Why female decision-makers shy away from promoting competition*

Olga Shurchkov† Alexandra V. M. van Geen‡

First Draft: July 2016
Current Draft: October 2018

Abstract:

Incentivizing subordinates is a crucial task of anyone in a decision-making role. However, little is known about the mechanisms behind selection of different types of incentives. Our laboratory experiment characterizes the ways in which male and female decision-makers assign incentives, and how these choices are perceived by those affected by them. We find that women are significantly less likely to select "competitive" incentives based on comparative performance of workers, particularly in the treatment where their workers can observe their gender. The results are not due to priming, but are rather consistent with the explanation that women conform to gender stereotypes in anticipation of subsequent evaluation by workers. Indeed, female decision-makers are significantly underrated relative to comparable males, even after controlling for incentive choice and an extensive set of individual characteristics. The gender difference in competency ratings can be attributed to male workers rating female decision-makers disproportionately lower relative to their male counterparts. The gender gap in ratings appears to arise because of gender *per se*, and not due to a differential impact of incentives by decision-maker's gender.

Key Words: personnel economics, incentives, competition, gender differences, discrimination, economic experiments, labor markets

JEL Classifications: C91, J16, J71, M50

---

* **Acknowledgements**: The authors are grateful to Katherine Baldiga Coffman, Kristin Butcher, Phillip Levine, Andrea Robbett, Gauri Kartini Shastri, and the participants of the Wellesley Economics Department work-in-progress seminar, the Erasmus School of Economics seminar, and the 2016 ASSA session "Experimental Gender Economics" for valuable comments and discussion. In addition, we would like to acknowledge the generous research support we received from ERIM and the Erasmus School of Economics to organize and run our experiment. Shurchkov would like to acknowledge the Wellesley College faculty award grant for further financial support that made this research possible. Gillian Courtney, Maximilian Kerk, and Li Song provided excellent research assistance. All remaining errors are our own.
† Corresponding author. Department of Economics, Wellesley College, 106 Central St., Wellesley, MA, USA.
e-mail: olga.shurchkov@wellesley.edu.
‡ Erasmus School of Economics, Erasmus University, Rotterdam, the Netherlands.
e-mail: avangeen@gmail.com.

1. Introduction

Provision of incentives is one of the most crucial aspects of decision-making. In the labor market, positive incentives (or "carrots") such as bonuses for high performance have been used to boost worker effort and productivity. Managers also use punishments (or "sticks"), which can range from losing an anticipated bonus to verbal scolding, when a worker underperforms. Incentives may also vary across the applied performance benchmarks. In particular, one's personal improvement relative to own past performance (or lack thereof) may serve as a basis for reward or punishment ("self-reference incentives"). Alternatively, performance may be evaluated and incentivized relative to the performance of other workers ("competitive incentives").

Despite their widespread use, little is known about how and why different incentive regimes are adopted, and how the decision-maker characteristics, such as gender, can affect the way in which chosen incentives are perceived. In his seminal work, Becker (1968) points out the inherent equivalency between carrots and sticks. However, the psychology literature is split: on the one hand, prospect theory suggests that sticks should be preferable because they are more effective due to loss aversion (Kahneman and Tversky 1979); on the other hand, "positive reinforcement" theory suggests the use of carrots in repeated interactions (Wiegand and Geller 2008; Denhardt, Denhardt, and Aristigueta 2013).

The experimental literature has thus far focused on the effects of exogenously imposed incentives in non-labor market settings (e.g. Andreoni, Harbaugh, & Vesterlund 2003) with tasks that do not involve real effort (e.g. Dickinson 2001; Frederickson & Waller 2005). Furthermore, the literature has ignored the choice of performance benchmarks in the assignment of incentives. Thus, we know relatively little about behavior in more realistic settings, where one might observe potential gender effects. Related research shows that females dislike competition against others and perform less well in competitive environments, especially in stereotypically male-oriented tasks (Niederle and Vesterlund, 2007; Shurchkov 2012). On the other hand, recent studies (for example, Apicella, Demiral, and Mollerstrom (2017) and Carpenter, Frank, and Huet-Vaughn (2018) show that women are as likely as men to "compete against oneself" (i.e., under self-reference type of incentive scheme). Whether the avoidance of competitive settings for themselves translates into female managers shying away from promoting competition amongst others remains an open question.

This paper in the first to design a labor-market laboratory experiment which fully considers both sides of the incentive contract: incentive selection by the decision-maker, as well as the impact on worker effort and productivity in a real-effort task (an addition task) and the subsequent evaluation of the decision-maker by the worker. Furthermore, we enrich the incentive selection set and move beyond the standard carrots and sticks. We consider incentives that spur competition among the workers ("competitive incentives") in

addition to the incentives that reward or punish performance in reference to the worker's own past performance ("self-reference incentives").

Studying how incentive selection and its evaluation varies by gender of the decision-maker in particular can shed light on the sources of gender gaps in representation among senior management of firms across a variety of industries (Catalyst 2016). One potential explanation for these gaps in naturally occurring data is that gender may be correlated with particular preferences or unobservable skills, such as the manager's willingness to punish or the manager's ability to process employee information, that are important for the selection of appropriate incentives and could result in lower performance and lower desirability of female managers ("statistical discrimination," Arrow 1973). Another is that, conditional on incentive choices, male managers are simply preferred over females ("taste-based discrimination," Becker 1957 or "implicit discrimination," Bertrand, Chugh, and Mullainathan 2005).

Our novel design allows us to separate taste-based/implicit discrimination from statistical discrimination in the workers' evaluation of the decision-maker, measured by a competency rating. First, we observe the decision-maker's choice of incentive regime, which she can use to boost worker effort and performance in a subsequent addition task. This choice is based on basic worker characteristics such as their baseline performance. This enables us to test for potential gender differences in incentive selection. Unlike the real-world, where manager quality is unobservable to the researcher as well as to the workers, in our experiment the choice of incentive is perfectly known by workers and can be therefore eliminated as a reason for differences in competency rating. Thus, by controlling for the choice of incentive, we are then able to test directly whether the decision-maker's gender impacts the perceived quality of the decision through the reported competency rating. Although our design does not allow us to distinguish between tastes and implicit bias, we can test whether any gender gaps in ratings can be due to differential impact of incentives assigned by women on performance, or if women are underrated purely because of their gender.

Our analysis yields 3 main findings. First, we find that women are, on average, significantly less likely than otherwise comparable men to select competitive incentives for their workers. This result is consistent with the previous literature on gender differences in attitudes toward competition against others, especially in mathematical environments (Niederle and Vesterlund 2007; Shurchkov 2012; Apicella, Demiral, and Mollerstrom (2017); see Niederle 2016 for review of the literature). Our study contributes to this literature by uncovering a novel result that women prefer non-competitive incentive schemes not only when applied to their own performance in a similar task, but also to the performance of others. If the labor market disproportionately rewards managers who promote competition rather than self-improvement among the

workers, our result can help explain the observed underrepresentation of women at the top of the corporate hierarchy.[1]

Second, conditional on individual characteristics, women decision-makers are disproportionately less likely to choose competitive incentives and more likely to choose self-reference carrots in the treatment where their gender is made salient to them and workers can observe the gender of the decision-maker. This result is consistent with previous findings in psychology that the anticipation of stereotype threat and the desire to conform to gender identity can depress ambition and assertiveness in women (Spencer, Steele, and Quinn 1999; Gupta and Bhawe 2007). Indeed, it is possible that the anticipation of future evaluation may trigger the stereotype threat in women to further explain our results. Our finding is also consistent with the literature on priming that finds that a gender prime may adversely impact outcomes for women (Shih, Pittinsky, and Ambady 1999).

Lastly, we show that, when workers know the gender of their decision-maker, they rate females as significantly less competent than males, on average. Importantly, women continue to be underrated even when we condition on incentive choice and a rich set of individual characteristics, which reveals that the ratings are not driven by gender differences in incentive selection. A further decomposition by the gender of the worker reveals an interaction effect. In particular, male workers rate female decision-makers disproportionately lower relative to their male counterparts, even when we condition on the choice of incentive – a finding that is robust to a variety of alternative specifications. We are the first to show that no such gender differences exist in the treatment where the gender of the decision-maker is unknown, which suggests that women are rated as less competent due to their gender *per se*.

Further analysis suggests that the gender gap in ratings does not arise due to the perceived or actual gender differences in the impact of the incentive on worker performance. Because we condition our estimates of the gender gap in ratings on the choice of incentive, and because we find no significant differential gender effect of incentives on worker productivity, we conclude that the observed gender difference in competency rating is suggestive of gender bias on the part of the male workers against females in a decision-making role. Further research is required in order to fully understand the mechanisms behind this result. For example, it would be worthwhile to investigate whether female decision-makers would still be underrated if the task assigned to workers was mores stereotypically female-oriented.

Our laboratory study complements existing literature that uses observational data to study the role of incentives (eg., Lazear 2000); on gender differences in leadership outcomes (eg., Atkinson, Baird and Frye (2003) for mutual fund managers; Adams and Ferreira (2009) for board members; Cardoso and Winter-

[1] There is field evidence to suggest that competitive incentives, such as sales competitions among subsets of stores, tend to increase sales growth (Delfgaauw et al. 2013). Bloom et al. (2015) use field data on public hospitals to show that productivity increases when management creates a competitive environment.

Ebmer (2010) for employers; Ferreira and Gyourko (2014) for U.S. mayors; and Dezsö and Ross (2012), Flabbi et al. (2016), and Gagliarducci and Paserman (2015) for firm executives); and on different attitudes of men and women toward leadership styles (Eagly, van Engen, and Johannesen-Schmidt 2003) and economic policy (May, McGarvey, and Kucera 2018). The advantage of an experimental setting is that we can avoid the confounding factors that inevitably result from endogenous selection into managerial roles. We can also exogenously vary the environment, revealing the gender of the decision-maker to only a random subset of the workers. Finally, we are able to control for a set of individual characteristics that are not easily observable in field data, such as risk and social preferences. It is particularly important to condition on these characteristics in our analysis because a large body of previous experiments has found that, in general, women are more risk averse than men and are more sensitive to some social cues (see Croson and Gneezy (2009), Eckel and Grossman (2008), and more recently Shurchkov and Eckel (2018) for detailed reviews of the literature).

Our experiment is conceptually related to Fehr and Schmidt (2007) and Fehr, Klein, and Schmidt (2007) insofar as we also explore the mechanisms behind incentive choice. Dohmen and Falk (2011) also study how behavioral traits such as risk aversion and social preferences interact with gender to explain individual sorting into fixed or variable incentive schemes. The most closely related paper is Price (2012) who explores gender differences in managerial choice between tournaments and piece-rates. The focus of that paper is on the effect of worker gender and information about worker's ability in the task on incentive choice – channels that we choose to shut down in our study. Instead, we focus on the gender of the decision-maker. Furthermore, by introducing punishments and incentives based on self-improvement, we present decision-makers with a less constrained incentive choice set which introduce a new source of gender differences – a result not present in Price (2012). Finally, we break ground on a relatively unexplored topic of gender bias in the evaluation of the chosen incentive regime.

Our work suggests that women primed with gender and expecting performance evaluation are more likely to manage in a way that more closely conforms to the gender stereotype. However, they are not rewarded for this choice. On the contrary, we find that workers reward men who go against their gender stereotype and choose non-competitive incentives. Our results are also consistent with the Gallup (2013) data that indicate that more Americans prefer a male boss to a female boss. In our experiment, female decision-makers are perceived as less competent than males even when they make identical incentive choices and are otherwise similar. Thus, there is a bias that may lead to explicit or implicit discrimination that is not based on any rational expectation of gender differentially affecting subsequent performance. Organizations seeking to overcome this bias and achieve equality in promotion decisions may pursue blind evaluation procedures (Goldin and Rouse 2000), increase the proportion of female evaluators (Zinovyeva and Bagues 2011), or introduce joint rather than separate evaluation procedures (Bohnet, van Geen, and Bazerman 2016).

The remainder of the paper is organized as follows. In Section 2, we present an overview of the experiment and describe the data patterns and summary statistics. Section 3 reports and discusses the results. Section 4 summarizes the implications of our study and concludes.

2. The Experiment

The experiment was conducted at the Erasmus School of Economics (ESE) in September-October of 2015 and in April 2016 with a total of 297 subjects. Subjects were undergraduate and graduate students from Erasmus University in the Netherlands.[2] We ran a total of 16 experimental sessions. Each session included multiple groups of three subjects: one "decision-maker" (DM hereafter) and two "workers." Approximately half the sessions were designed to reveal the gender of the DM to the workers, while the other half did not reveal this information. In total, 153 subjects participated in the gender treatment (GT) and 144 subjects participated in the no gender treatment (NGT). The October 2015 sessions consisted of three rounds of interactions, while the April 2016 sessions consisted of a single round. Table 1 provides a summary of sessions and treatments.

[TABLE 1 ABOUT HERE]

Upon arrival, subjects read and signed a consent form and were assigned to a cubicle. Subjects remained anonymous throughout the experiment both to the experimenters and to each other. No communication between participants was allowed.

2.1. Resume Questionnaire and Baseline Task Performance

After being seated, subjects received both computerized and oral instructions for the first part of the experiment (see Appendix A for sample instructions). Subjects received computerized instructions on each screen, programmed using the standard zTree software package (Fischbacher 2007). First, all subjects answered questions about several characteristics relevant to one's labor market "resume": student status, major, GPA range, city of residence, university, age range. Importantly, in the GT, subjects reported their gender as part of this battery of simple questions, while in the NGT, gender was not elicited. The advantage of asking for the other "filler" personal characteristics is that gender appears less salient to the respondent. We selected variables in which we expected little variation so that we could disentangle the effect of gender once analyzing respondents' decisions in the second part of the experiment. Note that by simply answering

[2] Laboratory experiments primarily use students as subjects for reasons of consistency across subject pools that allows replication, cost-effectiveness, and ease of access. Results from any laboratory experiment must be interpreted with caution due to the high degree of abstraction from the real world, but student subjects have been shown to not act in significantly different ways as compared to non-students (see for example, Frechette (2015) for a meta-analysis of papers that use student and professional subject pools). Replication of our results with professional subjects could present a useful direction for future research.

the question about gender, the subjects in the GT may have been primed to think of themselves as male and female. Psychology literature has found that gender priming may activate stereotype threat against women, lowering their performance in mathematical tasks (Shih, Pittinsky, and Ambady 1999; Gibson, Losee, and Vitiello 2014). We will investigate whether competitive choices may be similarly susceptible to stereotype activation.

The subjects then received information about the task they would have to perform. In particular, every subject had 5 minutes to complete an addition task of summing up five randomly chosen two-digit numbers – a task that is used conventionally in the literature, including Niederle and Vesterlund (2007), Price (2012), Apicella, Demiral, and Mollerstrom (2017) and numerous others. The subjects were informed that they would not be paid for performance in this part of the experiment, but that the total score in the addition task would be recorded as part of their "resume" to be used later in the experiment.

Summary statistics of resume characteristics are reported in Panel 1 of Table 2. On average, women in our experiment are more likely to live in Rotterdam and to study social sciences (rather than humanities or natural sciences).

[TABLE 2 ABOUT HERE]

2.2. Decision-Makers and Workers

In part 2 of the experiment, subjects were randomly assigned to groups of three individuals with one DM and two workers. The design of this part of the experiment differed across the sessions conducted in 2015 and 2016. In 2015, part 2 consisted of 3 rounds. In each round, the role of DM rotated randomly among the three subjects (while the other two were workers). This design was meant to model a situation where teams engage in continuing work relationships, rather than the more stylized one-time interactions. In order to avoid income effects and to minimize unwanted learning effects, we did not reveal performance, payoff, or any other information to the subjects between the rounds. In 2016, part 2 consisted of only 1 round where the role of the DM was randomly assigned to one of the three subjects. By keeping a single round, we eliminate any gender-specific reciprocity effects or confounds that may stem from DMs expecting gendered backlash in future rounds (for example, Paryevi, Bohnet and van Geen (2016) find male backlash against females in leadership positions). For further discussion of round effects, see the online appendix.

DM Decisions and Payoffs

The randomly selected DM of the group makes his or her decisions based on resume information of both workers, which included baseline addition task performance, student status, GPA range, field of study, city of residence, university, and age range. Note that the DM never saw the gender of the workers. We

made it clear that the DM would receive half of the points earned by each worker in the subsequent 5-minute addition task, and that therefore it was in the DM's best interest to maximize worker performance. The DM knew that she had the opportunity to incentivize the workers by picking one of the following 5 options:

1. Worker does not get or lose any points, beyond his/her own points in the subsequent addition task (No added incentive beyond piece-rate)
2. The worker gets 50 additional points if his or her total score in the subsequent addition task is greater than that of the other worker (Competitive carrot)
3. The worker gets 50 additional points if his or her total score in the subsequent addition task is greater than his/her own previous addition score (from Part 1) (Self-reference carrot)
4. The worker loses 50 points if his or her total score in the subsequent addition task is lower than that of the other worker (Competitive stick)
5. The worker loses 50 points if his or her total score in the subsequent addition task is lower than his/her own previous addition score (from Part 1) (Self-reference stick)

Choosing option 1 did not cost anything to the DM, but it did not provide any added incentive above and beyond the piece-rate in the subsequent task (no added incentive). Choosing any one of the options 2 through 5 cost the DM 10 points[3]. Option 2 can be interpreted as a positive incentive based on the comparison between workers in the subsequent task (competitive carrot), whereas option 4 is a negative incentive also based on the comparison between workers (competitive stick). Option 3 can be interpreted as a positive incentive based on the comparison of worker's subsequent performance to her own past performance (self-reference carrot), and option 5 as a negative incentive based on her own past performance (self-reference stick).

In our experiment, if the DM expects that carrots and sticks are on average similarly effective in boosting expected worker effort in the subsequent task, Option 3 (self-reference carrot) would be the optimal choice from the point of view of maximizing group welfare. This is because both workers could hypothetically improve their own performance and therefore receive the 50 point bonus. This is not the case with Option 2 (competitive carrot), where - by definition - only one worker can earn the extra 50 points. In their field experiment featuring a comparison between carrots and sticks, Fryer et al (2012) opt for a symmetric payoff design where, in the negative incentive condition, teachers are paid in advance and asked to give back the money if their students do not improve sufficiently. The important difference in our experiment is that all subjects view the same list of incentive options which includes both sticks and carrots. The Fryer et al study, on the other hand, randomizes which pairwise comparison each subject makes: sticks

[3] This means that for the DM the profit maximizing strategy is to provide incentives whenever the expected additional solved addition problems across both workers exceed 2.

vs. no incentive or carrots vs. no incentive. In our experiment, if subjects were to realize that carrots and sticks are essentially identical, DMs would be more likely to make choices at random and workers would have less impetus to give low ratings. Thus, we intentionally allow for incentives to vary in terms of "quality," with self-reference sticks as the potentially "worst" incentive type from the welfare point of view. Note however that, welfare considerations aside, individual attitudes toward the effectiveness of competition or the effectiveness of fear of punishment could theoretically result in the selection of any of the four incentive choices.

The DM was aware that workers would subsequently see her choice and that the choice would be applied to the payoffs of both workers. Following the incentive choice, the DM predicted subsequent performance of both workers.

DMs were informed of their payoffs in a series of on-screen slides that detailed the above information (slides provided in the online appendix). Here, we summarize DM i's payoff in points as:

$$\pi_i = 100 + \frac{1}{2}\sum_{j=1}^{2} 10Y_j - 10 \times Incent_i - \frac{1}{4}\sum_{j=1}^{2} |Y_j - E_{i,j}|$$

where i denotes the DM, $j \in \{1,2\}$ is the set of workers; Y_j is the part 2 addition score of worker j; $Incent_i$ is an indicator function that takes on the value of 1 if DM i chooses any of the incentive options (2-5), and 0 otherwise; and $E_{i,j}$ denotes the score prediction for worker j by DM i.

The last term in the DM payoff represents a "misprediction penalty" that we include in order to incentivize truth-telling and reduce the "noise" in the beliefs data (Gächter and Renner 2010). Because the DMs do not actually perform the task, and the workers do not predict performance, there is no reason to expect any hedging bias, most importantly in terms of the prediction influencing the choice of incentive by the DM.

Worker Decisions

Next, workers received information about their DM. In particular, they viewed the DM's resume, which included the DM's baseline addition task score, student status, field of study, city of residence, university, and age range. In the GT, the resume also included the DM's gender, while in the NGT, it did not. Following the resume information, workers saw the incentive option actually chosen by their DM from all the other possible options.[4]

Workers then were asked to rate the competency of the DM based on all the information they received. The five-point rating scale ranged from "not at all competent" to "very competent." The workers also provided hypothetical ratings for each of the five possible incentive choices their DM could have chosen.

[4] Workers did not see the other worker's resume when making rating decisions, because we wanted them to focus on the DM's resume which presented an already complicated set of information to process.

We chose to not incentivize competency ratings, so that the DM could choose any incentive without the fear of retaliation by the worker.[5] Because this paper focuses primarily on the mechanisms behind incentive choice, we prioritize minimizing any distortions to the DM's incentive decision. Still, workers had no reason to withhold their evaluation or rate DMs untruthfully. Furthermore, there is no reason to expect truth-telling in ratings to vary across treatments.

After the ratings were completed, workers once again had 5 minutes to perform the addition task of summing up five randomly chosen two-digit numbers.

Workers were informed of their payoffs in a series of on-screen slides that detailed the above information (slides provided in the online appendix). Here, we summarize Worker j's payoff in points as:

$$\pi_j = 25 + 10Y_j + 50 \times \mathbf{1_2} max\left(Y_j - Y_{-j}, 0\right) + 50 \times \mathbf{1_3} max\left(Y_j - Y_{j,base}, 0\right)$$
$$-50 \times \mathbf{1_4} max\left(Y_{-j} - Y_j\ , 0\right) - 50 \times \mathbf{1_5} max\left(Y_{j,base} - Y_j, 0\right)$$

where $j \in \{1,2\}$ is the set of workers; Y_j is the part 2 addition score of worker j; Y_{-j} is the part 2 addition score of the other worker; $Y_{j,base}$ is the baseline addition score in part 1 of worker j; $1_X(\cdot)$ are indicator functions that take on the value of 1 if the DM had chosen a particular incentive option X, and 0 otherwise.

2.3. Preference Elicitation and Post-Experiment Questionnaire

In parts 3 and 4 of the experiment, subjects once again made individual choices. Part 3 elicited risk preferences, loss aversion, and ambiguity aversion via an elicitation procedure due to Carpenter, Matthews, and Robbett (2017). For each of the three types of preferences, the subjects had to pick one out of six bags, each representing a lottery. Each bag had ten balls labeled with the numbers of points the subject could win, if that ball was drawn. Thus, subjects had to make three decisions of which bag (or lottery) to draw from. At the end of the experiment, the computer randomly chose one of the three decisions and played out that lottery. The earnings were then added to the part 2 payoffs

Part 4 elicited social preferences with subjects participating in three standard games: Trust Game (TG), Ultimatum Game (UG), and Dictator Game (DG). In the TG and the UG, we employed the strategy method in order to gather data on the way each subject would respond from the perspective of both: proposer and recipient. In particular, in the TG, we asked each subject to report her initial hypothetical transfer, which measures the willingness to trust others. Next, the subjects took on the role of the other player and had to report their hypothetical return transfers for each possible initial transfer. The mean return transfer for each subject measures the willingness to reciprocate. In the UG, we asked each subject to report her initial offer,

[5] Note that if DMs in the first round expected future DMs to remember their choice and base their future incentive decisions on what happened in the previous round, incentive decisions in the first round could be biased. We address this concern in the online appendix and find no evidence to suggest any round effects. We also eliminate rounds 2 and 3 from the 2016 sessions and find no difference between round-one behaviors across the two samples, which shows such strategic anticipation effects are unlikely to exist in our experiment.

which measures strategic altruism. We also use a similar strategy method to elicit each subject's minimum acceptance cut-off. In the DG, every subject acted as the proposer, which measures pure altruism.

Summary statistics of individual risk and social preferences are reported in Panel 2 of Table 2.[6] Women in our sample are significantly more risk and loss averse than men, on average. This is consistent with a large body of literature on gender differences in preferences (Gneezy and Croson 2009; Eckel and Grossman 2008; Shurchkov and Eckel 2018). Women are also significantly less likely to trust others in our experiment which is also in line with the majority of previous studies (Gneezy and Croson 2009). We do not observe significant gender differences in ambiguity aversion, reciprocity, strategic altruism, or pure altruism.

At the end of the session, subjects were informed of their performance and earnings in the experiment. They then filled out a post-experiment questionnaire that asked for detailed demographic information. Experimental instructions and questionnaire contents are available in the online appendix. Mean earnings in the experiment (including the show-up fee) equaled 17.13Eu with a standard deviation of 3.06Eu. Sessions lasted approximately one hour. Summary statistics of demographic characteristics are reported in Panel 3 of Table 2. Males in our sample are significantly more likely to be white and to have higher household income than women, on average. Questions about leadership style reveal that "transformational leadership style" is significantly more likely to be preferred by women, while "commanding leadership style" is significantly more likely to be preferred by men. This is consistent with the psychology literature on gender differences in attitudes toward leadership (Rosener 1990; Eagly, van Engen, and Johannesen-Schmidt 2003).

3. Gender differences in incentive choice: Women shy away from competitive incentives

This section reports the main results of the study concerning the DM's choice of a particular type of incentive scheme. Figure 1 demonstrates that DMs overwhelmingly choose carrots (positive incentives), which is consistent with previous literature (Luft 1994; Rigdon 2009). Figure A1 in the online appendix shows that this preference does not vary across rounds. Self-reference carrots are the most frequently chosen incentive option. The preference for self-reference carrots is not surprising from a utilitarian perspective, given that this incentive type allows for an increase in payoffs for both workers in the event that each one successfully improves on her own past performance, which makes this option the most desirable from the perspective of maximizing payoffs for the whole group.

[6] Experiences in part 2 of the experiment may have influenced decisions in parts 3 and 4. But a much graver concern would arise if we were to elicit risk and social preferences prior to the main decisions. For example, choosing among the lotteries could make risk aversion salient to our female subjects, leading them to make more risk-averse incentive choices, biasing our findings of the gender gap upwards. Still, to minimize spillovers from part 2 into parts 3 and 4 we only reveal outcome and payment information at the end of the experiment. We also never share the information on DM ratings with the subjects to prevent reciprocity effects in the social dilemma games.

[FIGURE 1 ABOUT HERE]

Figure 1 documents that competitive incentives, on average, represent about one third of all incentives, as well as of carrot incentives.

A variety of factors may explain individual differences in incentive choice. For example, men and women may differ in their preferred incentive types. On average, women select carrot incentives 86% of the time in the first round, while men select carrots 78% of the time. However, this first-round difference is not statistically significant (t-test p-value of 0.281) and is reversed when we consider all three rounds. Unconditional means comparisons are misleading in this case, because observable worker characteristics, such as the workers' scores in Part 1, as well as the DM's own individual characteristics can influence decisions. Furthermore, since the provision of incentives is associated with a certain cost and an uncertain outcome, risk aversion is likely to be correlated with the willingness to provide incentives. Finally, social and other preferences may affect the willingness to provide incentives. Thus, we turn to a model where we condition our estimates on these characteristics.

We use a linear probability model to estimate the effects of DM's gender on the likelihood of choosing a particular type of incentive, conditional on observable worker characteristics that are relevant to the decision, a set of DM's own characteristics, DM's self-reported risk and social preferences, and a set of session and round fixed effects.[7] Result 1 summarizes the main takeaway from the analysis.

Result 1: On average, men and women do not significantly differ in their willingness to incentivize workers in general or to choose carrot incentives. However, female DMs are, on average, significantly less likely than otherwise comparable men to select competitive incentives. This effect arises mostly prominently among women in the treatment group.

Main support for Result 1 comes from Table 3. We begin with a look at the entire sample in Columns 1, 4 and 7. Directionally, women are on more likely to choose no incentives and less likely to assign carrots. Column 7 shows that women are significantly less likely to choose competitive incentives.

[7] Because the outcome variable is binary, we acknowledge that a logistic model would in principle be better suited to this estimation. Furthermore, a multinomial logit model may be appropriate because the dependent variable is a choice between five different options. However, due to a limited sample size, we are unable to estimate the effects with a logit or probit model while also staying consistent in our approach across all incentive types (especially for the stick and no added incentive choices). Similarly, we are unable to estimate a multinomial logit model with the same rich set of controls and interactions. Note that the results are qualitatively similar when we use the logistic models to estimate the effects of DM gender on the choice of incentives using a subset of controls that allow the model to converge. Throughout the paper we also keep consistent in our approach of including all observable DM characteristics as controls, without making judgments on the relative importance of each characteristic.

The results are even stronger when we restrict the sample to the first round, which is our preferred specification that avoids a potential bias stemming from any possible reciprocity effects in rounds 2 and 3. Although the effects are present and significant when we pool the data, we choose to focus on the first round, because in the first round, past experiences as a worker do not confound the decision as a DM.[8]

Column 2 shows that, on average, women are marginally more likely to choose no added incentive. Decomposing the effect by treatment in Column 3, we find that women in the treatment group are marginally more likely than men in the treatment group to select no added incentives (f-test p-value of 0.057). Columns 4-6 show that the gender of the DM does not systematically impact the choice of carrots over sticks or no additional incentives.

The most robust finding concerns the effects of gender on the selection of competitive incentives (Columns 7-9). In all specifications, women are significantly less likely to choose competition for their workers. This result differs substantially from Price (2012) who finds no gender differences in incentive choices, on average. One possible explanation for the difference from the previous study is that we expand the list of incentives, introducing the self-reference option, which is preferred by female decision-makers.

Column 9 decomposes this effect by treatment. We find that when female DMs anticipate that their gender will be revealed to the workers, they are significantly less likely to select competitive incentives than their male counterparts who also anticipate that the workers will be aware of their gender (f-test p-value of 0.005). Importantly, conditional on worker and DM characteristics, treated female DMs are also less likely to select competition when compared to female DMs who are in the no gender treatment (f-test p-value of 0.073).[9] This finding suggests that the relative unwillingness of female DMs to offer competitive incentives might be tied to the reluctance of going against gender stereotypes, either because of pure priming effects or because the female DMs know that they will be subsequently evaluated by workers. In the control group, where the workers cannot observe the gender of the DM and the DM's were not asked for their gender in the resume building stage, female DMs are less unwilling to go against gender stereotypes and assign competitive incentives. On the other hand, when gender is revealed women conform to the gender stereotype and are less willing to assign competition to their workers.

[TABLE 3 ABOUT HERE]

[8] Appendix B provides additional analysis of the effects of multiple rounds on DM choice. First, we replicate the analysis in Table 3 for the full set of data, as well as in Round 3 only, to show that our results hold. We also show that pooling round 1 data from the 2015 multi-round sessions with round 1 data from the 2016 single-round sessions is justified because we do not find significant effects of the existence of multiple rounds on any of our outcome variables, including incentive choice.

[9] Even though the effect is marginal, it is robust to alternative specifications (such as ones that exclude DM controls, such as experience and preferences). The effect exists also in the pooled data across the three rounds, separating the analysis by treatment. See the online appendix for these robustness checks.

Table 3 also shows the effects of DM preferences on incentive choice. Individuals who exhibit greater preference for risk-taking are more likely to incentivize their workers (Columns 1-3). This is consistent with previous literature (Dohmen and Falk 2011) and with our expectations, as the provision of incentives entails paying a certain cost in exchange for an uncertain payoff. An increase in ambiguity aversion increases the likelihood of choosing no added incentive (Columns 1-3, significant in the first round) and decreases the likelihood of other incentive types (Columns 4-9). Social preferences do not systematically affect incentive selection. We also find no systematic interaction effects between the gender of the DM and preferences (see the online appendix for detailed discussion). Because prior decisions, such as the choice of incentive, may influence answers to the risk and social preferences surveys, we caution against causal interpretations of the relationships between preferences and Part 2 decisions. However, we note that our results are conceptually consistent with previous findings on risk-taking on behalf of others. Specifically, women are, on average, significantly less willing to make risky decisions on behalf of a group, but women who do sort into group decisions are not significantly more likely to take on risk than those who sort into deciding only for themselves (Ertac and Gurdal 2012).

Table 4 considers the effect of DM gender on the selection of each incentive type, conditional on the same set of controls as in Table 3.[10] On average, women are marginally less likely to choose competitive sticks (Column 3). Decomposing the effect of DM gender by treatment shows no significant effects for sticks (Columns 4 and 8), although the limited sample size of these incentive choices precludes us from making any definitive conclusions. Treated female DMs (GT) are less likely to select competitive carrots relative to treated male DMs (Column 2), and are more likely to select self-reference carrots relative to otherwise similar female DMs in the NGT (Column 6). These preference differences are consistent with the aggregate choices reported in Table 3.

[TABLE 4 ABOUT HERE]

The result that women are less likely to select competitive incentives with a mathematical task is consistent with the previous literature on gender differences in attitudes toward competition (Niederle and Vesterlund 2007). Shurchkov (2012) shows that task gender stereotypes represent an important mechanism for the unwillingness of women to enter competitions. Here, we contribute a novel finding to this literature: women prefer non-competitive incentive schemes not only when applied to their own performance, but also to the performance of others. Furthermore, the finding that women are less likely to promote competition

[10] The analysis in Table 4 is restricted to the first round. Table B6 in the Appendix reproduces the table with the entire sample, pooling data across the three rounds. Qualitatively the results are similar, but the effects are less pronounced. We believe that the cleanest incentive choice decision happens in the first round. First, DMs in the second and third round may recall the gender of their workers in the GT, because those workers were DMs in the preceding rounds. This could bias the choice of incentives. Second, subjects serving as DMs in subsequent rounds were workers in the preceding rounds. This experience with the environment could dilute the effects of gender.

in the treatment where their gender is revealed is consistent with the extensive past psychology literature. Specifically, the anticipation of stereotype threat activation and the desire to conform to gender identity have been found to depress ambition and assertiveness in women (Spencer, Steele, and Quinn 1999; Gupta and Bhawe 2007). Our finding may also be consistent with the interpretation that single women, representing 94 percent of our female subjects, may shy away from signaling competitiveness – a stereotypically male trait – due to a subconscious fear of being penalized in the marriage market (Bursztyn, Fujiwara, and Pallais 2017).

4. Gender differences in evaluations: Female DMs are rated as less competent

In this section, we investigate whether and why we may find gender differences in the way DMs are perceived by the workers.

The mean competency rating in our sample is 2.61 (ranging from 0 to 4, with 4 being the highest) with a standard deviation in ratings is 1.05. (See appendix A for detailed information on means and standard deviations for all decisions variables.)

Figure 2 shows that women are overall rated as significantly less competent than men in our experiment, across all sessions and treatments. The breakdown by rating in Figure 3 reveals that female DMs receive relatively more low ratings than men (1 and 2), while men receive more high ratings of 3 and especially of 4 – the highest competency rating.

[FIGURE 2 ABOUT HERE]
[FIGURE 3 ABOUT HERE]

In light of the significant gender differences in incentive selection, with women shying away from competitive incentives in favor of self-reference ones, we explore whether the gender gap in ratings may arise because workers especially prefer competition in our experiment. Figure 4 refutes this hypothesis. In fact, workers dislike competitive incentives in our experiment: DMs who choose competitive incentives receive a mean competency rating of 2.47, which is significantly lower than the rating for self-reference incentives of 2.83 (t-test p-value of 0.001). Workers also seem to prefer carrots over sticks or no added incentives. The mean rating for all carrot incentives is 2.80, while the mean rating for all other incentives is 1.70 (t-test p-value of <0.001).

[FIGURE 4 ABOUT HERE]

An important next question is whether the observed gender difference in ratings exists only when workers can observe the gender of the DM or if it exists regardless of treatment. Table 5 decomposes the mean competency rating by DM gender and treatment. The comparison of overall average ratings for male

and female DMs (top row) reveals that female DMs are indeed rated as significantly less competent than male DMs only when the workers know the gender of the DM. Furthermore, subsequent comparisons by incentive type show that gender gaps in the gender revealed treatment are significant for all incentives, carrot incentives, and self-reference incentives – precisely the incentive types that are favored by workers, in general.

[TABLE 5 ABOUT HERE]

Because the workers seem to rate DMs differently based on incentive choice, it is important to control for that choice in order to correctly estimate the effect of DM's gender on ratings. Thus, we can improve upon the simple means comparison by conditioning on the chosen incentive type, as well as controlling for the DM characteristics observable to the workers and for workers' own characteristics that may be relevant to the rating. In other words, we next ask whether female and male DMs rated differently even when the quality of the DM's decision and other observable characteristics are held constant. The results are summarized in Result 2.

Result 2: Conditional on the choice of incentive and an extensive set of individual characteristics, the gender difference in ratings can be attributed to male workers rating female DMs disproportionately lower relative to their male counterparts.

Support for Result 2 comes from Table 6 which reports estimates from fixed-effects OLS regressions of the effect of DM gender on the competency rating chosen by the worker. Column 1 reproduces the mean treatment effect of DM gender in the full sample: on average, female DMs receive significantly lower ratings when the gender of the DM is observed (GT). Column 6 shows no mean gender difference in ratings when the gender of the DM is unknown (NGT).

Specifications 2-5 and 7-10 condition on the choice of incentive by the DM, and introduce a rich set of controls that include DM resume characteristics observable by the worker, worker demographic characteristics, and worker social preference measures. We omit controls for worker risk preferences, because the rating is not associated with any risk, as it is not observed by the DM and does not affect payoffs, but the main result above is robust to their inclusion. Columns 4-5 and 9-10 further restrict the sample to the first round in order to check that the gender differences are not driven solely by reciprocity in the subsequent rounds.

We observe that the mean effect of gender becomes insignificant when we control for the choice of incentive in the full sample (Column 2). However, the magnitude of the coefficient decreases only slightly, which suggests that incentive choice is not the main driver of the gender gap in ratings. When the sample

is restricted to the first round – our preferred specification – the gender gap remains marginally significant even after we include controls (Column 4).

Columns 3 and 5 of Table 6 provide support for Result 2 by decomposing the effect of the DM's gender by the gender of the worker who performs the rating in the gender treatment. In particular, we show that male workers rate female DMs substantially lower than otherwise comparable male DMs, all else equal. In the full sample (Column 3), the reduction in ratings for women amounts to approximately one half of a standard deviation, or one half of 1.05. This result is robust to a variety of alternative specifications: dropping all controls and fixed effects; dropping all controls, but including fixed effects; dropping worker controls, but including the DM resume controls and fixed effects. When we restrict the sample to the first round, female DMs are penalized by male workers by about one standard deviation of ratings (Column 5). Once again, we observe a stronger gender difference in the first round, which is attributable to the fact that workers have not yet served as DMs in the first round. Column 8 and 10 confirm that these differences arise due to the knowledge of the DM's gender, because none of the interactions is significant and neither are the differences between the coefficients.

Table 6 offers several other noteworthy observations. Once again, we note that, in both treatments, DMs choosing carrots are rated as more competent than DMs choosing sticks (see the f-test p-values for all specifications in Table 6). In the full sample and in the NGT, DMs choosing carrots are also significantly favored relative to those choosing no added incentive (Columns 2, 3, 7-10).

Next, we ask whether female DMs who choose a certain type of incentive are more likely to get underrated relative to male DMs who also choose the same incentive. Table 7 reports the results from specifications that interact the gender of the DM with a particular incentive choice, conditional on the same set of controls as in Table 6. The omitted category in all regressions is the interaction of male DM with any other incentive (including no added incentive). Column 1 shows that men and women who choose competitive carrots are not rated differently by the workers. On the other hand, Column 3 reveals that female DMs who choose self-reference carrots are rated as less competent as compared to male DMs who also choose self-reference carrots. Note that differential gender effects are not present in the NGT (Columns 5-8).[11]

The relatively small sample of sticks in our experiment precludes us from drawing definitive conclusions about gender differences in ratings of DM choosing those incentives. Column 2 of Table 7 suggests that female DMs who assign competitive sticks are rated lower than male DMs making any other incentive choice. Furthermore, female DMs who assign competitive sticks seem to be rated marginally

[11] The results in Table 7 are robust and even strengthened when we consider the full sample, or when we exclude some or all controls. One additional finding with the full sample is that female DMs who choose competitive carrots are rated as less competent than female DMs choosing any other incentive. These results can be found in the online appendix and not in the main paper, because we prefer the more conservative specifications that restrict data to the first round, minimizing potential effects of learning and reciprocity.

lower than female DMs who choose any other incentive. However, these results should be interpreted with caution, given that there is not enough power to estimate the joint effect of male DM assigning competitive sticks in the GT (Column 2), female DM assigning self-reference sticks in the GT (Column 4), or male DM assigning self-reference sticks in the NGT (Column 8).

[TABLE 7 ABOUT HERE]

We conclude that, conditional on the DMs assigning the same incentive and having the same resume, we still find that women are under-rated relative to men – evidence suggestive of implicit bias or taste-based rather than statistical discrimination. In fact, female DMs are not awarded higher ratings for choosing the most optimal self-reference incentives and acting in a manner congruent with the societal gender norm (Eagly and Karau 2002). The results suggest an opposite effect: *men* who go against their gender stereotype of being competitive (see for example Cejka and Eagly 1999) by choosing self-reference carrots are rated as disproportionately more competent relative to women who also choose this incentive.

5. Potential explanations for the gender gap in ratings

It is possible for a particular incentive chosen by a woman to be perceived as differentially effective if it were chosen by an otherwise comparable man. While we are unable to gauge workers' beliefs about relative effectiveness of different incentive types, we can observe the impact of incentives on effort and performance, both overall and by gender of the DM.

Average numbers of entered solutions and average scores by incentive type shown in Figure 5 reveal that, in our experiment, some types of incentives perform better than others. For example, the total number of entered solutions is significantly greater under competitive carrots than under self-reference carrots or self-reference sticks (t-test p-values of 0.019 and 0.020, respectively). Score is, on average, significantly higher under competitive carrots than under self-reference sticks (t-test p-value of 0.031).

[FIGURE 5 ABOUT HERE]

However, added incentives do not substantially improve outcomes relative to the no added incentive baseline. The first reason could be that workers are already incentivized with a piece-rate incentive scheme, which reduces the importance of additional incentives. Second, the choice of incentive itself may reveal some expectation of future worker performance, because the DM observes relative baseline ability of workers in the task and can base her incentive decision on this information. Furthermore, workers' own characteristics may influence the ability of different incentive types to affect outcomes. In their review of the extensive literature on incentives, Gneezy, Meier, and Rey-Biel (2011) point out that the effectiveness

of extrinsic incentives in motivating desired behavior largely depends on the context and on the individual characteristics of the targeted parties. Finally, in treatments where the gender of the DM is observable, workers may respond differently to the same incentives if chosen by men than if chosen by women.

Table 8 documents the effect of different incentive types and gender on worker performance, conditional on the same rich set of controls as in Tables 6 and 7. All specifications also include controls for worker risk preferences, because Hannan, Hoffman, and Moser (2005) find that effort may interact with loss aversion and other preferences differentially under different incentive contracts (carrots vs. sticks, for example).[12] Columns 1 and 4 show the average effect of own gender and DM gender on score, for the gender treatment and the no gender treatment, respectively. When gender is known (Column 1), female workers score lower than males, while workers with female DMs score higher than workers with male DMs. This is not the case in the sessions where gender was not observable (Column 4). The explanation due to Spencer, Steele, and Quinn (1999) is that making gender salient in the gender treatments triggers stereotype against women, leading to lower performance. Gneezy, Niederle, and Rustichini (2003) point out similar effects when gender stereotypes are triggered by competition against men. It is also possible, however, that the gender difference found in columns 1 and 4 may arise due to priming effects. Because the gender difference in the Part 1 score does not seem to vary significantly by treatment (see the online appendix for estimates), we find this explanation less plausible.

Columns 2 and 3 show that gender differences in baseline task performance and individual-level differences may at least partially explain this effect. Indeed, the inclusion of the part 1 addition score and of individual characteristics renders the coefficient on worker gender insignificant and reduces it magnitude, implying that these controls serve as explanatory factors in driving worker performance. The results are qualitatively similar, and the effects are stronger, if we consider the total number of entered solutions (our proxy for effort) as the outcome variable (see the online appendix for estimates).

The gender of the DM significantly and positively affects worker performance, but only when we do not control for the kinds of incentives men and women tend to choose (Column 1). Column 2 reveals that the choice of incentive explains a large part of this effect. This makes sense as female DMs are more likely to assign the optimal self-referencing incentives.

Finally, while Column 2 of Table 8 shows that there are no significant average effects on worker performance of any of the incentive types relative to the no added incentive baseline if we pool the data across all three rounds, Column 3 shows that incentives, especially sticks, tend to reduce performance. Surprisingly, these negative incentive effects are only present in the GT, and in fact, directionally carrot incentives seem to have a positive effect on performance when DM gender is unobserved (Columns 5 and 6). However, the differences between treatments are not statistically significant.

[12] The impact of incentives varies across subjects with different preferences (loss aversion and reciprocity) in a manner consistent with Hannan, Hoffman, and Moser (2005). For detailed discussion, see the online appendix.

[TABLE 8 ABOUT HERE]

Table 9 decomposes the effects of incentives by the gender of the DM. The relevant comparison here is (A) and (B): how a given incentive affects score under female vs. male DM. In general, the f-tests find that the knowledge of the DM's gender does not significantly impact worker productivity under any incentive type. Unfortunately, power limitations prohibit us from making this comparison for competitive sticks. The results are not substantively different when we consider effort as our outcome, and are robust to a number of alternative specifications: when we consider the full sample, or when we exclude some or all controls. Specifications that include an interaction between the gender of the worker and the gender of the DM do not produce significant differences (female workers do not significantly benefit from having a female DM; results available upon request).

[TABLE 9 ABOUT HERE]

Although incentives in our experiment do not have differential effects on worker outcomes by DM gender, it is possible for the ratings to be influenced by the perceived "quality" of the incentive. We find that an imputed proxy for the value added by incentives cannot explain the gender differences in ratings (see the online appendix for discussion and estimates).

Our analysis has been able to show that gender differences in ratings must be due to the knowledge of the gender of the DM *per se*, rather than due to any gender differences in incentive choices or the impact of those choices. One remaining explanation is pure discrimination on behalf of male workers against female DMs. However, there are multiple other possible mechanisms for why the gender of the DM could matter. First, the evaluation of competency can be associated with the type of task workers expect to perform. Previous literature has found that women are more likely to compete and thrive under competition in tasks that are stereotypically perceived to favor women, such as verbal tasks (Shurchkov 2012). On the other hand, in tasks perceived to favor men, such as our task of addition, women shy away from competition and in some cases underperform men in tournaments (Gneezy, Niederle, and Rustichini 2003; Niederle and Vesterlund 2007). Similar effects may be at play in our experiment, where female managers might be stereotyped to be less competent in choosing incentives associated with male-oriented tasks. In a subsequent study, we plan to explore this dimension of the gender gap in ratings.

Another possible explanation for the gender gap in ratings can be that workers perceive men to be more "deserving" of the role of DM, which in our current experiment is assigned at random and is unrelated to past task performance. In order to explore this potential mechanism for the ratings gap, we plan to investigate whether workers continue to underrate female DMs even if they advanced to that position by

outperforming others in the Part 1 task. We speculate that women who become DMs due to high performance may in fact be punished even more relative to randomly chosen female DMs because male workers may dislike the feeling of having been outperformed by a woman.

Finally, worker ratings do not affect their own payoffs or the DM's payoffs in our experiment. This design feature allows for unbiased elicitation of preferences for incentives from the DMs, which is the main focus of this paper. If we were to incentivize competency ratings, the DMs could have responded by altering their preferred incentive choices in anticipation of subsequent retaliation. Thus, whether workers would pay to discriminate against women remains an open question – one that we leave for a future study where we simplify incentive choice but implement incentivized competency ratings by workers.

6. Conclusion

Our results indicate that men and women make different incentive choices when given a flexible list of possibilities. In particular, female decision-makers are significantly less likely to impose competition on their workers, on average. Furthermore, we find a treatment effect. In the control group, where the DMs know that workers will not view their gender, we find no gender difference in preferences for competitive incentives. On the other hand, when DMs know that gender will be revealed to the workers, women are significantly less willing to assign competition. The difference may be due to the reluctance of female DMs to go against gender stereotypes. In particular, gender serves as a reminder that women would be judged (rated) as "female managers" in the gender reveal treatment, as opposed to simply "managers" in the no gender reveal treatment. The difference may also arise purely due to gender priming.

Although women are less likely to choose competitive incentives for their workers, our analysis of the impact of incentives on effort and performance does not find any evidence to suggest that incentives chosen by female DMs result in lower worker productivity. Still, we find that workers rate female DMs as less competent as compared to male DMs making exactly the same decisions and who are otherwise comparable across a broad range of individual characteristics. Furthermore, the gender difference in ratings can be attributed to male workers rating female DMs disproportionately lower relative to their male counterparts. Importantly, these differences exist only in the treatment where the workers observe the gender of their decision-maker. Formal and informal evaluations by peers and subordinates constitute a significant part of the promotion process as part of the so-called 360-degree performance review, because the person up for promotion works most closely with these individuals (Gallagher 2008). Our results provide direct evidence that the gender composition of the evaluators matters for the unbiasedness of the evaluation. Although we show that the gender of the DM is a key determinant in this evaluation, we are cautious in interpreting the results as pure bias against female decision-makers on the part of the male workers. Other explanations can also be possible, such as the influence of our stereotypically male task, or the perception of merit (or lack thereof) in achieving DM status. We address these possibilities in future work.

The broad implication of our results is that the saliency of gender distorts both the manager's decision to provide incentives and the subsequent evaluation of managers. Thus, one direct policy implication may be to increase female share in managerial roles in order to avoid the "token" woman effect that would make gender extremely salient. Furthermore, female managers in particularly male-oriented domains would further benefit from a higher share of female workers on their teams, as we find that the most pronounced gender gap in competency ratings is amongst our male workers.

Regardless of the mechanisms behind the gender gap in competency ratings, any kind of gender discrimination in the workplace is illegal. In this paper, we have shown that it is also can be inefficient. If women with the same track record and whose impact on worker performance is equivalently strong as their male counterparts are consistently underrated and therefore overlooked for promotion, then there would be an efficiency loss in the labor market. In addition, the anticipation of the stigma associated with choosing particular (stereotypically masculine) methods to incentivize their workers may cause them to opt for suboptimal methods. Measuring the size of this efficiency loss is beyond the scope of this study, but may prove to be a useful avenue for future research.

Because of the experimental nature of our work, addressing its external validity is important. First, managers are aware of the gender of their workers, which may play a role in incentive selection. Note that the decision-maker in the third round of the gender treatment may recall the genders of the workers because those subjects were themselves decision-makers in the previous rounds. However, the sample is currently too small to produce any conclusive evidence, and an extension to explore these effects may be warranted. Second, our work abstracts away from an individual's own decision to take on a leadership role, because the decision-maker in our experiment is randomly assigned. Further exploration of gender differences in selection into the leadership role would enhance our understanding of the complex nature of the promotion process. Third, the evaluation in our experiment is performed by the workers rather than by third-party observers. However, promotion decisions are made by the members of top management who are not directly involved in the team production process. Experiments where evaluation of the decision-maker is conducted by such "third-party observers" would complement the evidence from our study and present a fruitful avenue for future research.

References

Adams, Renée B. and Daniel Ferreira. (2009). Women in the Boardroom and Their Impact on Governance and Performance. *Journal of Financial Economics* 94: 291-309.

Andreoni, James, Harbaugh, William, and Lise Vesterlund. (2003). The Carrot or the Stick: Rewards, Punishments, and Cooperation. *American Economic Review* 93 (3): 893-902.

Apicella, Coren L., Demiral, Elif E., and Johanna Mollerstrom. (2017). No Gender Difference in Willingness to Compete When Competing against Self. *American Economic Review Papers & Proceedings* 107 (5): 136-40.

Arbak, Emrah and Marie Claire Villeval. (2013). Voluntary Leadership: Selection and Influence. *Social Choice and Welfare* 40 (3): 635-62.

Arrow, Kenneth J. (1973). The Theory of Discrimination, in O. Ashenfelter and A. Rees (eds.), *Discrimination in Labor Markets*, Princeton, NJ: Princeton University Press.

Atkinson, Stanley M., Baird, Samantha B. and Melissa B. Frye. (2003). Do Female Mutual Fund Managers Manage Differently? *Journal of Financial Research* 26 (1): 1-18.

Becker, Gary. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.

Becker, Gary. (1968). Crime and Punishment: An Economic Analysis. *Journal of Political Economy* 78 (2): 169-217.

Bertrand, Marianne, Chugh, Dolly, and Sendhil Mullainathan. (2005). Implicit Discrimination. *American Economic Review* 95 (2): 94-98.

Bloom, Nicholas, Propper, Carol, Seiler, Stephan, and John Van Reenen. (2015). The Impact of Competition on Management Quality: Evidence from Public Hospitals. *Review of Economic Studies* 82 (2): 457-89.

Bohnet, Iris, Geen, Alexandra V.M. van, and Bazerman, Max H. (2016). When Performance Trumps Gender Bias: Joint Versus Separate Evaluation. *Management Science* 62 (5): 1225-1531.

Brandts, Jordi, Cooper, David J., and Roberto A. Weber. (2014). Legitimacy, Communication and Leadership in the Turnaround Game. *Management Science* 61 (11): 2627-45.

Bursztyn, Leonardo, Fujiwara, Thomas, and Amanda Pallais. (2017). 'Acting Wife': Marriage Market Incentives and Labor Market Investments. *American Economic Review* 107 (11): 3288-3319.

Carpenter, Jeffrey, Frank, Rachel, and Emiliano Huet-Vaughn. (2018). Gender Differences in Interpersonal and Intrapersonal Competitive Behavior. *Journal of Behavioral and Experimental Economics*, in press.

Carpenter, Jeffrey, Matthews, Peter Hans, and Andrea Robbett. (2017). Compensating Differentials in Experimental Labor Markets. *Journal of Behavioral and Experimental Economics*, 69: 50-60.

Cardoso, Ana R. and Rudolf Winter-Ebmer. (2010). Female-Led Firms and Gender Wage Policies. *Industrial and Labor Relations Review* 64 (1): 143-63.

Catalyst. (2016. *Pyramid: Women in S&P 500 Companies*. New York: Catalyst.

Cejka, Mary Ann and Alice H. Eagly, 1999. Gender-Stereotypic Images of Occupations Correspond to the Sex Segregation of Employment, *Personality and Social Psychology Bulletin*, 25(4): 413-423.

Croson, Rachel and Uri Gneezy. (2009). Gender Differences in Preferences. *Journal of Economic Literature* 47 (2): 1-27.

Denhardt, Robert B., Denhardt, Janet V, and Maria P. Aristigueta. (2013). *Managing Human Behavior in Public and Nonprofit Organizations*. SAGE Publications, London, United Kingdom.

Dickinson, David L. (2001). The Carrot vs. the Stick in Work Team Motivation. *Experimental Economics* 4 (1): 107-24.

Delfgaauw, Josse, Dur, Robert, Sol, Joeri, and WillemVerbeke, W. (2013). Tournament Incentives in the Field: Gender Differences in the Workplace. *Journal of Labor Economics* 31 (2): 305-26.

Dezsö, Cristian L. and David Gaddis Ross. (2012). Does Female Representation in Top Management Improve Firm Performance? A Panel Data Investigation. *Strategic Management Journal* 33 (9): 1072-89.

Dohmen, Thomas and Armin Falk. (2011). Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. *American Economic Review* 101 (2): 556-90.

Eagly, Alice H., Engen, Marloes L. van, and Mary C. Johannesen-Schmidt. (2003. "Transformational, Transactional, and Laissez-faire Leadership Styles: A Meta-Analysis Comparing Women and Men." *Psychological Bulletin* 129(4): 569-91.

Eagly, Alice H. and Steven J. Karau. (2002). Role Congruity Theory of Prejudice Toward Female Leaders. *Psychological Review*, 109 (3): 573–598.

Eckel, Catherine C. and Grossman, Philip J. (2008). Differences in the Economic Decisions of Men and Women: Experimental Evidence. In the *Handbook of Experimental Economics Results*, Vol. 1, edited by Charles Plott and Vernon Smith, 509-19. New York: Elsevier.

Ertac, Seda and Mehmet Y. Gurdal. (2012). Deciding to Decide: Gender, Leadership and Risk-Taking in Groups. *Journal of Economic Behavior and Organization* 83 (1): 24-30.

Fehr, Ernst and Klaus M. Schmidt. (2007). Adding a Stick to the Carrot? The Interaction of Bonuses and Fines. *American Economic Review* 97 (2): 177-81.

Fehr, Ernst, Klein, Alexander, and Klaus M. Schmidt. (2007). Fairness and Contract Design. *Econometrica* 75 (1): 121-54.

Flabbi, Luca, Macis, Mario, Moro, Andrea, and Fabiano Schivardi. (2016). Do Female Executives Make a Difference? The Impact of Female Leadership on Gender Gaps and Firm Performance. NBER Working Paper No. 22877.

Fischbacher, Urs. (2007). z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics* 10: 171-78.

Ferreira, Fernando and Joseph Gyourko. (2014). Does Gender Matter for Political Leadership? The Case of U.S. Mayors. *Journal of Public Economics* 112: 24-39.

Fréchette, Guillaume R. (2015). Laboratory Experiments: Professionals versus Students, in *Handbook of Experimental Economic Methodology*, Guillaume R. Fréchette and Andrew Schotter (eds), Oxford University Press, pp: 360-390.

Frederickson, James R. and William S. Waller. (2005). Carrot or Stick? Contract Frame and Use of Decision-Influencing Information in a Principal-Agent Setting. *Journal of Accounting Research* 43 (5): 709-33.

Fryer, Roland G., Jr., Levitt, Steven D., List, John, and Sally Sadoff. (2012). Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. NBER Working Paper 18237.

Gächter, Simon and Elke Renner. (2010). The Effects of (Incentivized) Belief Elicitation in Public Good Experiments. *Experimental Economics* 13 (3): 364–77.

Gagliarducci, Stefano and M. Daniele Paserman. (2015). The Effect of Female Leadership on Establishment and Employee Outcomes: Evidence from Linked Employer-Employee Data. *Research in Labor Economics* 41: 343-75.

Gallagher, Tracy. (2008). 360-Degree Performance Reviews Offer Valuable Perspectives. *Financial Executive* 24 (10): 61.

Gallup. (2013). Americans Still Prefer a Male Boss. Work & Education Poll.

Gibson, Carolyn E., Losee, Joy, and Christine Vitiello. (2014). A Replication Attempt of Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance. *Social Psychology* 45: 194-98.

Gneezy, Uri, Meier, Stephan, and Pedro Rey-Biel. (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives* 25 (4): 191-210.

Grosse, Niels, Riener, Gerhard, and Markus Dertwinkel-Kalt. (2014). Explaining Gender Differences in Competitiveness: Testing a Theory on Gender-Task Stereotypes. Working Paper.

Grossman, Philip J., Komai, Mana, and Jensen, James E. (2015). Leadership and Gender in Groups: An Experiment. *Canadian Journal of Economics* 48 (1): 368-88.

Gupta, Vishal K. and Nachiket M. Bhawe. (2007). The Influence of Proactive Personality and Stereotype Threat on Women's Entrepreneurial Intentions. *Journal of Leadership & Organizational Studies* 13 (4): 73–85.

Hannan, R. Lynn, Hoffman, Vicky B., and Donald V. Moser. (2005). Bonus Versus Penalty: Does Contract Frame Affect Employee Effort? In *Experimental Business Research*, Vol. 2, edited by Rami Zwick and Amnon Rapoport, Springer, US, 151-69.

Kahneman, Daniel and Amos Tversky. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47 (2): 263-92.

Lazear, Edward. (2000). Performance, Pay and Productivity. *American Economic Review* 90(5): 1346-61.

Luft, Joan. (1994. Bonus and Penalty Incentives: Contract Choice by Employees. *Journal of Accounting and Economics* 18 (2): 181-206.

May, Ann Mari, McGarvey, Mary G., and David Kucera. (2018). Gender and European Economic Policy: A Survey of the Views of European Economists on Contemporary Economic Policy. *KYKLOS*, 71 (1): 162–183.

Niederle, Muriel and Lise Vesterlund. (2007). Do Women Shy Away from Competition? Do Men Compete too Much? *Quarterly Journal of Economics* 122 (3): 1067–1101.

Niederle, Muriel. (2016). Gender. In *the Handbook of Experimental Economics,* Vol.2, John Kagel and Alvin E. Roth (eds.), Princenton University Press.

Paryevi, M., Bohnet, Iris, and Alexandra V. M. van Geen. (2016). Descriptive Norms and Gender Diversity: Reactance from Men. Working Paper.

Price, Curtis R. (2012). Gender, Competition, and Managerial Decisions. *Management Science* 58 (1): 114-122.

Rigdon, Mary. (2009). Trust and Reciprocity in Incentive Contracting. *Journal of Economic Behavior and Organization* 70 (1): 93-105.

Rosener, Judy B. (1990). Ways Women Lead. *Harvard Business Review* 68: 119-25.

Shih, Margaret, Pittinsky, Todd L., and Nalini Ambady. (1999). Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance. *Psychological Science* 10: 80–83.

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): 1189-1213.

Shurchkov, Olga, and Catherine C. Eckel. (2018). Gender Differences in Behavioral Traits and Labor Market Outcomes, in *The Oxford Handbook of Women and the Economy*, Susan L. Averett, Laura M. Argys, and Saul D. Hoffman, eds. (Oxford University Press, New York, NY.)

Spencer, Steven J., Steele, Claude M., and Diane M. Quinn. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology* 35 (1): 4-11.

Wiegand, Douglas M. and E. Scott Geller. (2008). Connecting Positive Psychology and Organizational Behavior Management. *Journal of Organizational Behavior Management* 24 (1-2): 3-25.

Zinovyeva, Natalia and Manuel F. Bagues. (2011). Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment. Institute for the Study of Labor (IZA) Discussion Paper No. 5537.

Table 1
Treatment summary

Treatment	Sessions with 3 Rounds			Sessions with 1 Round			Total		
	# Sessions	#Subjects		# Sessions	#Subjects		#Subjects		
		Male	Female		Male	Female	Male	Female	All
Gender	3	39	33	5	51	30	90	63	153
No gender	3	31	32	5	42	39	73	71	144
Total	6	70	65	10	93	69	163	134	297

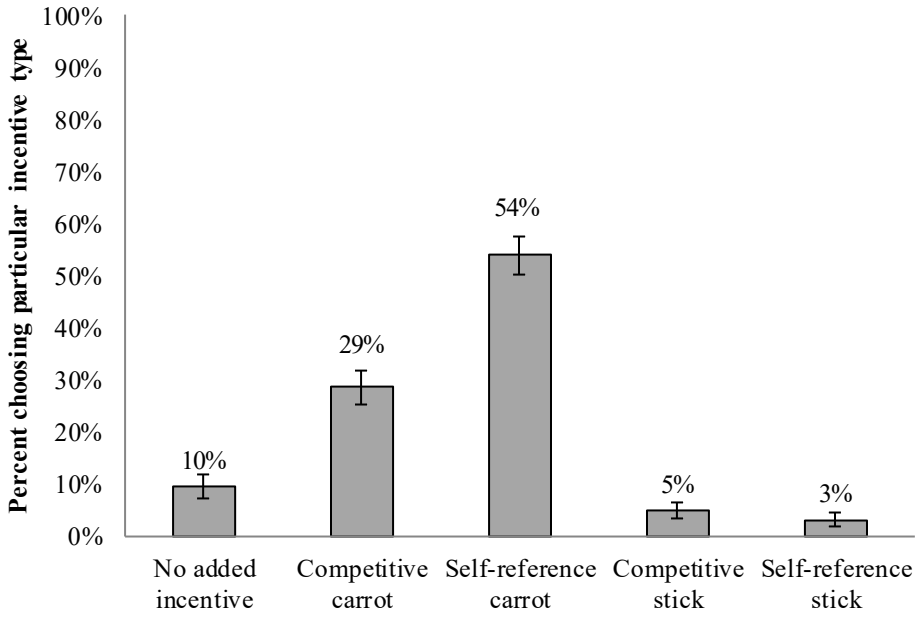

Figure 1: *Percentage of DMs choosing given incentive type (all sessions and treatments; 95% confidence intervals)*

Table 2
Summary statistics by gender

	Females		Males		Comparison	
Variable	Value	# Obs.	Value	# Obs	t-test p-val	U-test p-val
			Panel 1: Part 1 (Resume) Characteristics			
Mean resume addition score	9.88	131	9.97	162	0.843	0.967
% respondents living in Rotterdam	86%	134	74%	163	0.013**	0.014**
% respondents studying at Erasmus	99%	134	96%	163	0.248	0.247
% respondents in 19-25 age range	90%	134	85%	163	0.149	0.149
% respondents in social sciences	92%	134	81%	163	0.008***	0.008***
% respondents in Masters program	45%	134	39%	163	0.288	0.287
% respondents with GPA >7.5/10	38%	134	31%	163	0.183	0.182
			Panel 2: Preferences			
Mean risk lottery choice (out of 6 lotteries)	2.75	134	3.56	163	0.000***	0.000***
Loss aversion (loss - risk choice)	1.04	134	0.52	163	0.013**	0.004***
Ambiguity aversion (risk - ambiguity choice)	-0.23	134	-0.02	163	0.287	0.327
Mean pl. 1 trust transfer (trust) (Eu/10)	4.26	134	5.40	163	0.008***	0.015**
Mean pl. 2 trust transfer (reciprocity) (Eu/10)	4.95	134	5.12	163	0.592	0.362
Mean pl. 1 ultim. transfer (strategic) (Eu/10)	4.08	134	4.06	163	0.896	0.315
Mean dictator transfer (pure) (Eu/10)	2.10	134	2.29	163	0.472	0.992
			Panel 3: Demographic Characteristics			
Mean GPA (out of 10)	7.21	132	7.19	157	0.782	0.752
Mean age (years)	21.88	134	22.20	162	0.372	0.972
% citizen of the Netherlands	58%	134	65%	163	0.230	0.229
% White respondents	63%	134	72%	163	0.075*	0.075*
% with fluent or native English proficiency	58%	134	63%	163	0.445	0.444
% with family income >Eu65,000	19%	134	42%	163	0.000***	0.000***
% married	6%	134	6%	163	0.869	0.869
% with previous leadership experience	68%	134	67%	163	0.938	0.938
% with previous experience with games	63%	134	69%	163	0.340	0.339
% transformational leadership	83%	134	74%	163	0.058*	0.058*
% democratic leadership	87%	134	81%	163	0.141	0.141
% affiliative leadership	31%	134	25%	163	0.193	0.193
% commanding leadership	6%	134	13%	163	0.032**	0.033**
Final profit in the experiment (Eu)	16.98	134	17.26	163	0.429	0.475

Notes: See online appendix for detailed explanations of elicitation procedures and survey questions. Significance levels: * 10 percent, ** 5 percent, *** 1 percent.

Table 3
Determinants of incentive choice, aggregated incentive categories

Outcome variable:	Pr [No Added Incentive]			Pr [Carrot]			Pr [Competitive]		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
female DM dummy	0.063	0.157*		-0.0334	-0.0498		-0.157**	-0.380**	
	(0.055)	(0.089)		(0.0773)	(0.126)		(0.0776)	(0.141)	
treatment (gender known = 1)	0.105	0.133		-0.258	0.426		0.149	-0.548*	
	(0.132)	(0.237)		(0.164)	(0.451)		(0.308)	(0.314)	
female DM x treatment			0.345			0.305			-0.946**
			(0.278)			(0.445)			(0.418)
female DM x no treatment			0.144			-0.096			-0.304
			(0.123)			(0.187)			(0.252)
male DM x treatment			0.143			0.431			-0.499
			(0.246)			(0.478)			(0.366)
f-test p-value (treated vs. untreated female DM)			[0.428]			[0.385]			**[0.073]**
f-test p-value (treated female vs. treated male DM)			**[0.057]**			[0.444]			**[0.005]**
DM Preferences									
risk preference	-0.028	-0.083**	-0.082*	0.025	0.038	0.020	-0.001	0.050	0.057
	(0.027)	(0.039)	(0.045)	(0.030)	(0.053)	(0.065)	(0.037)	(0.064)	(0.069)
loss aversion	0.005	-0.041	-0.051	-0.010	0.014	0.027	0.027	0.038	0.043
	(0.018)	(0.036)	(0.038)	(0.023)	(0.047)	(0.056)	(0.028)	(0.043)	(0.053)
ambiguity aversion	0.018	0.064**	0.064**	-0.026	-0.082**	-0.057	-0.033	-0.094**	-0.098**
	(0.019)	(0.026)	(0.031)	(0.022)	(0.037)	(0.042)	(0.028)	(0.035)	(0.040)
dictator transfer	0.0001	0.003	0.012	0.004	0.054	0.038	-0.031	-0.041	-0.046
	(0.009)	(0.018)	(0.020)	(0.014)	(0.038)	(0.046)	(0.026)	(0.032)	(0.038)
ultimatum offer	-0.017	-0.051*	-0.053*	-0.017	-0.014	-0.022	-0.010	0.040	0.041
	(0.016)	(0.025)	(0.028)	(0.022)	(0.048)	(0.050)	(0.028)	(0.035)	(0.033)
ultimatum acceptance cutoff	-0.011	-0.016	-0.012	0.014	0.004	0.000	0.021	0.063**	0.060*
	(0.019)	(0.021)	(0.021)	(0.026)	(0.040)	(0.041)	(0.026)	(0.030)	(0.034)
trust	-0.006	0.001	-0.002	0.004	0.002	0.006	-0.020	-0.026	-0.027
	(0.007)	(0.016)	(0.016)	(0.009)	(0.023)	(0.025)	(0.015)	(0.022)	(0.026)
reciprocity	-0.009	-0.031	-0.028	0.021	0.044	0.028	0.042**	0.016	0.016
	(0.011)	(0.020)	(0.023)	(0.016)	(0.034)	(0.035)	(0.019)	(0.032)	(0.033)
Dependent Var. Mean:	0.09	0.07	0.07	0.83	0.83	0.83	0.33	0.36	0.36
Sample	All	Round 1	Round 1	All	Round 1	Round 1	All	Round 1	Round 1
Observations	175	94	94	175	94	94	175	94	94
R-squared	0.30	0.73	0.75	0.38	0.67	0.71	0.43	0.85	0.85

Notes: Coefficients estimated using a linear probability (fixed effects OLS) model. All treatments. Specifications include: Worker resume characteristics observable to the DM and relevant to incentive selection: Part 1 addition scores, worker major indicator, and worker GPA range; DM characteristics: race, Part 1 addition score, GPA, major, age, student status, university, city of residence, marital status, English language proficiency, income bracket, citizenship, leadership experience, game experience; and Session fixed effects. In specifications where the full sample is used, round fixed effects are included. Robust standard errors in parentheses. Significance levels: * 10 percent, ** 5 percent, *** 1 percent.

Table 4
Determinants of incentive choice in the first round, separate incentive categories

Outcome variable:	Pr [Comp Carrot]		Pr [Comp Stick]		Pr [Self Carrot]		Pr [Self Stick]	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
female DM dummy	-0.219		-0.162*		0.169		0.054	
	(0.175)		(0.092)		(0.170)		(0.091)	
treatment (gender known = 1)	-0.357		-0.190		0.783*		-0.368	
	(0.412)		(0.252)		(0.398)		(0.244)	
female DM x treatment		-0.622		-0.324		0.927*		-0.326
		(0.459)		(0.268)		(0.494)		(0.254)
female DM x no treatment		-0.154		-0.149		0.058		0.101
		(0.303)		(0.134)		(0.301)		(0.138)
male DM x treatment		-0.278		-0.220		0.710		-0.354
		(0.452)		(0.256)		(0.466)		(0.277)
f-test p-value (treated vs. untreated female DM)		[0.278]		[0.494]		**[0.078]**		[0.131]
f-test p-value (treated female vs. treated male DM)		**[0.043]**		[0.227]		[0.236]		[0.808]
Dependent Var. Mean:	0.31	0.31	0.05	0.05	0.52	0.52	0.04	0.04
Worker resume controls	YES	YES	YES	YES	YES	YES	YES	YES
DM demographic controls	YES	YES	YES	YES	YES	YES	YES	YES
DM risk and social preferences	YES	YES	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES	YES	YES
Observations	94	94	94	94	94	94	94	94
R-squared	0.80	0.83	0.64	0.69	0.77	0.78	0.63	0.65

Notes: Coefficients estimated using a linear probability (fixed effects OLS) model; round 1 data only. All treatments. Specifications include: Worker resume characteristics observable to the DM and relevant to incentive selection: Part 1 addition scores, worker major indicator, and worker GPA range; DM characteristics: Part 1 addition score, race, GPA, major, age, student status, university, city of residence, marital status, English language proficiency, income bracket, citizenship, leadership experience, game experience; DM risk and social preferences (risk preference, loss aversion, ambiguity aversion, dictator transfer, strategic altruism, trust, and reciprocity). Robust standard errors in parentheses. Significance levels: * 10 percent, ** 5 percent, *** 1 percent.

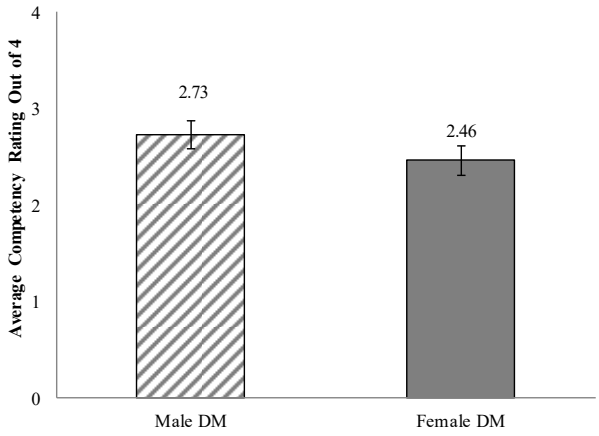

Figure 2: *Average DM competency rating (all sessions, incentives and treatments; 95% confidence intervals; t-test p-value of 0.011)*

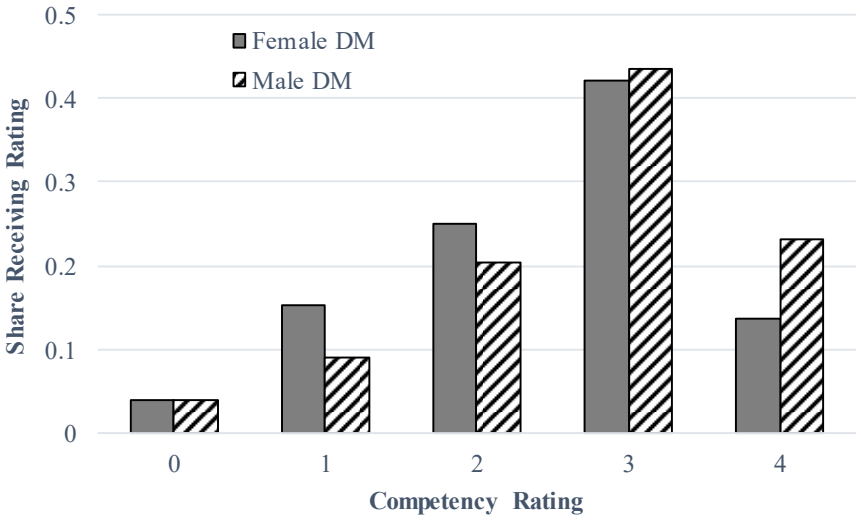

Figure 3: *Distribution of competency ratings by DM gender (all sessions and treatments)*

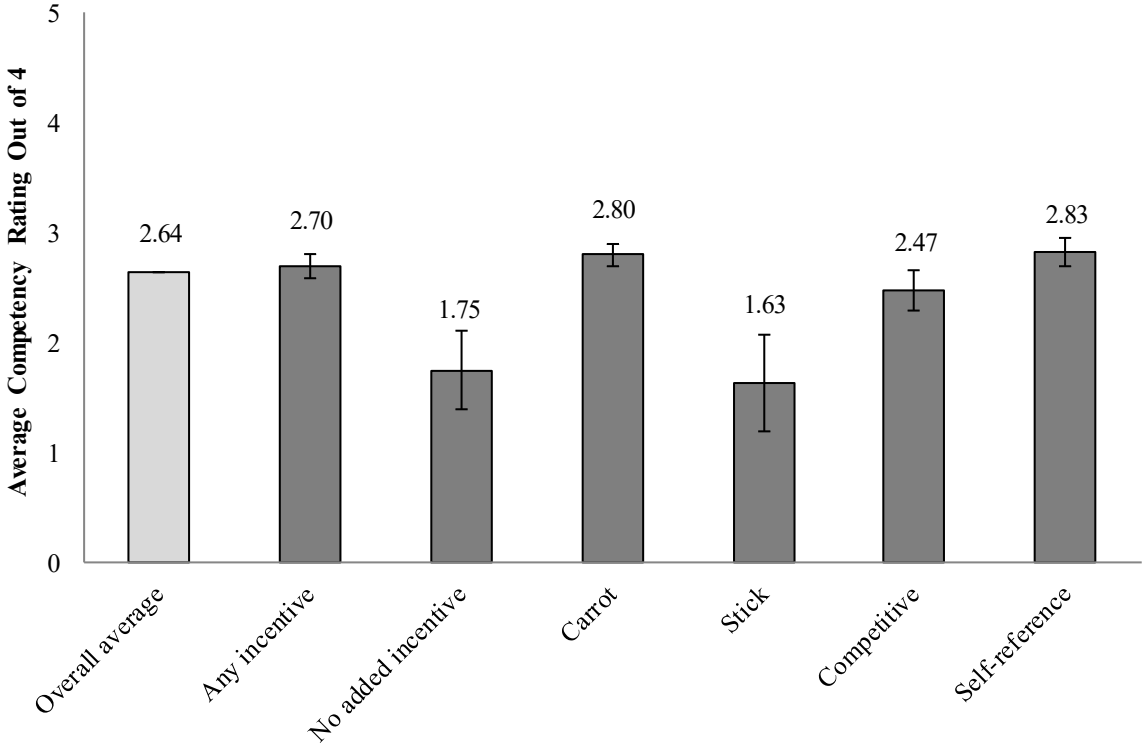

Figure 4: *Average DM competency rating by incentive type (all sessions and treatments; 95% confidence intervals)*

Table 5
Average competency rating by incentive type, treatment, and gender of the DM

Competency Rating	Male DM	Female DM	t-test p-value	Male DM	Female DM	t-test p-value
	Gender Treatment			*No Gender Treatment*		
Overall average	2.79	2.44	0.019**	2.66	2.48	0.251
Any incentive	2.92	2.54	0.015**	2.69	2.55	0.350
No added incentive	1.50	2.00	0.232	1.00	1.75	0.452
Carrot	2.96	2.62	0.024**	2.80	2.74	0.669
Stick	2.00	1.25	0.524	1.63	1.64	0.971
Competitive	2.61	2.23	0.20	2.61	2.39	0.36
Self-reference	3.04	2.73	0.0784*	2.75	2.67	0.70

Notes: Data pooled across all sessions and rounds. Significance levels: * 10 percent, ** 5 percent, *** 1 percent.

Table 6

Workers' rating of DM competency in the first round by treatment and worker gender, conditional on incentive choice

Outcome variable:	DM competency rating (0 - not at all competent; 4 - very competent)									
	Gender Treatment					*No Gender Treatment*				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
DM gender (1 = Female)	-0.353**	-0.312		-0.558*		-0.178	0.112		-0.09	
	(0.148)	(0.193)		(0.308)		(0.154)	(0.185)		(0.298)	
Worker gender (1 = Female)		0.206		0.399			0.345		0.536*	
		(0.186)		(0.314)			(0.211)		(0.267)	
DM female & worker male			-0.458**		-1.100***			0.248		-0.127
			(0.230)		(0.387)			(0.270)		(0.375)
DM female & worker female			-0.068		-0.046			0.472		0.433
			(0.256)		(0.381)			(0.315)		(0.470)
DM male & worker female			0.057		-0.206			0.474*		0.483
			(0.255)		(0.371)			(0.280)		(0.428)
DM choice:										
competitive carrot		0.930***	0.931***	0.204	0.255		1.068**	1.055**	1.689**	1.719**
		(0.333)	(0.324)	(0.590)	(0.593)		(0.458)	(0.465)	(0.705)	(0.723)
competitive stick		0.301	0.294	-1.652	-1.841		0.489	0.450	0.786	0.824
		(0.784)	(0.818)	(1.227)	(1.187)		(0.514)	(0.525)	(0.820)	(0.843)
self-reference carrot		1.442***	1.449***	0.551	0.626		1.226***	1.209***	1.586**	1.612**
		(0.273)	(0.267)	(0.478)	(0.498)		(0.434)	(0.439)	(0.760)	(0.769)
self-reference stick		-0.342	-0.274	-2.223*	-2.029*		0.040	0.001	1.185	1.217
		(0.915)	(0.910)	(1.236)	(1.137)		(0.611)	(0.638)	(1.139)	(1.163)
f-test p-value (carrots vs. sticks)		**[0.033]**	**[0.039]**	**[0.005]**	**[0.001]**		**[0.004]**	**[0.004]**	**[0.138]**	**[0.153]**
Dependent Variable Mean:	2.64	2.63	2.63	2.56	2.56	2.57	2.59	2.59	2.71	2.71
Treatment (DM gender shown?)	YES	YES	YES	YES	YES	NO	NO	NO	NO	NO
DM resume controls	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Worker characteristics	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Session FE	NO	YES	YES	YES	YES	NO	YES	YES	YES	YES
Round FE	NO	YES	YES	N/A	N/A	NO	YES	YES	N/A	N/A
Round	All	All	All	1	1	All	All	All	1	1
Observations	198	182	182	94	94	180	170	170	90	90
R-squared	0.03	0.47	0.47	0.74	0.78	0.01	0.41	0.42	0.73	0.73

Notes: Coefficients estimated using a fixed effects OLS model. Omitted category in specifications 3, 5, 8, and 10: DM male & male worker. DM resume controls include Part 1 addition score, GPA range, age bracket, student status (Bachelors or Masters), major, city of residence, and university. Worker characteristics include Part 1 addition score, race, age, GPA, major, student status, university, city of residence, marital status, English language proficiency, income bracket, citizenship, game experience and social preference measures (pure altruism, strategic altruism, trust, and reciprocity). Robust standard errors in parentheses. Significance levels: * 10 percent, ** 5 percent, *** 1 percent.

Table 7
Workers' rating of DM competency in the first round by treatment and choice of incentive

Outcome variable:	DM Competency Rating (0 - Not at all competent; 4 - Extremely competent)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Gender Treatment				*No Gender Treatment*			
	comp. carrot	comp. stick	self. carrot	self. stick	comp. carrot	comp. stick	self. carrot	self. stick
incentive & female DM (A)	0.153	-2.564**	0.038		0.435	-0.680	0.386	0.072
	(0.697)	(1.245)	(0.467)		(0.393)	(0.853)	(0.450)	(0.862)
incentive & male DM (B)	-0.134		0.754	-2.771**	0.752	-1.112	0.003	
	(0.547)		(0.466)	(1.128)	(0.536)	(0.843)	(0.595)	
different incentive & female DM (C)	-0.547	-0.402	-0.134	-0.484*	0.306	-0.062	-0.037	0.055
	(0.369)	(0.305)	(0.537)	(0.287)	(0.586)	(0.277)	(0.430)	(0.324)
f-test p-value (A vs B)	[0.665]	N/A	**[0.066]**	N/A	[0.513]	[0.750]	[0.545]	N/A
f-test p-value (A vs C)	[0.262]	**[0.095]**	[0.749]	N/A	[0.807]	[0.447]	[0.357]	[0.984]
f-test p-value (B vs C)	[0.391]	N/A	[0.101]	**[0.044]**	[0.292]	[0.227]	[0.933]	N/A
Dependent Variable Mean:	2.56	2.56	2.56	2.56	2.71	2.71	2.71	2.71
Treatment (DM gender shown?)	YES	YES	YES	YES	NO	NO	NO	NO
DM resume controls	YES	YES	YES	YES	YES	YES	YES	YES
Worker characteristics	YES	YES	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES	YES	YES
Round	1	1	1	1	1	1	1	1
Observations	94	94	94	94	90	90	90	90
R-squared	0.68	0.69	0.68	0.72	0.67	0.67	0.65	0.64

Notes: Coefficients estimated using a fixed effects OLS model. Omitted category in all specifications: different incentive & male DM. DM resume controls include Part 1 addition score, GPA range, age bracket, student status (Bachelors or Masters), major, city of residence, and university. Worker characteristics include Part 1 addition score, gender, race, age, GPA, major, student status, university, city of residence, marital status, English language proficiency, income bracket, citizenship, game experience and social preference measures (pure altruism, strategic altruism, trust, and reciprocity). Robust standard errors in parentheses. Significance levels: * 10 percent, ** 5 percent, *** 1 percent.

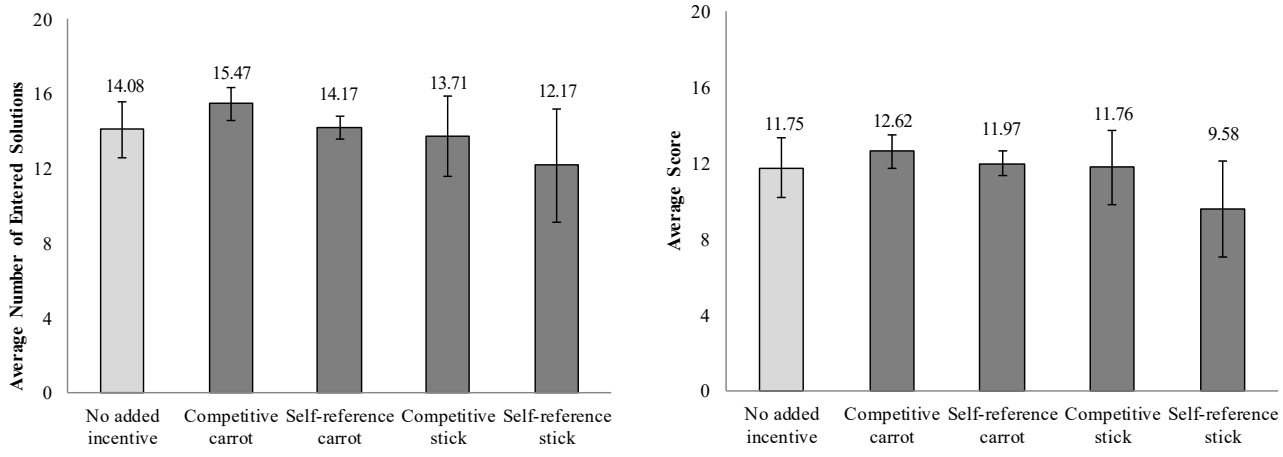

Figure 5: *Average number of entered solutions and average score by incentive type (all sessions and treatments; 95% confidence intervals)*

Table 8
Impact of DM gender and incentive choice on worker performance

Outcome variable:	Score (Number correctly solved additions)					
	Gender Treatment			*No Gender Treatment*		
	(1)	(2)	(3)	(4)	(5)	(6)
DM gender (1 = Female)	1.431**	0.788	-0.616	0.649	-0.163	0.399
	(0.633)	(0.690)	(0.843)	(0.700)	(0.663)	(2.290)
worker female	-1.796***	-0.763	-1.397	1.153	-0.202	-0.227
	(0.637)	(0.631)	(0.959)	(0.703)	(0.785)	(1.420)
DM choice:						
competitive carrot		1.031	-3.008*		0.256	3.131
		(1.175)	(1.515)		(2.127)	(3.576)
competitive stick		-1.316	-7.260**		-1.222	2.313
		(1.864)	(3.055)		(2.215)	(3.691)
self-reference carrot		-0.064	-1.834		0.279	0.575
		(1.098)	(1.280)		(1.858)	(3.567)
self-reference stick		0.787	-6.017*		-1.983	-0.970
		(1.971)	(3.242)		(2.065)	(3.427)
f-test p-value (carrots vs. sticks)		[0.535]	**[0.040]**		**[0.013]**	[0.553]
worker part 1 addition score		0.824***	0.902***		0.673***	0.674***
		(0.086)	(0.151)		(0.101)	(0.139)
Dependent Variable Mean:	11.79	11.70	11.83	12.33	12.23	11.40
Treatment (DM gender shown?)	YES	YES	YES	NO	NO	NO
DM resume controls	NO	YES	YES	NO	YES	YES
Worker characteristics	NO	YES	YES	NO	YES	YES
Session FE	NO	YES	YES	NO	YES	YES
Round FE	NO	YES	N/A	NO	YES	N/A
Round	All	All	1	All	All	1
Observations	197	181	93	178	168	89
R-squared	0.06	0.70	0.90	0.02	0.69	0.80

Notes: Coefficients estimated using a fixed effects OLS model. DM resume controls include Part 1 addition score, GPA range, age bracket, student status (Bachelors or Masters), major, city of residence, and university. Worker characteristics include Part 1 addition score, gender, race, age, GPA, major, student status, university, city of residence, marital status, English language proficiency, income bracket, citizenship, game experience, and risk and social preference measures (risk, loss, and ambiguity aversion, pure altruism, strategic altruism, trust, and reciprocity). Robust standard errors in parentheses. Significance levels: * 10 percent, ** 5 percent, *** 1 percent.

Table 9

Impact of different incentives on worker performance by gender of the DM

Outcome variable:	Score (Number correctly solved additions)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Gender Treatment				*No Gender Treatment*			
	comp. carrot	comp. stick	self. carrot	self. stick	comp. carrot	comp. stick	self. carrot	self. stick
incentive & female DM (A)	-0.980	-4.952*	-0.327		2.476	-3.284	-1.690	-2.080
	(1.328)	(2.816)	(1.012)		(2.801)	(3.592)	(2.172)	(2.534)
incentive & male DM (B)	-0.090		0.295	-3.687	1.504	3.731	-2.471	
	(1.460)		(1.242)	(2.821)	(2.169)	(2.950)	(2.488)	
different incentive & female DM (C)	-0.302	-0.339	0.200	-0.497	-0.597	1.590	-0.132	1.073
	(1.158)	(0.847)	(1.680)	(0.877)	(2.496)	(2.096)	(2.403)	(2.377)
f-test p-value (A vs B)	[0.544]	N/A	[0.587]	N/A	[0.737]	[0.165]	[0.788]	N/A
f-test p-value (A vs C)	[0.610]	**[0.097]**	[0.713]	N/A	[0.147]	[0.119]	[0.433]	[0.244]
f-test p-value (B vs C)	[0.871]	N/A	[0.956]	[0.282]	[0.338]	[0.507]	[0.478]	N/A
Dependent Variable Mean:	11.83	11.83	11.83	11.83	11.40	11.40	11.40	11.40
Treatment (DM gender shown?)	YES	YES	YES	YES	NO	NO	NO	NO
DM resume controls	YES	YES	YES	YES	YES	YES	YES	YES
Worker characteristics	YES	YES	YES	YES	YES	YES	YES	YES
Session FE	YES	YES	YES	YES	YES	YES	YES	YES
Round	1	1	1	1	1	1	1	1
Observations	93	93	93	93	89	89	89	89
R-squared	0.88	0.88	0.88	0.88	0.80	0.78	0.79	0.78

Notes: Coefficients estimated using a fixed effects OLS model. DM resume controls include Part 1 addition score, GPA range, age bracket, student status (Bachelors or Masters), major, city of residence, and university. Worker characteristics include Part 1 addition score, gender, race, age, GPA, major, student status, university, city of residence, marital status, English language proficiency, income bracket, citizenship, game experience, and risk and social preference measures (risk, loss, and ambiguity aversion, pure altruism, strategic altruism, trust, and reciprocity). Robust standard errors in parentheses. Significance levels: * 10 percent, ** 5 percent, *** 1 percent.