 tokens 2/10 chance of 160 tokens	8/10 chance of 385 tokens 2/10 chance of 10 tokens	
9	9/10 chance of 200 tokens 1/10 chance of 160 tokens	9/10 chance of 385 tokens 1/10 chance of 10 tokens	
10	10/10 chance of 200 tokens 0/10 chance of 160 tokens	10/10 chance of 385 tokens 0/10 chance of 10 tokens	

Incentivized Ambiguity Aversion Elicitation

Instructions for Task T

In addition to the Instructions, this envelope contains a Decision Sheet. Please look on to your Decision Sheet as you read these Instructions to ensure that you understand the procedure of the experiment. If you have a question at any point, please raise your hand.

The Decision Sheet contains 20 separate Decisions numbering 1 through 20. Each of these Decisions is a choice between drawing a ball from "Urn A" or "Urn B". One of these decisions will be randomly selected, depending upon a roll of a one 20-sided die, to determine your earnings. You will select a color, Red or Black, and this will be your **Success Color**. Once a decision is selected, your earnings will be determined by whether the ball drawn from the Urn matches your Success Color.

You will make one choice in the "Switch" column on the right. This choice indicates that you would like to switch from drawing a ball out of Urn A to drawing out of Urn B. Making a choice to switch means that every decision **before** your choice, a ball will be drawn from Urn A. For each decision **after** your decision, including the decision where the choice was made, a ball will be drawn from Urn B.

For example, if the die roll is 9, Decision No. 9 will determine payment.

1. If your "Switch" number is **after** Decision No. 9, a ball will be drawn from Urn A, and if the color of the ball matches the Success Color, then you will 200 tokens. If it does not match, you will earn 0 tokens.
2. If your "Switch" number is **before or at** Decision No. 9, a ball will be drawn from Urn B, and if the color of the ball matches the chosen Success Color, then you will earn 228 tokens. If it does not match, you will earn 0 tokens.

Namely, once we select a decision to determine your earnings, if that decision came **before** your choice to switch, a ball will be drawn from Urn A. If that decision came **after** or **at** your choice, a ball will be drawn from Urn B.

In each of the 20 decisions, Urn A has 50 Red balls and 50 black balls, and pays 200 tokens if the ball drawn from Urn A matches your Success Color, and 0 tokens if it does not match. Since each color has a ½ chance of being drawn, this means that drawing from Urn A pays 200 tokens with a chance of ½ , and pays 0 with a chance of ½.

Urn B, on the other hand, has an unknown number of Red and Black balls (with a total of 100 balls). It pays a positive amount if the ball drawn from Urn B matches your Success Color, and 0 tokens if it does not match. Since the chance of each color being drawn is unknown, the chance of Urn B paying a positive number of tokens is

unknown as well. The only difference between the 20 options is the amount paid when a ball matching your Success Color is drawn from Urn B.

When you have made your choice to switch, please place these instructions and your Decision Sheet back into the envelope marked T. Once you are done with both tasks H and T, you will proceed to another room where an experimenter will flip a coin. If the outcome is Tails, the experimenter will throw one 20-sided die to select which of the 20 decisions will be used. The experimenter will then draw a ball from the Urn you had selected for that Decision to determine your payoff. Earnings in tokens will be converted to dollars such that 20 tokens = $1, so if your payoff was 200 tokens you would earn $10. You will then be paid in cash.

Please raise your hand if you have any questions. If you do not have any questions, please proceed to the Decision Sheet and mark your choices.

Decision Sheet

My Success Color is (please circle one): **Red** **Black**

Please indicate at which decision you would like to switch from Urn A to Urn B by putting a check mark in the box of the Switch column. You should have 1 check mark total. For any decisions before this check mark, a ball will be drawn from Urn A. For any decisions after and including the check mark, a ball will be drawn from Urn B

	Urn A	**Urn B**	
No.	**50 Red balls, 50 Black balls**	**? Red balls, ? Black balls**	**Switch**
1	200 tokens if Chosen Color 0 tokens if not	164 tokens if Chosen Color 0 tokens if not	
2	200 tokens if Chosen Color 0 tokens if not	172 tokens if Chosen Color 0 tokens if not	
3	200 tokens if Chosen Color 0 tokens if not	180 tokens if Chosen Color 0 tokens if not	
4	200 tokens if Chosen Color 0 tokens if not	188 tokens if Chosen Color 0 tokens if not	
5	200 tokens if Chosen Color 0 tokens if not	196 tokens if Chosen Color 0 tokens if not	
6	200 tokens if Chosen Color 0 tokens if not	204 tokens if Chosen Color 0 tokens if not	
7	200 tokens if Chosen Color 0 tokens if not	212 tokens if Chosen Color 0 tokens if not	
8	200 tokens if Chosen Color 0 tokens if not	220 tokens if Chosen Color 0 tokens if not	
9	200 tokens if Chosen Color 0 tokens if not	228 tokens if Chosen Color 0 tokens if not	
10	200 tokens if Chosen Color 0 tokens if not	236 tokens if Chosen Color 0 tokens if not	
11	200 tokens if Chosen Color 0 tokens if not	244 tokens if Chosen Color 0 tokens if not	
12	200 tokens if Chosen Color 0 tokens if not	252 tokens if Chosen Color 0 tokens if not	
13	200 tokens if Chosen Color 0 tokens if not	260 tokens if Chosen Color 0 tokens if not	
14	200 tokens if Chosen Color 0 tokens if not	268 tokens if Chosen Color 0 tokens if not	
15	200 tokens if Chosen Color 0 tokens if not	276 tokens if Chosen Color 0 tokens if not	
16	200 tokens if Chosen Color 0 tokens if not	284 tokens if Chosen Color 0 tokens if not	
17	200 tokens if Chosen Color 0 tokens if not	292 tokens if Chosen Color 0 tokens if not	
18	200 tokens if Chosen Color 0 tokens if not	300 tokens if Chosen Color 0 tokens if not	
19	200 tokens if Chosen Color 0 tokens if not	308 tokens if Chosen Color 0 tokens if not	
20	200 tokens if Chosen Color 0 tokens if not	316 tokens if Chosen Color 0 tokens if not	

References

Altonji, Joseph G., and Rebecca M. Blank. 1999. "Race and Gender in the Labor Market. In *Handbook of Labor Economics*. Vol. Volume 3, Part C, Elsevier, 3143-3259.

Andersen, Steffen, Seda Ertac, Uri Gneezy, John A. List, and Sandra Maximiano. 2013. Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society. *Review of Economics and Statistics* 95 (4): 1438-43.

Babcock, Linda, and Sara Laschever. 2003. *Women don't ask: Negotiation and the gender divide.* Princeton University Press.

Balafoutas, Loukas, and Matthias Sutter. 2012. "Affirmative action policies promote women and do not harm efficiency in the laboratory". *Science,* 335 (6068) (Feb 3): 579-82.

Balafoutas, Loukas, Rudolf Kerschbamer, and Matthias Sutter. 2012. "Distributional preferences and competitive behavior". *Journal of Economic Behavior & Organization* 83 (1): 125-35.

Bertrand, Marianne, and Kevin F. Hallock. 2001. "The gender gap in top corporate jobs". *Industrial and Labor Relations Review* 55 (1).

Bertrand, Marianne. 2011. "New Perspectives on Gender." In Handbook of Labor Economics, vol. 4B, edited by Orley Ashenfelter and David Card. Elsevier, 1543 – 1590.

Booth, Alison, and Patrick Nolen. 2012. "Choosing to compete: How different are girls and boys?" *Journal of Economic Behavior & Organization* 81 (2): 542-55.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2013. "*Gender, competitiveness and career choices*." Forthcoming in The Quarterly Journal of Economics.

Cason, Timothy N., William A. Masters, and Roman M. Sheremeta. 2010. "Entry into winner take-all and proportional-prize contests: An experimental study." *Journal of Public Economics* 94 (9–10) (10): 604-11.

Charness, Gary, Uri Gneezy, and Michael Kuhn. 2012. "*Experimental Method: Between-Subject and Within-Subject Design*." Journal of Economic Behavior & Organization, 81, 1-8.

Croson, Rachel, and Uri Gneezy. 2009. "Gender differences in preferences." *Journal of Economic Literature*: 448-74.

Datta Gupta, Nabanita, Anders Poulsen, and Marie Claire Villeval. 2013. "Gender matching and competitiveness: Experimental evidence". *Economic Inquiry* 51 (1): 816-35.

Dohmen, Thomas, and Armin Falk. 2011. "Performance pay and multidimensional sorting: Productivity, preferences, and gender." *The American Economic Review* 101 (2): 556-90.

Dohmen, Thomas, David Huffman, Uwe Sunde, Jurgen Schupp, and Gert G. Wagner. 2011. "IndividualRisk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association* 9 (3): 522–550.

Dargnies, Marie-Pierre. 2009. "Team competition: Eliminating the gender gap in competitiveness." *Paris School of Economics, Mimeo*.

Ellison, Glenn, and Ashley Swanson. 2010. "The gender gap in secondary school mathematics at high achievement levels: Evidence from the American mathematics competitions." *Journal of Economic Perspectives* 24 (2): 109-28.

Ertac, Seda, and Balazs Szentes. 2010. The effect of performance feedback on gender differences in competitiveness: Experimental evidence. *Working Paper, Koc Univ., Turkey*.

Fryer, Roland G., and Steven Levitt. 2010. "An empirical analysis of the gender gap in mathematics." *American Economic Journal: Applied Economics* 2 (2): 210-40.

Grosse, Niels D., and Gerhard Riener. 2010. "Explaining Gender Differences in Competitiveness: Gender-Task Stereotypes." *Jena Economic Research Papers*.

Gneezy, Uri, Kenneth L. Leonard, and John A. List. 2009. "Gender differences in competition: Evidence from a matrilineal and a patriarchal society." *Econometrica* 77 (5): 1637-64.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in competitive environments: Gender differences." *The Quarterly Journal of Economics* 118 (3): 1049-74.

Gneezy, Uri, and Aldo Rustichini. 2004. “Gender and competition at a young age.” *American Economic Review*: 377-81.

Goldin, Claudia, Lawrence F. Katz, and Ilyana Kuziemko. 2006. “The homecoming of american college women: The reversal of the college gender gap.” *The Journal of Economic Perspectives* 20 (4): 133.

Günther, Christina, Neslihan Arslan Ekinci, Christiane Schwieren, and Martin Strobel. 2010. "Women can’t jump? An experiment on competitive attitudes and stereotype threat." *Journal of Economic Behavior & Organization* 75 (3): 395-401.

Healy, Andrew, and Jennifer Pate. 2011. “Can teams help to close the gender competition gap?.” *The Economic Journal* 121 (555): 1192-204.

Holt, Charles A., and Susan K. Laury. 2002. “Risk aversion and incentive effects.” *American Economic Review* 92 (5): 1644-55.

Hausmann, Ricardo, Laura D Tyson, and Saadia Zahidi, 2010. “The Global Gender Gap Report”, World Economic Forum.

Niederle, Muriel, Carmit Segal, and Lise Vesterlund. 2013. “How costly is diversity? Affirmative action in light of gender differences in competitiveness.” *Management Science* 59 (1):1–16

Niederle, Muriel, and Lise Vesterlund. 2010. "Explaining the gender gap in math test scores: The role of competition." *The Journal of Economic Perspectives* 24 (2): 129-144.

Niederle, Muriel and Lise Vesterlund. 2007. “Do Women Shy Away from Competition? Do Men Compete Too Much?” *The Quarterly Journal of Economics* 122 (3):1067–1101.

Pope, Devin G., and Justin R. Sydnor. 2010. “Geographic variation in the gender differences in test scores.” *The Journal of Economic Perspectives* 24 (2): 95-108.

Rigdon, Mary L. 2012. An experimental investigation of gender differences in wage negotiations. Working paper *Available at SSRN 2165253*.

Shurchkov, Olga. 2012. “Under pressure: Gender differences in output quality and quantity under competition and time constraints.” *Journal of the European Economic Association* 10 (5): 1189-213.

Sutter, Matthias and Daniela Rützler. 2010. “Gender Differences in Competition

Emerge Early in Life". IZA DP No. 5015.

Vandegrift, Donald, and Abdullah Yavas. 2009. "Men, women, and competition:An experimental test of behavior." *Journal of Economic Behavior & Organization* 72 (1):554-70.

Wozniak, David, William T. Harbaugh, and Ulrich Mayr. "The menstrual cycle and performance feedback alter gender differences in competitive choices." *Journal of Labor Economics* 32, no. 1 (2014): 161-198.