I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12611

Do Workers Discriminate against Female Bosses?

Martin Abel

SEPTEMBER 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12611

Do Workers Discriminate against Female Bosses?

Martin Abel
Middlebury College and IZA

SEPTEMBER 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Do Workers Discriminate against Female Bosses?*

I hire 2,700 workers for a transcription job, randomly assigning the gender of their (fictitious) manager and provision of performance feedback. While praise from a manager has no effect, criticism negatively impacts workers' job satisfaction and perception of the task's importance. When female managers, rather than male, deliver this feedback, the negative effects double in magnitude. Having a critical female manager does not affect effort provision but it does lower workers' interest in working for the firm in the future. These findings hold for both female and male workers. I show that results are consistent with gendered expectations of feedback among workers. By contrast, I find no evidence for the role of either attention discrimination or implicit gender bias.

JEL Classification: J50, J70

Keywords: gender discrimination, gig economy, female leadership

Corresponding author:
Martin Abel
Middlebury College
14 Old Chapel Rd
Middlebury, VT 05753
USA
E-mail: mabel@middlebury.edu

* This paper greatly benefited from discussions with and comments by Willa Brown, Tanya Byker, Jeff Carpenter, Erik Gong, Caitlin Myers, and Oyebola Okunogbe. Daniel Buchman provided superb research assistance. The experiment was registered under registry number AEARCTR-0003810 and IRB approval was obtained from Middlebury College. All errors and omissions are my own.

1 Introduction

Women have overtaken men in educational attainment and score higher on leadership competencies (Zenger and Folkman, 2019). Yet, while 45% of S&P 500 workers are female, women only make up 37% of managers at the mid-level, 26% at the senior-level and 5% of CEOs (Catalyst, 2019). In addition to raising equity concerns, this misallocation of talent can have severe negative effects on productivity and growth (Hsieh et al. (forthcoming)). Why then are women less likely to climb the corporate ladder? Explanations range from discrimination in hiring to the "mommy tax" to gender differences in competitiveness and risk-aversion (for reviews see Bertrand (2011), Blau and Kahn (2017) and Neumark (2018)).

A nascent literature investigates the role of discrimination by subordinates (Grossman et al., 2016; Ayalew et al., 2018). While hiring discrimination disadvantages firms in a competitive market (Arrow, 1971), discrimination by subordinates can create an equilibrium in which gender discrimination becomes a self-fulfilling prophecy (Loury, 2009): women may become less effective managers precisely because workers perceive them to be less effective, implying that it is rational for firms to promote men over equally or more qualified women.

I investigate whether workers discriminate against female managers through a real effort field experiment set in the "gig economy" .[1] Specifically, I employ a fictitious firm, which hires 2,700 U.S.-based workers via Amazon's Mechanical Turk (MTurk) platform to transcribe receipts. At the beginning of the task, I introduce workers to a fictitious manager who will remotely monitor their performance. I randomly assign the manager either a female or male name. At the task's halfway mark, this manager provides a random subset of workers with either positive or negative feedback based on their actual (above *or* below average) performance in the first part of the task. I then measure workers' performance (i.e. effort) in the second part of the task and elicit workers' job satisfaction, perception of the task's importance, and interest in working for the firm in the future (i.e. attitudes) at the end of the task.

Following a registered pre-analysis plan, the main findings are three-fold. First, in the absence of receiving any feedback, workers exert the same amount of effort regardless of manager gender. Second, feedback has a net negative effect on workers' attitudes; this is driven by the large negative effect of criticism, which outweighs the very modest positive effect of praise. Third, the negative effect of criticism on attitudes of both female and male workers is much larger when a female manager delivers the feedback; in fact, criticism from female managers doubles the share of workers not interested in working for the firm in the future and leads to a 70% larger reduction in job satisfaction than criticism from male managers.

These results have important implications for firms. Alexander (2006) points out that a

[1] The term refers to flexible work arrangements mediated through online platforms (Abraham et al., 2018).

worker's preoccupation with the identity of the person providing the feedback may prevent learning and diminish the worker's commitment and productivity. While I find evidence for reduced commitment to both the task and the firm, this does not translate into short-term behavior change. Neither female nor male workers change the amount of effort they exert in response to either praise or criticism, regardless of manager gender. MTurk's reputation system may be the reason why the drop in attitudes does not translate into a reduction of effort. Receiving criticism signals to workers that they are at the risk of getting their task rejected, which can have severe consequences for future work opportunities. [2] The literature, however, finds a robust positive relationship between job satisfaction and performance, suggesting the deterioration of attitudes are more detrimental for longer term supervisor-subordinate relationships, which rely on cooperation and learning (Judge et al., 2001; Eagly and Chaiken, 1993). In addition, worker retention is regarded as a key measure of match quality across many industries (Hoffman et al., 2017).

Why does criticism from women make workers more dissatisfied with the task and firm? I document that both female and male workers on MTurk have gendered expectations of certain management styles, confirming evidence from more traditional work settings by Carli (2001) and others. They are about three times more likely to associate giving praise and appropriate use of tone with female managers. By contrast, they are about twice more likely to associate giving criticism and strict expectations with male managers. Female managers who criticize workers are thus violating expectations, which may explain the large negative effects of such criticism from female managers on workers' attitudes.[3]

Notably, for my experiment, the extent to which these gendered expectations actually affect workers' perception of feedback appear to differ by worker gender. Male workers tend to dismiss the competence of their female managers after receiving negative feedback. They judge verbatim the same criticism 0.3 s.d. less accurate if it comes from a female manager. By contrast, female workers' perception of feedback does not vary by manager gender. These results suggest that the reason for *why* workers discriminate against female managers is gender-specific.

The rich data I collect enables me to test three other prominent hypotheses for gender discrimination that have not been empirically tested in the context of discrimination by subordinates. First, I provide evidence against the importance of so-called "attention discrimination" (Bartos et al., 2016), which holds that female managers are less effective because workers pay less attention to them, limiting their ability to motivate and change behavior. In fact, I find workers spend about about 8% *more* time processing feedback from female

[2] Most MTurk employers require an approval percentage of at least 98% to work on their tasks.

[3] This captures what Jamieson et al. (1995) called the femininity/competence double bind of female leadership. In the US, the stereotypical image of a leader is that of a man (Rudman and Kilianski, 2000). Women who adhere to this image face accusations of being "brusque" or cold, putting into question the success of strategies proposed in *Lean In* (Sandberg, 2015) and recent similar works. At the same time, those who choose a more feminine approach, however, often face accusations of incompetence or challenges to their authority (Jamieson et al., 1995).

managers. This holds across worker gender.

Second, I explore the role of implicit biases. Glover et al. (2019) find that implicit racial biases explain why white managers are less effective if matched with black workers. Following the transcription task, 830 workers in my sample complete an implicit associations test, which measures the extent to which they subconsciously associate women with family life and men with professional characteristics. While the test provides evidence of moderate levels of implicit gender bias within the sample, workers with stronger biases do not respond differently to female supervisors.

Third, I test for the role of previous experiences. Unfamiliarity with women in leadership positions, for instance, may explain why people respond more negatively to female managers (Beaman et al., 2009). This, however, does not hold in the sample, in which 86% of workers have worked under female supervisors. In fact, workers do not assess that previous female supervisors were less effective than their male counterparts. It is thus unsurprising that previous experience is not correlated with workers' responses to being assigned a female manager. These results also suggest that *"discrimination from below"* (Ayalew et al., 2018) may not disappear as workers grow more accustomed to (competent) female managers. I do find, however, that younger workers react less negatively to critical female managers and that gender discrimination completely disappears among workers in the 20s. In line with the importance of gendered expectations, these young workers are also less likely to associate critical feedback with male managers.

It is important to acknowledge limitations to this experimental design. While MTurk is one of the largest players in the gig economy, using this platform for research raises concerns about Hawthorne effects where workers are aware they are participating in research and thus change their behavior. I address this by leading workers to believe the MTurk requester is an actual firm hiring workers for a real effort task. Even after I disclose that the task was part of a research project, only 13% of workers guess the research focuses on the role of feedback and virtually no workers suspect the study is related to manager gender. I also show that treatment effects do not differ for people who suspect the task is related to research. Lastly, workers' awareness of participating in research testing for gender-based discrimination would likely bias the results *against* finding an effect.

An additional concern relates to the external validity of this study's findings. One key advantage of MTurk is that it offers a more controlled setting for studying the supervisor-subordinate relationship than more traditional workplaces where research is limited to using observational data.[4] Jobs in the gig economy are, however, distinct from "traditional" work

[4] This limitation also holds for education settings, in which recent studies document biases against female instructors in student evaluations (Buser et al., 2019; Boring, 2016). One exception is Sinclair and Kunda (2000) who conduct an experiment providing 54 male students with feedback on interpersonal skills. Similar to my results, students viewed women as less competent than men after receiving critical evaluations, while there is no difference for positive feedback.

arrangements. One important feature is that relationships between managers and workers tend to be more short-term and not involve face-to-face interactions. The extent to which results can be generalized to more traditional settings is therefore debatable. The gig economy, however, is an expanding part of the economy.[5] In addition, increasingly common remote work arrangements have similar limitations in supervisor-subordinate interactions (Bloom et al., 2014). In particular, it is harder for firms to monitor and incentivize worker effort, especially for tasks that involve worker discretion. My results highlight that managing workers remotely may be particularly challenging for female supervisors.

Existing studies on discrimination by subordinates use lab experiments to test whether workers follow the advice of a player of a randomly assigned gender (Grossman et al., 2016; Ayalew et al., 2018). In line with these studies, evidence from my field experiment shows that female managers face discrimination with important consequences for subordinates' job satisfaction and consequently firms' ability to retain workers. Discrimination is limited to negative feedback, confirming observational work suggesting women are disproportionately penalized for being disagreeable (Mueller and Plug, 2006). My findings also show that supervisors' tendency to give female workers less frequent and more vague feedback (Correll and Simard, 2016) because of concerns about adverse reaction to criticism is unfounded. In fact, women respond similarly to criticism with respect to attitudes and, if at all, *less* negatively with respect to effort.

This study also adds to an established literature on the effect of performance feedback. Evidence is mixed, with some papers finding negative effects on productivity (Eriksson et al., 2009; Kuhnen and Tymula, 2012; Hannan et al., 2012; Barankay, 2012) and other studies finding positive ones (Charness et al., 2014; Azmat and Iriberri, 2010; Bandiera et al., 2015; Blanes i Vidal and Nossol, 2011). I contribute to this literature by providing some of the first large-scale experimental evidence from the field and exploring novel causal mechanisms that may explain why results differ across study settings. This is also one of the first studies to test, experimentally, whether the effect of feedback depends on the identity of the feedback source.

Similar to other experimental studies in the literature, I randomly assign feedback receipt but not feedback content, which is endogenous and determined by a worker's individual task performance. Results should therefore be interpreted as effects of feedback for strong versus poor performers rather than the effect of positive versus negative feedback. As a methodological contribution, my experimental design includes a sharp discontinuity for positive and negative feedback around the average performance threshold, which I can exploit to estimate the causal effect of quasi-random feedback content. These results are similar in magnitude but less precise.

[5] The exact magnitude of the recent rise in popularity of the gig economy is hard to estimate given that these activities are under-reported in surveys such as the monthly CPS (Katz and Krueger, 2018). Most studies find at least a modest upward trend of U.S. workforce participation in alternative work arrangements throughout the 2000s (Katz and Krueger, 2018).

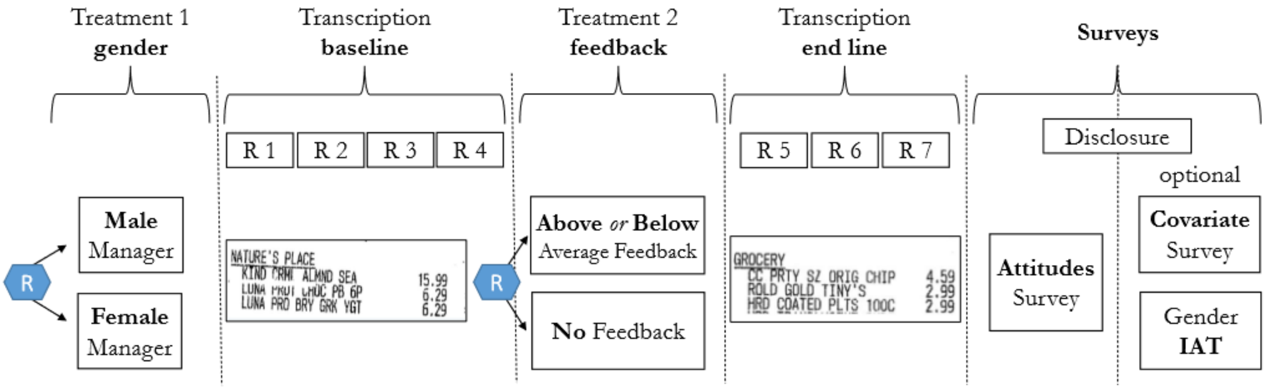
Given the availability of large amounts of data collected in real-time, organizations increasingly have the option to provide tailor-made, automated, instant performance feedback to their workers (Cecchi-Dimeglio, 2017). This study highlights potential limitations of this strategy. It also raises questions of how to mitigate gender-based discrimination among feedback recipients. Future research should explore the effectiveness of interventions such as raising awareness of discriminatory behavior, counseling workers on how to receive feedback, or sharing more information about the qualifications of their female managers.

The rest of the paper is structured as follows: Section 2 discusses the experimental design and empirical strategy. Section 3 presents results and Section 4 explores mechanisms. Section 5 concludes with a discussion of the findings.

2 Experimental Design

Figure 1 provides an overview of the experimental design. I recruit people on MTurk for a transcription task and randomize whether they work under a female or male manager. After transcribing four receipts, a random subset of workers receives feedback about whether they performed above or below average. Workers then transcribe an additional three receipts before completing an endline survey that collects data about their attitudes. I then disclose the purpose of the study, before collecting data on background characteristics and administering a gender and career IAT.

Figure 1: Design Overview



2.1 Recruitment and Sample Characteristics

Mechanical Turk (MTurk) is an online platform that allows firms to recruit workers for simple tasks, known as Human Intelligence Tasks (HITs). MTurk has been widely used in market research and is becoming increasingly popular as a platform for academic research (see for example (DellaVigna and Pope, 2017, 2018; List, 2017)).

To ensure data quality and prevent bots from completing the task, I follow existing studies and exclude workers with HIT approval ratings below 95% and limited previous experience (less than 100 completed HITs). Data collection is further divided into small batches of 30 to 100 workers posted at different times on different days. These batches filled up quickly, minimizing risks that our HIT is discussed and shared in online fora.

Workers are paid a flat rate of $1.75 rather than a piece rate for each completed receipt.[6] Paying a flat rate relies more on workers' intrinsic motivation.[7] Activating intrinsic motivation is regarded as one of the key characteristics of effective managers (Grant, 2008). This contractual arrangement also mimics trends in work arrangements, especially when individual output is difficult to measure. However, workers still have an incentive to perform well because the HIT requester can choose not to pay MTurk workers if we deem the quality of work inadequate, which has both direct financial harms as well as first order reputational effects.

Table 1 (col. 1 and 2) provides descriptive characteristics of the study sample. I recruit a total of 2,714 workers of which 97% complete the task. As is typical for workers in the gig economy, the sample is younger and more educated than the average worker in the labor force. 53% of workers are male, the average age is 35.6 and 62% (49%) completed at least a two (four) year college. 71% of our workers are white and 19% are black.[8] The average gender implicit bias score is 0.33 indicating moderate degrees of bias with respect to gender and career. 86% of workers have had a female manager in the past, slightly below the rate for male managers (96%).

In reviewing the literature on the effects of feedback, Straub et al. (2014) highlights the need for field experiments using realistic tasks and real-world settings to increase external validity. Participants in this study are therefore led to believe that they are working for an actual firm. The introduction of the task states that the transcription task can "*help businesses understand ... how to spend money better so they can increase their bottom line*". The HIT recruitment does mention that "*performance data will be recorded for research pur-*

[6] The implied average hourly rate of $8.75 is in line with the wage typically paid on MTurk (List, 2017).

[7] Existing studies find conflicting evidence on how the payment scheme affects the impact of feedback. Azmat and Iriberri (2010) and Hannan et al. (2012) find that feedback improves performance when paid a piece rate but has no effect under fixed rates. By contrast, Charness et al. (2014) and Eriksson et al. (2009) find no effect of feedback under a piece rate scheme.

[8] I over-sample black workers to facilitate analysis by worker race used in a companion paper.

poses" (see Appendix A.3 for the complete recruitment and introduction protocol). However, with the sharp increase in the use of "people analytics", these type of data collection arrangements become increasingly common, not just from customers but also from workers (Adler-Bell and Miller, 2018; Collins et al., 2017).

It appears that participants took the task seriously, regularly emailing with questions and comments about our HIT. As discussed in more detail below, the majority of participants did not suspect that this task was part of a research project and very few guessed the specific research question once we disclosed this fact.

Table 1: Sample Characteristics and Balance Test

			Manager Gender			Feedback		
	N	Mean	Female	Male	p-value	Yes	No	p-value
Age	2487	35.6	35.7	35.5	.59	35.6	35.7	.82
Worker male	2641	.53	.51	.55	.02	.54	.52	.34
College degree	2462	.62	.62	.61	.46	.62	.61	.35
Worker Black	2537	.19	.19	.19	.84	.18	.2	.35
Worker White	2537	.71	.71	.72	.61	.72	.71	.57
Prior female manager	1315	.86	.86	.86	.78	.85	.86	.58
Prior male manager	1314	.96	.96	.95	.39	.96	.95	.89
Gender Implicit Bias	822	.32	.33	.32	.55	.32	.32	.95
Quit job early	2714	.03	.03	.03	.18	.03	.02	.1
Manager black	2707	.24	.24	.24	.74	.24	.24	.93
Female manager	2707	.5	1	0	0	.5	.5	.79
Feedback	2708	.59	.6	.59	.79	1	0	.
Joint Significance					.48			0.80

Notes: The first two column show the sample size for each variable and the sample mean. The next three column compare characteristics across the random *manager gender* assignment and provide a p-value for a test of equal means. The last three column compare characteristics across the random *feedback* assignment and provide a p-value for a test of equal means.

2.2 Manager Names and Gender

At the end of the introduction, I randomize half of workers to have a female versus a male supervisor. Specifically, the introduction reads "*Our manager NAME might check in with you during the task.*" Table 1 (col 3, 4, 5) suggest that the randomization was successful. Of 10 covariates, only one is imbalanced. We can reject that the differences between the two groups are jointly significant (p-value: 0.48).

A common criticism of audit studies is that employers have more associations with names than merely the characteristic that researchers try to manipulate. For example, the seminal study on race-based discrimination in hiring by Bertrand and Mullainathan (2004) compares applications from resumes with distinctively White names (e.g. Emily and Greg) to identical resumes with distinctively Black names (e.g. Jamal and Lakisha) and finds that the former group receives about 50% more employer callbacks. One methodological critique is that aside from race, employers may associate job seekers with distinctively Black names with observable characteristics other than race (e.g. lower educational attainment) (Fryer Jr and Levitt, 2004).

I address this important concern by eliciting race, age, and education associations people have with various names in an out-of-sample MTurk survey.[9] I then match eight pairs of male and female names so that associations with respect to race, age, and education are balanced across gender (see Appendix Table A1): Brittany, Chloe, Christine, Ebony, Emily, Jennifer, Lynn and Shanice for women and Darius, Doug, Ethan, Josh, Justin, Michael and Tyrone for men. This matching exercise makes it less likely that any differences in worker behavior are driven by statistical discrimination.

2.3 Transcription Task and Outcomes

One challenge is to choose a real-effort task that is gender-neutral as women are perceived to perform worse in male-stereotyped domains (Sarsons, 2017a). The transcription task includes both language and math components and may thus be seen as more gender neutral (Niederle and Vesterlund, 2007). Women and men perform indeed equally well in this task (Table A2). Transcription scores are also not correlated with age or education, suggesting that performance depends on effort rather than experience or skills. Lastly, transcription is attractive for research purposes, because it is a common MTurk task and provides an objective measure of effort and quality.

2.3.1 Effort outcomes

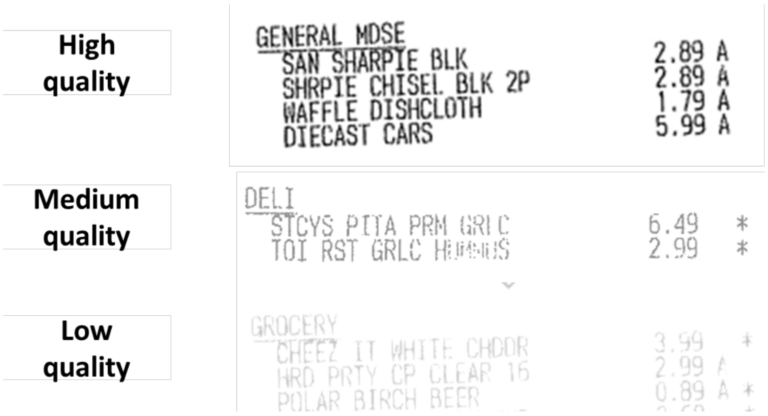
I use actual receipts but shorten them to include 3-5 items and manipulate their legibility (see Figure 2 for examples). Following List (2017), I first ask workers whether a given receipt is legible. The task description states that workers should only skip the receipt "*...if it is genuinely illegible*". Across all seven receipts, an average of 70% of receipts were deemed legible.

[9] Specifically, we chose 32 names with similar average annual income levels and asked for each name (in random order) *"Imagine you heard a name without any additional information. Based on the name, what do you think is the AGE / EDUCATION / RACE of that person? If it could be any, select 'UNCLEAR'."*

The share of correctly transcribed items provides a measure of accuracy. I convert performance to a score with each correctly transcribed item counts as 1 and each price as 0.5.[10] Across all receipts, the average score is 64%. For each receipt that is deemed legible, I tell workers "*If you have a calculator available, add the prices of all the items.*" This framing provides workers with an excuse not to exert this extra effort. I also emphasize that this is "*not a required task and will not affect your payment*". Over all seven receipts, workers completed this voluntary task for 48% of all receipts (of which almost all were added correctly).

These measures of effort serve different purposes: accurate transcription of receipts is the mandatory part of the task. By contrast, adding of numbers is a voluntary task that measures whether workers go above and beyond what is required from them. Activating intrinsic motivation so that workers exert effort on tasks that are hard to measure or difficult to contractually specify is a key quality of successful managers. The combination of voluntary and mandatory tasks also allows us to test whether (strict) monitoring on contractually specified output crowds out intrinsic motivation (Deci, 1971) and address concerns about multitasking (Holmstrom and Milgrom, 1991).

Figure 2: Receipt Examples


High quality
Medium quality
Low quality

GENERAL MDSE
SAN SHARPIE BLK 2.89 A
SHRPIE CHISEL BLK 2P 2.89 A
WAFFLE DISHCLOTH 1.79 A
DIECAST CARS 5.99 A

DELI
STCYS PITA PRM GRLC 6.49 *
TOI RST GRLC HUMMUS 2.99 *

GROCERY
CHEEZ IT WHITE CHDDR 3.99 *
HRD PRTY CP CLEAR 16 2.99 A
POLAR BIRCH BEER 0.89 A *

Notes: The graph shows receipts of varying quality that workers are asked to transcribe

2.3.2 Attitude

One of the goals of management is to "evoke emotions which can be used to organizational ends" (Ashforth and Humphrey, 1995). I collect data on attitudinal outcomes in a survey

[10] For receipts deemed illegible, we set values for subsequent effort measures (adding, transcription accuracy) to zero to account for potential selection effects. Results are unchanged in magnitude and significance if I use the scoring rule used for the worker feedback, which assigns 60% of possible points for skipping a receipt. This rule reflects that very poor performance is often more detrimental than not completing the task at all. In fact, companies using gig economy workers spend substantial resources on verifying completed work.

after worker completed the task and before the debrief.[11]

Worker retention is a key challenge for many firms. Turnover is particularly costly in the presence of learning on the job and if there are high training or selection costs (Hoffman et al., 2017). I measure effects on retention by asking workers if they are interested in working for our firm in the future. In addition, I collect two attitude outcomes pertaining to the actual task: whether workers were satisfied with the task and whether they were convinced that the task was important.[12] Table A2 shows that female workers are on average more satisfied with the job and are more likely to report that the task was important. By contrast, workers with a college degree are less satisfied and do not find the task to be important.

Last, I collect data on the perception of the manager feedback. Specifically, I ask workers in the treatment group if they think the feedback was accurate and if the tone was appropriate.

2.4 Feedback

After transcribing four receipts, 60% of workers are randomly assigned to receive either positive or negative feedback, depending on whether their performance was above or below average.[13] Table 1 suggests that the randomization was successful. I can reject that the differences between the two groups are jointly significant (p-value: 0.85).

The exact text for **positive** / **negative** feedback reads as follows:

> *Hello,*
> *This is NAME. As mentioned in the task introduction, I'm overseeing your performance in transcribing the receipts.*
>
> *I just went over some of the receipts. Your performance has been* ***above*** */* ***below*** *average. I was* ***pleased with*** */* ***disappointed by*** *your effort and attention to detail.*
>
> *Going forward, remember that your* ***continued commitment will improve*** */* ***lack of commitment will harm*** *the quality of our services.*
> *NAME*

This text includes standard components of feedback. It starts with a reminder about the manager's supervisory role in the task. Then, worker are informed about their performance

[11] 99.2% of workers complete the survey. To account for order effects, I randomize the order in which these questions are asked. See Appendix A.3 for the complete survey.

[12] I initially aimed to also collect data on stress levels. However, I updated the pre-analysis plan to exclude this measure as effects on stress among workers is an ambivalent outcome. At low levels of stress it can be beneficial, at high levels it can be detrimental (Muse et al., 2003).

[13] The scoring was automated so that we could provide instant feedback.

(above vs. below), followed by an associated sentiment (being pleased vs. disappointed). Finally, the feedback lays out consequences of actions for future outcomes (improve vs. harm services).

2.5 Estimation Strategy

As described, the experimental design features two stages of randomization: the gender of the manager (*FemMgr*) and whether workers receive feedback (*Feedback*). One can easily estimate reduced form effects by comparing outcomes between groups with female and male managers and between groups that did and did not not receive feedback. To test whether the effect of feedback depends on the gender of the manager, I estimate:

$$y_i = \beta_0 + \beta_1 FemMgr_i + \beta_2 Feedback_i + \beta_3 FemMgr\, x\, Feedb_i + \epsilon_i \tag{1}$$

y_i measures outcomes y for worker i. As specified in the pre-analysis plan, I separate outcomes y into measures of effort (legibility, adding, score) and attitude (interest in future work, task satisfaction, task importance). I also combine the three attitude measures into a standardized index to allay multiple hypothesis testing concerns (Kling et al., 2007). β_1 estimates the effect of a female manager who does not provide feedback, β_2 the effect of feedback from a male supervisor, and β_2+β_3 the effect of feedback from a female supervisor. All of these estimates are well-identified.

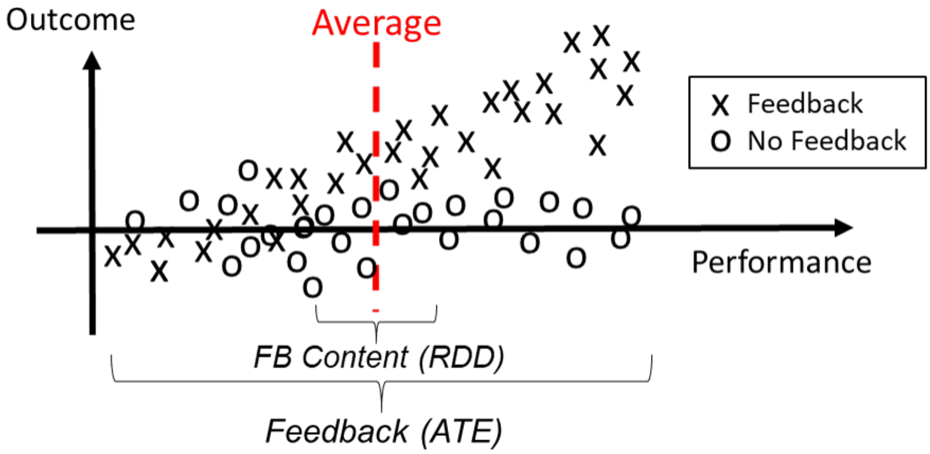
To estimate the effect of feedback *content*, I use the following specification:

$$y_i = \gamma_0 + \gamma_1 AboveAv_i + \gamma_2 Feedback_i + \gamma_3 AboveAv\, x\, Feedb_i + \epsilon_i \tag{2}$$

AboveAv is a dummy indicating whether the worker performed above average. Even though feedback content is determined by previous performance and thus endogenous, γ coefficients are well-identified, but require nuanced interpretation. Figure 3 provides a schematic diagram: for each worker in the treatment group (**x**), there is a worker with the same baseline performance in the control group (**o**) who does not receive feedback. γ_2 therefore captures the average treatment effect (ATE) of feedback on *low-performing* workers and γ_2+γ_3 the ATE of feedback on *high-performing* workers rather than the effect of negative and positive feedback, respectively.

However, the study design also allows me to estimate the effect of (quasi-)random performance feedback by exploiting the sharp discontinuity in feedback content (above vs. below average) at the threshold through a regression discontinuity design (RDD). Intuitively, I can compare responses of workers to feedback around the cutoff as performance just below or above the threshold is as good as random.

Figure 3: Identification (Schematic Diagram)



Notes: Data example to estimate the average treatment effect of feedback as well as the effect of feedback content through RDD from observations around the average threshold.

3 Results

Following the pre-analysis plan, the analysis is structured as follows: I first show the effect of manager gender on effort in the absence of feedback. Next, I look at the effect of feedback and the role of feedback content. Last, I explore the role of manager gender and how it interacts with feedback content. For each section results are presented for the full sample and then dis-aggregated by worker gender.

3.1 Manager Gender and Effort Provision

Do effort levels depend on manager gender in transcribing the first four receipts, i.e. before workers receive feedback? Table 2 (col. 1, 3, 5) shows that the coefficients of having a female managers are positive but statistically insignificant and small in magnitude (less than 0.05 standard deviations). I also do not find that the effect of having a female manager depends on the gender of the worker (col. 2, 4, 6). Female workers tend to put in slightly more effort working under a female manager, but these differences are not significant (p-value reported in bottom row). These estimates are very precise. I can rule out that there are even modest levels of gender discrimination in effort provision in the absence of feedback.

Table 2: Effect of Manager Race on Effort (Baseline)

	Legible		Adding		Transcription Score	
	(1)	(2)	(3)	(4)	(5)	(6)
Female manager	0.012	0.012	0.007	-0.006	0.157	0.068
	(0.008)	(0.012)	(0.016)	(0.022)	(0.169)	(0.244)
Female worker	0.010	0.010	0.019	0.006	0.209	0.113
	(0.008)	(0.012)	(0.016)	(0.022)	(0.168)	(0.238)
Fem mgr x Fem wkr		0.000		0.026		0.191
		(0.017)		(0.032)		(0.337)
Observations	2642	2642	2642	2642	2642	2642
Sample Mean	0.81	0.81	0.60	0.60	14.97	14.97
Std Dev	0.25	0.25	0.42	0.42	4.61	4.61
P-v: $\beta_2+\beta_3=0$		0.27		0.36		0.26

Notes: The dependent variable in Column (1) and (2) is whether the worker says the receipt is legible. The dependent variable in Column (3) and (4) measure if the worker is willing to add up the amounts. The dependent variable in Column (5) and (6) captures the accuracy of transcribing receipts All estimations are OLS. Robust standard errors are in parentheses. *P-v* presents the p-value of testing if female managers have a positive effect for female workers. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.2 Effect of Feedback Content

This section explores the effect of feedback, both in aggregate and split by praise vs. criticism. I will first look at the effects on worker attitude and then whether effects on attitude translate into changes in effort.

3.2.1 Attitude

Table 3 shows that aggregate effects of feedback are negative and statistically significant at the 1 percent level (Panel A, col. 1, 3, 5, 7). Effect sizes are similar for all three outcomes, ranging between 0.15 and 0.2 s.d..

This overall negative effect is driven by the large negative effects of criticism: it reduces interest in working in the future by 0.35 s.d., job satisfaction by 0.5 s.d., task importance by 0.25 s.d. and the aggregate index by 0.45 s.d.. (Panel A, col. 2, 4, 6, 8). Interestingly, the effect of positive feedback on outcomes is very small for all outcomes and not statistically different from zero (Panel A, col. 2, 4, 6, 8).[14] One explanation for these results is that

[14] This is unlikely to be solely a mechanical ”ceiling effect” since we observe the same for task importance

workers expect positive feedback for their work and thus do not change their attitudes in response to praise. Indeed, I will show below that workers deem negative feedback to be very inaccurate.

One surprising result is that feedback has a negative effect on the perceived importance of the task (col. 5 and 6). This is driven by workers downplaying the importance of the task in response to criticism, which appears to be particularly common among female workers, although the gender difference is not statistically significant significant (p-value: 0.18).

The effects of praise vs. criticism on attitudes are similar in magnitude for regression discontinuity estimates (Table A4, Panel B), suggesting that the effect of feedback for high vs. low-performing workers is similar to the causal effects of praise versus criticism. However, these coefficients are estimated off a subset of workers and are thus less precise.

3.2.2 Effort

Next, I test whether lower attitudes in response to feedback in general and criticism in particular affects the level of effort workers put into transcribing receipts. Table 4 shows that the aggregate effect of feedback is negative and statistically insignificant (Panel A, col. 1, 3, 5). Compared to equivalent estimates for attitude outcomes of around 0.2 s.d., coefficients on effort are small in magnitude: 0.05 s.d. for the voluntary task and 0.03 s.d. for the mandatory task. Effects of praise on effort tend to be more positive, but these differences are small and not statistically significant (Panel A, col. 2, 4, 6). RDD estimates are very similar in magnitude (Table A4), suggesting that this is not the mechanical result of ceiling effects or reversion to the mean.

There are some notable differences between female (Panel B) and male workers (Panel C). While aggregate effects of feedback on effort are small and positive for women, they are negative and moderate in magnitude (about 0.1 s.d.) for male workers. While only marginally significant, results suggest that men lower effort in response to criticism. These gender differences are strongest for the voluntary task: male workers lower their effort by 0.15 s.d. while women increase effort by 0.1 s.d. (p-value of gender difference: 0.029).

It is noteworthy that praise does not affect attitudes nor behavior. One explanation is that workers are expecting positive feedback. In fact, workers perceive praise as a highly accurate reflection of their performance as discussed in more detail below.

which has a lower average response.

Table 3: Effect of Feedback Content on Attitudes by Worker Gender

	Work Future		Job Satisf.		Task Import		Index (std)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Full Sample								
Feedback	-0.032***	-0.074***	-0.231***	-0.507***	-0.240***	-0.406***	-0.167***	-0.346***
	(0.008)	(0.013)	(0.043)	(0.070)	(0.064)	(0.097)	(0.029)	(0.047)
Abv Avg x Feedb		0.083***		0.545***		0.323**		0.353***
		(0.015)		(0.085)		(0.128)		(0.057)
Observations	2642	2642	2642	2642	2642	2642	2642	2642
P-v: pos FB=0		0.24		0.43		0.32		0.83
Panel B: Female Workers								
Feedback	-0.033***	-0.080***	-0.256***	-0.574***	-0.330***	-0.565***	-0.193***	-0.410***
	(0.012)	(0.021)	(0.062)	(0.102)	(0.088)	(0.135)	(0.043)	(0.071)
Abv Avg x Feedb		0.092***		0.622***		0.455***		0.424***
		(0.024)		(0.122)		(0.176)		(0.085)
Observations	1234	1234	1234	1234	1234	1234	1234	1234
Panel C: Male Workers								
Feedback	-0.031***	-0.069***	-0.209***	-0.452***	-0.161*	-0.272*	-0.144***	-0.292***
	(0.010)	(0.017)	(0.060)	(0.096)	(0.092)	(0.139)	(0.039)	(0.062)
Abv Avg x Feedb		0.076***		0.483***		0.217		0.294***
		(0.020)		(0.120)		(0.184)		(0.078)
Observations	1408	1408	1408	1408	1408	1408	1408	1408
Sample Mean	0.96	0.96	4.84	4.84	3.99	3.99	-0.05	-0.05
Std Dev	0.20	0.20	1.15	1.15	1.70	1.70	0.75	0.75
P-v: FB: M=F	0.906		0.585		0.182		0.401	
P-v: Neg FB: M=F		0.691		0.382		0.130		0.213
P-v: Pos FB: M=F		0.711		0.866		0.743		0.852

Notes: Work future is a binary variable whether worker are interested to work for the firm in the future. *Job satisf.* and *Task Import* measure how strongly workers agree that they were satisfied with the task and that the task was important, respectively (1=strongly disagree, 6= stronly agree). *Abv Avg* is a dummy if the worker performed above average. Panel B and C estimates effects separately by worker gender. All estimations are OLS. Robust standard errors are in parentheses. P-values report tests of equal coefficients across worker gender.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

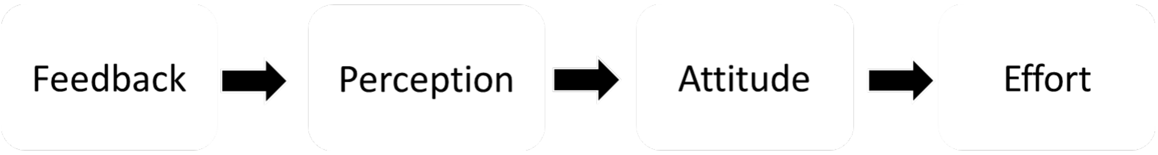
Table 4: Effect of Feedback Content on Effort by Worker Gender

	Legible		Adding		Transcription Score	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Full Sample						
Feedback	-0.008	-0.018	-0.020	-0.014	-0.170	-0.333
	(0.010)	(0.016)	(0.016)	(0.023)	(0.168)	(0.262)
Abv Avg x Feedb		0.025		-0.007		0.401
		(0.019)		(0.032)		(0.312)
Observations	2642	2642	2642	2642	2642	2642
P-v: pos FB=0		0.49		0.33		0.69
Panel B: Female Workers						
Feedback	0.010	0.005	0.003	0.039	0.063	-0.029
	(0.013)	(0.022)	(0.023)	(0.033)	(0.231)	(0.363)
Abv Avg x Feedb		0.013		-0.067		0.222
		(0.025)		(0.046)		(0.435)
Observations	1234	1234	1234	1234	1234	1234
Panel C: Male Workers						
Feedback	-0.024*	-0.038	-0.041*	-0.061*	-0.378	-0.585
	(0.014)	(0.023)	(0.022)	(0.031)	(0.242)	(0.376)
Abv Avg x Feedb		0.034		0.046		0.536
		(0.027)		(0.044)		(0.445)
Observations	1408	1408	1408	1408	1408	1408
Sample Mean	0.76	0.76	0.55	0.55	11.44	11.44
Std Dev	0.26	0.26	0.41	0.41	4.51	4.51
P-v: FB: M=F	0.075		0.18		0.187	
P-v: Neg FB: M=F		0.178		0.029		0.287
P-v: Pos FB: M=F		0.249		0.748		0.472

Notes: The dependent variable in Column (1) and (2) is whether the worker says the receipt is legible. The dependent variable in Column (3) and (4) measure if the worker is willing to add up the amounts. The dependent variable in Column (5) and (6) captures the accuracy of transcribing receipts Column (7) and (8) use the score as the dependent variable and divide the sample by worker gender. All estimations are OLS. Robust standard errors are in parentheses. P-values report tests of equal coefficients across worker gender. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.3 Effect of Manager Gender

This section explores the role of manager gender and how it interacts with feedback content and worker gender. Results are organized following the causal chain sketched out below to test how feedback may affect behavior differently depending on the gender of the manager. First, manager gender may affect how workers *perceive* feedback. Differences in perception may translate into differences in *attitude*, which may in turn affect *effort.*



This will be estimated through specifications with triple interactions of manager gender, feedback, and feedback content.[15] For greater transparency, I will also present non-parametric results graphically. One caveat for this analysis is that estimates will be less precise, although the average sub-group size of 330 workers is still large compared to other studies in this literature.

3.3.1 Perception of Feedback

Workers are asked how much they agree with the following statements (0=strongly disagree, 6=strongly agree): *"The feedback I received was accurate."* and *"The tone of the feedback was appropriate."* In aggregate, feedback from female managers is seen as equally accurate (Table 5, col. 1) and appropriate (col. 4).[16] However, differences emerge when I look at the interaction of content and manager gender. Criticism is assessed to be 0.17 s.d. less accurate (col. 2) and 0.11 s.d. less appropriate (col. 6) if it comes from female managers (coefficient only significant for accuracy). By contrast, there are no manager gender differences for praise.

Differences in response to criticism by female managers are exclusively driven by male workers. The left panel of Figure 4 shows that male workers rate verbatim the same negative feedback 0.3 s.d. less accurate if it comes from a female manager (p-value: 0.015). By

[15] Specifically, I will estimate the following equation.

$$\begin{aligned} y_i = \delta_0 + \delta 1_1 AboveAv_i + \delta_2 FB_i + \delta_3 FemMgr_i + \delta_4 AboveAv\, x\, FB_i + \delta_5 AboveAv\, x\, FemMgr_i + \\ \delta_6 FemMgr\, x\, FB_i + \delta_7 FemMgr\, x\, FB_i\, x\, AboveAv + \epsilon_i \end{aligned} \tag{3}$$

For readability reasons, I will only report coefficients of interest.

[16] Unsurprisingly, accuracy depends on the content of the feedback: positive feedback is perceived to be 1.3 s.d. more accurate than negative feedback. There is a similar level of perceived accuracy and appropriateness for workers in the left and right tail of the performance distribution suggesting that the wording of the two messages are perceived similarly fair for very strong and very weak performers.

contrast, female workers assess criticism independent of manager gender (p-value: 0.95). [17] For positive feedback, workers assess praise as 0.1 s.d. more accurate if it comes from a manager of the opposite sex. These results are also presented in regression format in Table 5 column 3 and 4. One caveat is that these results are estimated imprecisely and that the worker gender difference is only marginally significant (p-value: 0.102).

Table 5: Effect of Feedback Content on Feedback Attitude

	Feedback Accurate (std)				Feedback Appropriate (std)			
	Full	Full	Women	Men	Full	Full	Women	Men
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female manager	0.024	-0.166*	-0.008	-0.303**	0.063	-0.113	-0.015	-0.193
	(0.069)	(0.090)	(0.130)	(0.124)	(0.069)	(0.094)	(0.138)	(0.128)
Above x Fem Mgr		0.177*	-0.076	0.392***		0.149	-0.027	0.287**
		(0.101)	(0.146)	(0.137)		(0.101)	(0.147)	(0.140)
Above Average		1.276***	1.414***	1.173***		1.293***	1.432***	1.191***
		(0.072)	(0.101)	(0.101)		(0.072)	(0.104)	(0.098)
Observations	846	846	389	457	846	846	389	457
p-value: pos FB=0		0.799	0.202	0.133		0.347	0.403	0.092

Notes: Dependent variables are how much worker agree that the feedback was accurate (col. 1-4) and appropriate (col. 5-8). All estimations are OLS. Robust standard errors are in parentheses. The p-value reported in the last row tests if the effect of positive feedback is different from zero. The table is estimated for participants who received feedback. A programming error prevented collecting data on this variable for the second half of the sample. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Results on the perceived appropriateness of feedback are similar (Figure A2). Male workers assess criticism to be 0.2 s.d. less appropriate and praise 0.1 s.d. more appropriate if it comes from a female manager. The appropriateness assessment of female workers is independent of the manager gender (Figure A2 and Table 5 column 7 and 8).

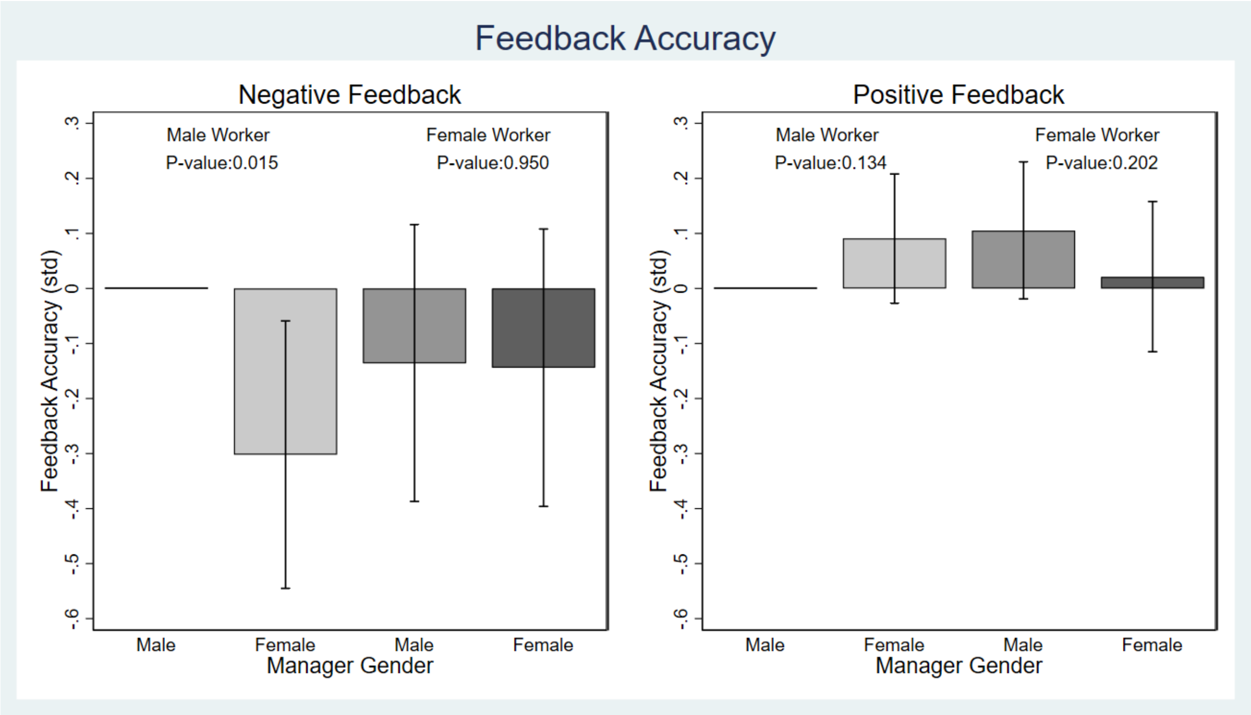
3.3.2 Attitude

Table 6 disaggregates the overall negative effect of feedback on attitudes both by worker gender and feedback content. Looking at aggregate effects of feedback, I find that negative effects tend to be stronger for female managers (col. 1, 3, 5, 7). These manager gender differences are sizable (about 50% for the index), but not statistically significant.

The interaction of feedback content and manager gender shows that the aggregate negative effect of female feedback is exclusively driven by large effects of criticism. Importantly, the

[17] One explanation for this worker gender differences is that men react more negatively if they perceive a threat to their manhood (Netchaeva et al., 2015). For a review of other examples of motivated stereotyping see Fiske and Neuberg (1990).

Figure 4: Feedback Accuracy



Notes: Figure 4 shows how workers perceive the accuracy of positive (left panel) and negative (right panel) feedback. Bars show the difference of each combination relative to the assessment of a male worker paired with a male manager. Effects are reported for a standardized measure of feedback accuracy (mean=0, s.d.=1).

negative effects of criticism by women are for most outcomes *twice* as large as for men: willingness to work for the firm in the future decreases by 10 p.p. instead of 5p.p., job satisfaction drops by 0.55 instead of 0.35 s.d., and task importance decreases by 0.32 instead of 0.16 s.d.. In aggregate, attitudes in response to female criticism drop by 0.6 s.d. compared to 0.3 for male.

While these patterns are broadly similar across female and male workers, there is suggestive evidence that workers are more sensitive to criticism from managers of the opposite sex. While this pattern holds for each outcome, this difference is only significant for task importance. By contrast, there are no manager gender differences for praise across worker gender (p-value worker gender difference for attitude index: 0.8).

These results can also be depicted by comparing average outcomes across the eight groups that are effectively created through the triple interaction. The bottom panel of Figure 5 shows that the effect of praise is very small and similar across both manager and worker gender. The top panel shows a significant drop in attitude in response to criticism across manager gender. These effects do not differ significantly by worker gender. However, as

Table 6: Effect of Feedback Content on Attitudes by Worker Gender

	Work Future		Job Satisf.		Task Import		Index (std)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Full Sample								
Feedback	-0.020*	-0.050***	-0.190***	-0.373***	-0.237**	-0.268*	-0.134***	-0.233***
	(0.011)	(0.019)	(0.061)	(0.098)	(0.092)	(0.141)	(0.042)	(0.068)
Fem Mgr x FB	-0.024	-0.049*	-0.081	-0.270*	-0.008	-0.276	-0.066	-0.228**
	(0.015)	(0.027)	(0.086)	(0.140)	(0.127)	(0.194)	(0.058)	(0.094)
Above x FB		0.060***		0.373***		0.066		0.203**
		(0.023)		(0.122)		(0.185)		(0.084)
Fem x FB x Above		0.046		0.346**		0.510**		0.301***
		(0.031)		(0.171)		(0.256)		(0.115)
Observations	2642	2642	2642	2642	2642	2642	2642	2642
Panel B: Female Workers								
Feedback	-0.021	-0.060*	-0.214**	-0.419***	-0.357***	-0.539***	-0.168***	-0.324***
	(0.018)	(0.031)	(0.092)	(0.146)	(0.130)	(0.199)	(0.064)	(0.102)
Fem Mgr x FB	-0.022	-0.040	-0.082	-0.305	0.050	-0.056	-0.048	-0.171
	(0.024)	(0.042)	(0.124)	(0.203)	(0.176)	(0.271)	(0.086)	(0.142)
Above x FB		0.078**		0.422**		0.371		0.318**
		(0.035)		(0.180)		(0.260)		(0.125)
Fem x FB x Above		0.028		0.390		0.173		0.208
		(0.048)		(0.243)		(0.353)		(0.170)
Observations	1234	1234	1234	1234	1234	1234	1234	1234
Panel C: Male Workers								
Feedback	-0.019	-0.043*	-0.172**	-0.351***	-0.127	-0.042	-0.103*	-0.162*
	(0.015)	(0.025)	(0.081)	(0.131)	(0.129)	(0.200)	(0.056)	(0.091)
Fem Mgr x FB	-0.025	-0.055	-0.077	-0.218	-0.057	-0.453	-0.081	-0.267**
	(0.020)	(0.035)	(0.121)	(0.193)	(0.183)	(0.278)	(0.078)	(0.126)
Above x FB		0.047		0.357**		-0.161		0.121
		(0.030)		(0.164)		(0.261)		(0.113)
Fem x FB x Above		0.059		0.273		0.774**		0.362**
		(0.040)		(0.240)		(0.368)		(0.156)
Observations	1408	1408	1408	1408	1408	1408	1408	1408
Sample Mean	0.95	0.95	4.94	4.94	4.15	4.15	-0.00	-0.00
Std Dev	0.21	0.21	1.14	1.14	1.65	1.65	0.77	0.77

Notes: The dependent variable in Column (1) and (2) is whether the worker says the receipt is legible. The dependent variable in Column (3) and (4) measure if the worker is willing to add up the amounts. The dependent variable in Column (5) and (6) captures the accuracy of transcribing receipts Column (7) and (8) use the score as the dependent variable and divide the sample by worker gender. All estimations are OLS. Robust standard errors are in parentheses. The p-value reported in the last row tests if the sum of the two feedback coefficients are different from zero. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

discussed, the effect of criticism is about twice as large for female (top right panel) compared to male managers (top level panel). This is in part driven by the fact that workers report slightly more positive attitudes working under a female manager in the absence of feedback.

In sum, I find that both male and female workers lower attitudes much more in response to criticism from a female manager. However, only male workers *perceive* criticism from female managers as less accurate (and appropriate), suggesting that for female workers the change in attitude operates through a different channel than feedback perception. The next section explores if the overall drop in attitude affects effort levels.

3.3.3 Effort

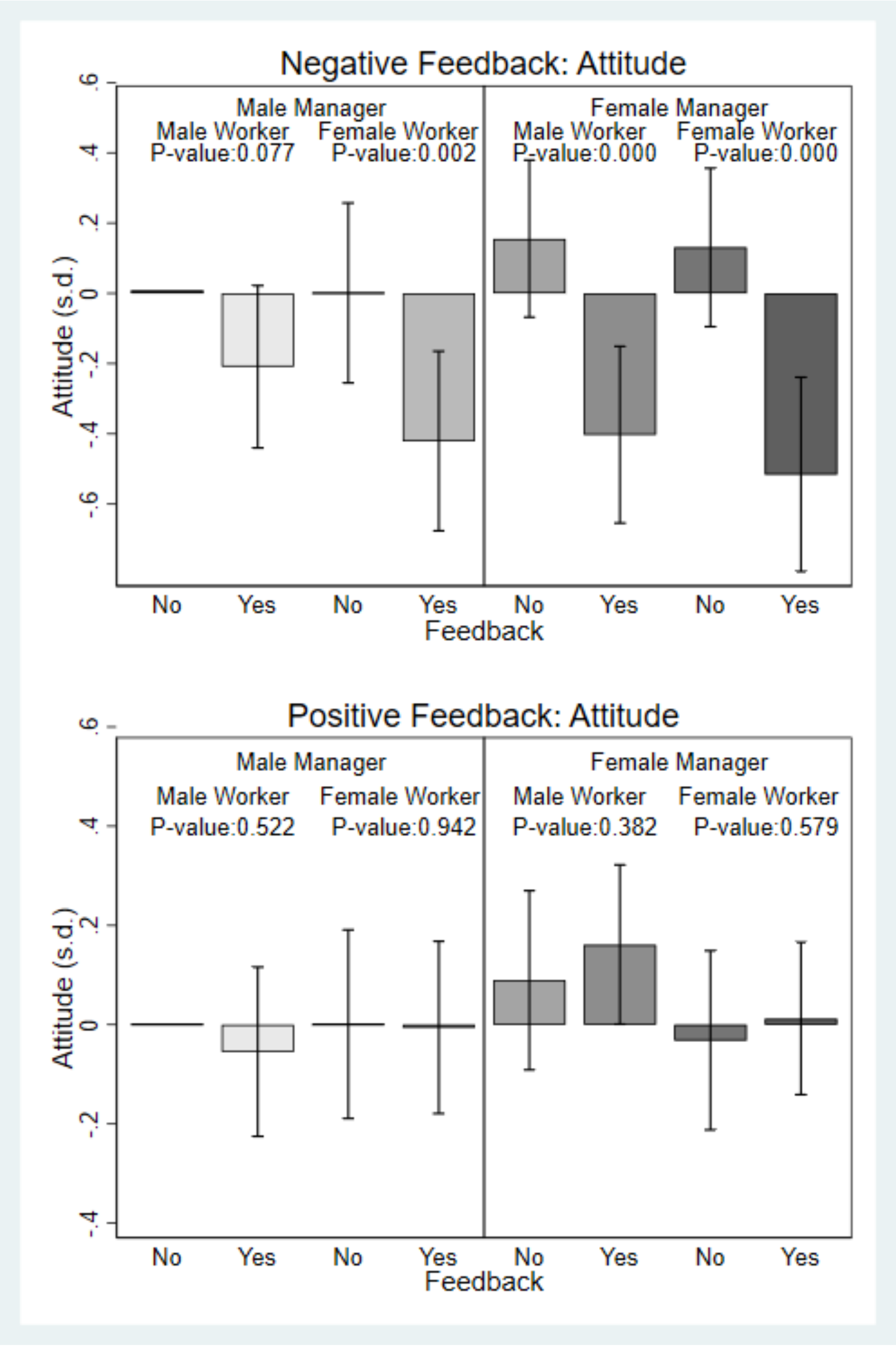
Section 3.2. concluded that female workers do not change effort in response to either type of feedback. By contrast, there is suggestive evidence that male workers reduce effort in response to feedback, driven by their negative response to criticism. This section tests how these results vary with the gender of the manager.

Table 7 shows that the (aggregate) effect of feedback does not differ by manager gender (Panel A, col. 1, 3, 5). However, there are important difference by worker gender. Men lower effort (Panel B, col. 1, 3, 5) whereas women increase effort slightly (Panel C, col. 1, 3, 5) in response to male feedback (gender difference p-values reported in row *i)*). By contrast, women and men react similarly to feedback by female supervisors (p-values row *ii)*).

Next, I test how these effects differ by feedback *content*. First, neither men nor women react to either positive or negative feedback by female managers (see top right and bottom right panel of Figure 6). Second, male workers reduce performance by 0.23 s.d. if criticized by a men (Panel C, col. 6), whereas we see a small increase among female workers (worker gender difference p-value: 0.10). Gender differences in response to male criticism are even more pronounced for the voluntary task: men lower efforts by 0.2 s.d. (Panel B, col 4), while women increase it by 0.13 s.d. (Panel C, col. 4) (worker gender difference p-value: 0.03).

In summary, results resemble those on attitude in that praise has no effect for either manager gender and either worker gender. Results differ on the effect of criticism: neither male nor female workers lower effort in response to criticism by a woman. By contrast, only men lower efforts if criticized by a men, especially on voluntary tasks.

Figure 5: Feedback and Attitude, Manager Gender



Notes: This graph shows the effect of negative (top) and positive feedback (bottom) for both women and men working under male (left) and female managers on the standardized attitude outcome. This attitude outcome is normalized so that zero presents the attitude of a male worker assigned to male managers who does not receive feedback. P-values are reported for a test of equal means across feedback assignment for each the respective worker manager gender combination. 90% confidence intervals are displayed.

Table 7: Effect of Feedback Content on Effort by Worker Gender

	Legible		Adding		Transcrip Score	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Full Sample						
Feedback	-0.011	-0.024	-0.025	-0.023	-0.248	-0.448
	(0.014)	(0.022)	(0.023)	(0.032)	(0.239)	(0.369)
Fem Mgr x FB	0.006	0.011	0.009	0.017	0.155	0.237
	(0.019)	(0.032)	(0.032)	(0.046)	(0.336)	(0.524)
Above x FB		0.034		0.005		0.569
		(0.026)		(0.044)		(0.438)
Fem x FB x Above		-0.018		-0.024		-0.336
		(0.037)		(0.063)		(0.624)
Observations	2642	2642	2642	2642	2642	2642
Panel B: Female Workers						
Feedback	0.016	0.016	0.009	0.052	0.198	0.206
	(0.019)	(0.031)	(0.034)	(0.046)	(0.343)	(0.529)
Fem Mgr x FB	-0.011	-0.022	-0.012	-0.025	-0.271	-0.494
	(0.026)	(0.043)	(0.047)	(0.067)	(0.463)	(0.725)
Above x FB		0.006		-0.081		0.076
		(0.036)		(0.066)		(0.631)
Fem x FB x Above		0.017		0.026		0.342
		(0.051)		(0.092)		(0.869)
Observations	1234	1234	1234	1234	1234	1234
Panel C: Male Workers						
Feedback	-0.034*	-0.055*	-0.053*	-0.082*	-0.632*	-0.996*
	(0.019)	(0.031)	(0.031)	(0.044)	(0.331)	(0.515)
Fem Mgr x FB	0.020	0.036	0.025	0.046	0.514	0.839
	(0.028)	(0.046)	(0.045)	(0.062)	(0.485)	(0.752)
Above x FB		0.055		0.071		0.957
		(0.036)		(0.060)		(0.607)
Fem x FB x Above		-0.043		-0.056		-0.875
		(0.054)		(0.088)		(0.892)
Observations	1408	1408	1408	1408	1408	1408
Sample Mean	0.76	0.76	0.55	0.55	11.72	11.72
Std Dev	0.28	0.28	0.42	0.42	4.30	4.30
i) P-v: M Mgr: M=F	0.062		0.171		0.081	
ii) P-v: F Mgr: M=F	0.498		0.578		0.925	
iii) P-v: M Mgr, Neg: M=F	0.103			0.034		0.102
iv) P-v: F Mgr, Neg: M=F		0.780		0.337		0.858

Notes: The dependent variable in Column (1) and (2) is whether the worker says the receipt is legible. The dependent variable in Column (3) and (4) measure if the worker is willing to add up the amounts. The dependent variable in Column (5) and (6) captures the accuracy of transcribing receipts Column (7) and (8) use the score as the dependent variable and divide the sample by worker gender. All estimations are OLS. Robust standard errors are in parentheses. The p-value reported in the last row tests if worker gender effects are different from zero. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Mechanism

The literature is inconclusive on the effect of feedback, especially for studies outside the lab. For example, Blanes i Vidal and Nossol (2011) find a positive effect of feedback on employee productivity while Barankay (2012) finds the opposite. This highlights the need to shed light on underlying mechanisms in order to better understand why effects differ across contexts. As outlined in the pre-analysis plan, I will explore the role of attention discrimination, implicit gender bias, and gendered expectations.

4.1 Attention Discrimination

Bartos et al. (2016) point out that negative stereotypes may result in people allocating less attention to certain groups. It has long been proposed that female leaders may face such attention discrimination, e.g. they are more likely to be ignored when trying to influence people (Carli, 2001). This is critical for the success of managers since attention of subordinates is important to lead and change behavior (Dansererau et al. 1975). This study provides a unique opportunity to test this hypothesis since I can measure the exact time that workers spend reading and thinking about managers' feedback.

In the presence of attention discrimination, workers would proceed faster to the next task if feedback comes from a female supervisor. I instead find the opposite: on a sample mean of 11 seconds, workers spend 0.84 seconds (0.13 s.d., p-value: 0.007) *longer* on the feedback of female supervisors (Table 8, col. 1). The difference is not driven by outliers (see distribution Figure A1). There are no significant differences by worker gender (col. 2), feedback content (col. 3) or the interaction of the two (col. 4, 5).

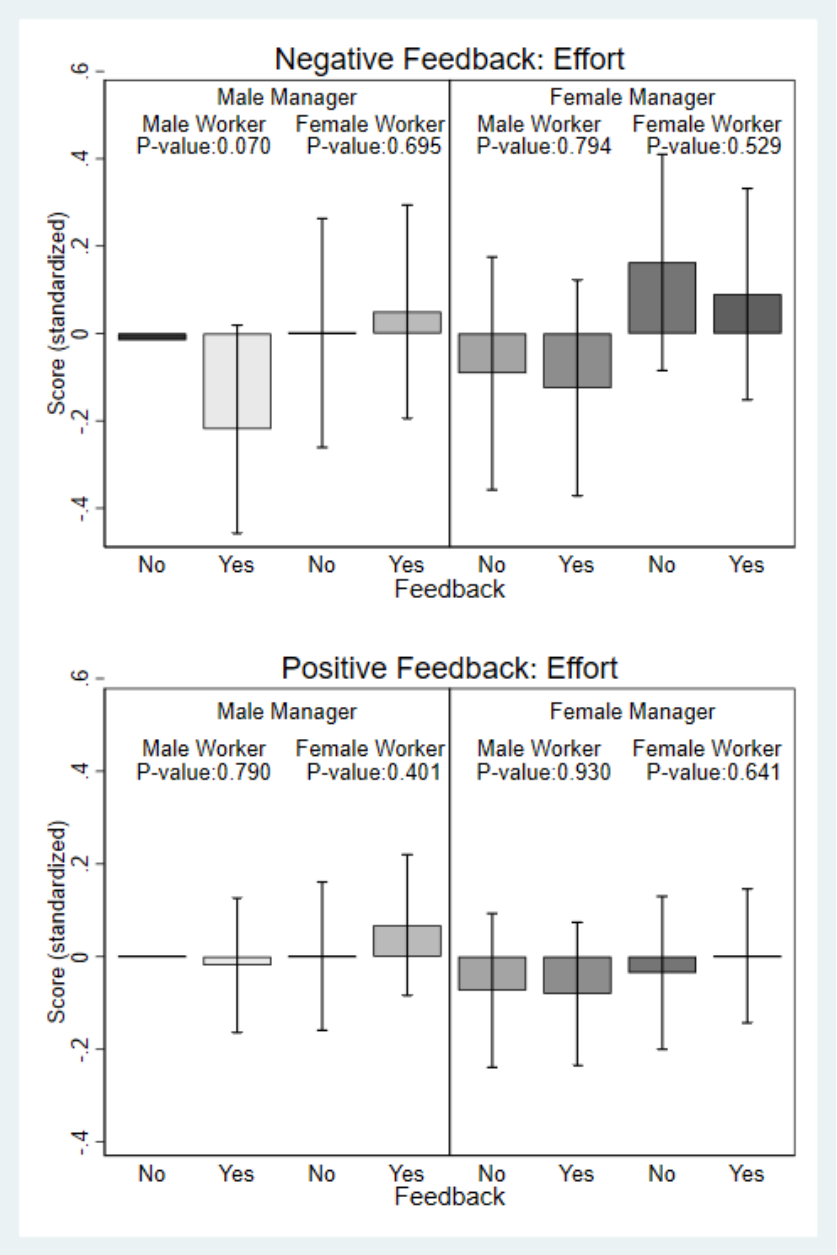
4.2 Implicit Bias

There is an active debate to what extent implicit biases are important drivers of behavior. One side claims that IATs are poor predictors of actual behavior (Oswald et al., 2013). The other side argues that implicit biases are far more common and predictive than explicit biases (Greenwald et al., 2015; Grossman et al., 2016). For example, Reuben et al. (2014) find that employers' IAT scores predict biased updating of expectations upon receiving performance information. Along the same lines, Glover et al. (2019) find that manager bias lowers job performance of minority workers.

After disclosing the study purpose to workers, I offer a small bonus for completing a "Gender-Career" IAT.[18] The IAT measures the speed with which someone associates male

[18] The IAt is based on Nosek et al. (2005). I use a propriety technology developed by Carpenter et al.

Figure 6: Attitudes and Feedback Content



Notes: This graph shows the effect of negative (top) and positive feedback (bottom) for both women and men working under male (left) and female managers on the standardized transcription score. This score is normalized so that zero presents the score of male workers assigned to male managers who do not receive feedback. P-values are reported for a test of equal means across feedback assignment for each the respective worker manager gender combination. 90% confidence intervals are displayed.

Table 8: Attention to Feedback

	(1)	Gender (2)	Content (3)	Fem Wkr (4)	Male Wkr (5)
Female manager	0.836***	0.688	1.015**	1.309*	0.613
	(0.310)	(0.425)	(0.470)	(0.699)	(0.637)
Female worker		0.360			
		(0.433)			
Fem mgr x Fem wkr		0.193			
		(0.624)			
Above average			-0.507	-0.674	-0.311
			(0.426)	(0.663)	(0.556)
Abover Av x Fem Mgr			-0.342	-0.769	0.148
			(0.620)	(0.915)	(0.851)
Observations	1575	1565	1575	719	846
Sample Mean	11.10	11.10	11.10	11.10	11.10
Std Dev	6.16	6.16	6.16	6.16	6.16

Notes: The dependent variable in Column (1) and (2) is whether the worker says the receipt is legible. The dependent variable in Column (3) and (4) measure if the worker is willing to add up the amounts. The dependent variable in Column (5) and (6) captures the accuracy of transcribing receipts Column (7) and (8) use the score as the dependent variable and divide the sample by worker gender. All estimations are OLS. Robust standard errors are in parentheses. The p-value reported in the last row tests if the sum of the two feedback coefficients are different from zero. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

names with professional terms (e.g. "office," "manager," and "salary") and female names with family terms ("marriage", "home" and "children"). The difference in speed is transformed to a standardized score ranging between 1 and -1 where 0 indicates no bias (Greenwald et al., 2003). Slight, moderate, and strong biases correspond to scores of 0.15, 0.3, and 0.6, respectively (Banaji, 2013).

The average score in our sample is 0.32 indicating moderate levels of bias.[19] However, implicit gender bias does not explain why workers discriminate against female managers.[20]

(2018) to integrate the IAT into Qualtrics surveys. Workers receive a $1 USD bonus for completing this test, which takes approximately 6 minutes. Of those that we offered it to, more than 90% chose to take the IAT. After 830 workers I stopped administering the test for budgetary reasons.

[19] Individuals' implicit bias scores are not correlated with the assigned manager gender, feedback, or their interaction (Table 2.1), confirming that implicit biases are based on cultural associations that are unlikely to move as a result of a single experience (Ma et al., 2013). There are, however, significant relationships between worker characteristics and implicit bias scores. Consistent with Nosek et al. (2002), I find that female workers exhibit a 0.2 s.d. larger implicit bias than men, significant at the 1% level even after controlling for factors such as age, race, education, treatment group, and manager gender. Age and, interestingly, education levels are also positively correlated with implicit gender bias.

[20] Models are fully interacted, but only coefficients of interest are reported for readability reasons.

Those working under female supervisors do not change attitude and effort differently depending on their implicit gender bias (Table 9, col. 1, 5). Neither does gender bias predict how workers respond to feedback from a female manager (col. 2, 6). Looking at this effect separately by feedback content, I find that, if at all, workers with larger gender biases tend to react more negatively to *praise* from female managers (col. 3, 7) but not to criticism (col. 4, 8). This is the opposite of what one would expect if implicit bias explained why the negative effect of criticism on attitude is twice as large for female managers (3).

Table 9: Implicit Bias

	Transcrip. Score				Attitude Index			
	Aggreg. (1)	Feedb. (2)	Pos FB (3)	Neg FB (4)	Aggreg. (5)	Feedb. (6)	Pos FB (7)	Neg FB (8)
Fem Mgr x IAT	-0.083 (0.360)	0.262 (0.606)	0.331 (0.643)	0.408 (0.933)	0.199 (0.155)	0.272 (0.212)	0.404 (0.280)	0.112 (0.315)
IAT x FB		-0.070 (0.573)	-0.577 (0.603)	-0.066 (0.861)		0.160 (0.224)	0.423* (0.250)	-0.231 (0.364)
Fem Mgr x FB x IAT		-0.592 (0.756)	-0.726 (0.820)	-0.120 (1.182)		-0.182 (0.301)	-0.597* (0.349)	0.370 (0.480)
Observations	822	822	442	380	822	822	442	380
Sample Mean	13.94	13.94	13.94	13.94	-0.00	-0.00	-0.00	-0.00
Std Dev	2.112	2.112	2.112	2.112	0.768	0.768	0.768	0.768

Notes: The dependent variable in Column (1) and (2) is whether the worker says the receipt is legible. The dependent variable in Column (3) and (4) measure if the worker is willing to add up the amounts. The dependent variable in Column (5) and (6) captures the accuracy of transcribing receipts Column (7) and (8) use the score as the dependent variable and divide the sample by worker gender. All estimations are OLS. Robust standard errors are in parentheses. The p-value reported in the last row tests if the sum of the two feedback coefficients are different from zero. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.3 Gendered Expectations and Perception of Feedback

Previous studies have found that the impact of feedback depends on the subjects' expectations about their own performance (Gjedrem, 2018; Kuhnen and Tymula, 2012). Workers' response to feedback may also depend on expectations they hold about the leadership style of their respective manager. Numerous studies find that people have gendered expectations: women are perceived to be more agreeable, altruistic, warm, compliant, modest, and sympathetic than men (Blau and Kahn, 2017; Bertrand, 2011). Eagly et al. (1995) conjectures that these "*gender role expectations spill over into leadership roles and produce important consequences for the effectiveness of leaders*". In our context, workers may react more negatively to criticism from female managers precisely because they associate women with giving praise.[21]

[21] Research shows that feedback is more negatively received if there is an incongruence between the content and the facial expressions / tone of the person giving the feedback (Newcombe and Ashkanasy, 2002). The

Table 10: Management Style Expectations

	N	Female	Male	Either	M/F Ratio
Panel A: Full Sample					
Strict Expectation	303	.15	.29	.55	1.93
Give Criticism	303	.13	.32	.55	2.46
Give Praise	303	.37	.11	.52	0.3
Appropriate Tone	303	.32	.12	.57	0.37
Communicate Expectations	303	.25	.18	.57	0.72
Give Feedback	303	.16	.14	.7	0.87
Panel B: Female Respondents					
Strict Expectation	120	.18	.24	.57	1.33
Give Criticism	120	.14	.3	.56	2.14
Give Praise	120	.4	.13	.48	0.32
Appropriate Tone	120	.33	.09	.58	0.27
Communicate Expectations	120	.33	.13	.54	0.39
Give Feedback	120	.2	.12	.68	0.6
Panel C: Male Respondents					
Strict Expectation	183	.13	.33	.54	2.54
Give Criticism	183	.13	.33	.55	2.54
Give Praise	183	.34	.1	.55	0.29
Appropriate Tone	183	.31	.13	.56	0.42
Communicate Expectations	183	.2	.21	.59	1.05
Give Feedback	183	.14	.15	.72	1.07

Notes: Results from an out of sample survey asking people whether they associate different management styles with female or male managers or either gender. M/F reported the response ratio of male versus female association.

I collect data on gendered expectations with regard to management styles through an out-of-sample survey with 303 MTurk workers.[22] Results confirm evidence from other settings. Respondents are about twice more likely to associate strict expectations giving criticism with male managers. By contrast, female managers were about three times more likely to be associated with giving praise and appropriate tone (Table 10, Panel A). Workers in our study may thus regard criticism from female managers as a much stronger negative signal because women needed to overcome their "natural inclination" to be nice and agreeable.

While there are no differences in gendered expectations for praise across worker gender, the association of criticism with male managers is larger among male workers, especially with regard to having strict standards (M/F ratio, Panel B and C). This may explain why

authors stress the importance to test these results outside the lab in a "*more complex design that permits investigation of gender effects in leader-member relationship*".

[22] whether they associate characteristics such as "giving praise", "strictness" and "giving criticism" more with female or male supervisor.

male workers tend to perceive female criticism to be more inaccurate and inappropriate than female workers.

In sum, gendered expectations offer a mechanism consistent with the study results. However, as with implicit bias, this mechanism test is inconclusive: control group participants provide a counterfactual, but gendered expectations may be correlated with other factors that drive discriminatory behavior. Additional research is therefore needed to establish causality more conclusively, e.g. by exogenously manipulating gender expectations or factors that increase reliance on implicit biases such as fatigue or ambiguity (Ma et al., 2013).

5 Discussion

It has been proposed that the rise of the gig economy will particularly benefit women as it offers more job with flexibility (Goldin, 2014). However, these work arrangements have also raised concerns about discrimination due to lack of regulatory oversight and equal opportunity protections. Existing evidence on discrimination in the gig economy is mixed and depends on the specific context and platform.[23] The present study provides evidence from one of the largest platforms and sheds light on mechanisms, which may help to understand why discrimination varies across contexts.

One additional rationale for conducting this study on a gig economy platform like MTurk is that it offers researchers a setting with actual employees, while maintaining experimental control one would lose in more traditional work settings. As a result, MTurk is getting more popular as a platform to do research. This raises concerns about potential Hawthorne effects. Before the study debrief, I inform participants that the purpose of the task was research and ask them what they guess the main thing I try to learn through this study. 55% said they do not know. The most common answers among those that guessed was related to the actual transcription task (e.g. testing the quality and speed of transcription). 13% (or 6% of the population) correctly answered that the research is about feedback. Nobody guessed that the research was related to gender.

Importantly, those that thought they knew the purpose behind the research behave very similarly than others. Taking, for example, the key finding of how workers change attitudes in response to female feedback, the response to both criticism and praise are very similar between these groups (p-values of equal coefficients: 0.93 and 0.83, respectively). This makes it unlikely that results are driven by surveyor demand effects or other Hawthorne effects. In addition, one would expect participants *not* to discriminate if they suspected to be part of a research project.

[23] Edelman et al. (2017) finds evidence for racial discrimination on AirBnB, for example. By contrast, Cook et al. (2018) conclude that the gender earning gap among Uber drivers is not driven by discrimination.

The external validity of these findings is, however, more debatable. With regards to observable characteristics, MTurk workers in the US tend to be Whiter, more educated, and lower income than average (Paolacci et al., 2010). However, MTurk offers a more realistic work environment than conducting lab experiments with undergraduate students and allows researchers to conceal the true purpose of their research within Internal Review Board (IRB) restrictions. Furthermore, research shows that MTurk workers share the same heuristics and biases as subjects in other alternative study populations (Horton et al., 2010).

Numerous studies uncovering examples of how female workers and managers are treated differently have received a lot of attention in recent years. However, not all of these cases are necessarily the result of gender discrimination. For example, Correll and Simard (2016) find that women receive more vague feedback than men, which hinders their ability to learn and excel. In theory, this can be an optimal management strategy, if women in fact react more negatively to criticism. However, I instead find that, if at all, female workers are less likely to reduce effort after receiving criticism. Other studies document men are more likely to be aggressive and assertive when communicating with female supervisors (Vandello and Bosson, 2013; Netchaeva et al., 2015).

My results show that this is not merely the result of different leadership styles across manager gender. Male workers are more dismissive of criticism from women, even if the feedback is *identical* - a clear case of gender discrimination.

What can be done to address discrimination against female managers? One possibility is that such behavior will abate once workers get more accustomed to women in power (Beaman et al., 2009). However, 86% of workers in my sample previously had a female supervisor compared to 96% who had male supervisor.[24] When asked to rate their effectiveness on a scale from very ineffective (0) to very effective (4), female and male managers get similar scores (3.14 and 3.00, respectively). Workers with more positive experiences with female managers are also not less likely to discriminate against female managers.[25] It is therefore unlikely that increased exposure to (effective) female managers is sufficient to address gender discrimination.

There is evidence, however, that gender discrimination may be lower among younger workers. Following the pre-analysis plan, Table A5 shows how results differ for people below versus above the median worker age of 34. Consistent across all attitude outcomes, gender discrimination in response to criticism is only significant among older workers. Differences are large in magnitude: coefficients for the attitude index are almost three times larger for older workers (col. 8) (although the worker age difference is significant at the 5% level only

[24] After completing the task, we ask workers: *"Did you have a male / female supervisor in the past?"* If they answer yes, we ask: *"How effective was this supervisor?"*

[25] Table A3 confirms that having had a female manager is not correlated with how much effort people put in (col. 1) nor with their attitudes about the task (col. 5). In aggregate I find that those with prior experience react more *negatively* to feedback from female managers (col. 2, 6). There is no clear pattern whether this is driven by positive feedback (col. 3, 7) or by negative feedback (col. 4, 8).

for job satisfaction). While this specification was not included in the pre-analysis plan, it is noteworthy that workers in their 20s respond equally to female than male criticism. People in their 20s are also equally likely to associate the specific criticism used in the study with male and female supervisors, while older workers are three times more likely to associate it with male supervisors.[26] This provides additional support for the importance of gendered expectations.

What strategies may be effective in reducing gender discrimination from below? There is evidence suggesting that increasing awareness of biases can decrease prejudice (Devine et al., 1991; Pope et al., 2018). Recent work also finds that reduction of information asymmetries disproportionately benefits female job applicants (Abel et al. (forthcoming), Botelho and Abraham (2017)) and that gender discrimination reverses once women receive positive public evaluations (Bohren et al., 2018). Indeed, Ayalew et al. (2018) finds that people are more likely to follow advice of women if they are presented as highly skilled. Yet, other research shows gender bias in the assessments of highlty qualified professionals such as doctors (Sarsons, 2017b).

The ability to provide critical feedback is a key tool for managers to change behavior of subordinates. Results in this study show that using this tool is more likely to backfire for female managers. As a result they may adopt less effective management strategies or become altogether less interested in holding leadership positions. More research is needed to find strategies for how this can be mitigated, both to promote gender equity and to increase the efficiency of firms.

[26] This result comes from an out-of-sample survey with 150 respondents. I provide respondents the exact feedback text and ask if they think it is more likely that this statement comes from a female vs male supervisor.

References

Abel, M., R. Burger, and P. Piraino (forthcoming). The value of reference letters: Experimental evidence from south africa. *American Economic Journal: Applied Economics*.

Abraham, K. G., J. C. Haltiwanger, K. Sandusky, and J. R. Spletzer (2018). Measuring the gig economy: Current knowledge and open issues. Technical report, National Bureau of Economic Research.

Adler-Bell, S. and M. Miller (2018). The datafication of employment. Report on surveillance and privacy, The Century Foundation.

Alexander, D. M. (2006). How do 360 degree performance reviews affect employee attitudes, effectiveness and performance. Seminar Research Paper Series 8, Schmidt Labor Research Center.

Arrow, K. The theory of discrimination. Technical report.

Ashforth, B. E. and R. H. Humphrey (1995). Emotion in the workplace: A reappraisal. *Human relations 48*(2), 97–125.

Ayalew, S., S. Manian, and K. Sheth (2018). Discrimination from below: Experimental evidence on female leadership in ethiopia. WPS 079, Center for Effective Global Action.

Azmat, G. and N. Iriberri (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics 94*(7-8), 435–452.

Banaji, M. R. (2013). The implicit association test at age 7: A methodological and conceptual review. *Social psychology and the unconscious: The automaticity of higher mental processes 265*.

Bandiera, O., V. Larcinese, and I. Rasul (2015). Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics 34*, 13–25.

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. Working paper, Wharton Business Economics and Public Policy Papers.

Bartos, V., M. Bauer, J. Chytilov, and F. Matjka (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review 106*(6), 1437–1475.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova (2009). Powerful women: does exposure reduce bias? *The Quarterly journal of economics 124*(4), 1497–1540.

Bertrand, M. (2011). New perspectives on gender. In *Handbook of Labor Economics*, Volume 4, pp. 1543–1590. Elsevier.

Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review 94*(4), 991–1013.

Blanes i Vidal, J. and M. Nossol (2011). Tournaments without prizes: Evidence from personnel records. *Management Science 57*(10), 1721–1736.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Bloom, N., J. Liang, J. Roberts, and Z. J. Ying (2014). Does working from home work? evidence from a chinese experiment. *The Quarterly Journal of Economics 130*(1), 165–218.

Bohren, J. A., A. Imas, and M. Rosenberg (2018). The dynamics of discrimination: Theory and evidence.

Boring, A. (2016). Gender biases in student evaluations of teaching. *Journal of Public Economics 145*, 27–41.

Botelho, T. L. and M. Abraham (2017). Pursuing quality: how search costs and uncertainty magnify gender-based double standards in a multistage evaluation process. *Administrative Science Quarterly 62*(4), 698–730.

Buser, W., J. Hayter, and E. C. Marshall (2019). Gender bias and temporal effects in standard evaluations of teaching. *AEA Papers and Proceedings 109*, 261–265.

Carli, L. L. (2001). Gender and social influence. *Journal of Social Issues 57*(4), 725–741.

Carpenter, T., R. Pogacar, C. Pullig, M. Kouril, S. J. Aguilar, J. P. LaBouff, N. Isenberg, and A. Chakroff (2018). Survey-based implicit association tests: A methodological and empirical analysis.

Catalyst (2019). Pyramid: Women in sp 500 companies. Technical report.

Cecchi-Dimeglio, P. (2017). How gender bias corrupts performance reviews, and what to do about it. *Harvard Business Review*.

Charness, G., D. Masclet, and M. C. Villeval (2014). The dark side of competition for status. *Management Science 60*, 38–55.

Collins, L., D. R. Fineman, and A. Tsuchida (2017). People analytics: Recalculating the route. *Deloitte Insights*.

Cook, C., R. Diamond, J. Hall, J. A. List, and P. Oyer (2018). The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. Technical report, National Bureau of Economic Research.

Correll, S. and C. Simard (2016). Vague feedback is holding women back. *Harvard Business Review*.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology 18*(1), 105.

DellaVigna, S. and D. Pope (2017). Predicting experimental results: Who knows what? *Journal of Political Economy 126*(6), 2410–2456.

DellaVigna, S. and D. Pope (2018). What motivates effort? evidence and expert forecasts. *Review of Economic Studies 85*(2), 1029–1069.

Devine, P. G., M. J. Monteith, J. R. Zuwerink, and A. J. Elliot (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology 60*(6), 817.

Eagly, A. H. and S. Chaiken (1993). *The psychology of attitudes.* Harcourt Brace Jovanovich College Publishers.

Eagly, A. H., S. J. Karau, and M. G. Makhijani (1995). Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin 117*(1), 125.

Edelman, B., M. Luca, and D. Svirsky (2017). Racial discrimination in the sharing economy: evidence from a field experiment. *American Economic Journal: Applied Economics 9*(2), 1–22.

Eriksson, T., A. Poulsen, and M. C. Villeval (2009). Feedback and incentives: Experimental evidence. *Labour Economics 16*(6), 679–688.

Fiske, S. T. and S. L. Neuberg (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology*, Volume 23, pp. 1–74. Elsevier.

Fryer Jr, R. G. and S. D. Levitt (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics 119*(3), 767–805.

Gjedrem, W. G. (2018). Relative performance feedback: Effective or dismaying? *Journal of Behavioral and Experimental Economics 74*, 1–16.

Glover, D., A. Pallais, and W. Pariente (2019). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *The Quarterly Journal of Economics 132*(3), 1219–1260.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review 104*(4), 1091–1119.

Grant, A. M. (2008). Does intrinsic motivation fuel the prosocial fire? motivational synergy in predicting persistence, performance, and productivity. *Journal of applied psychology 93*(1), 48.

Greenwald, A. G., M. R. Banaji, and B. A. Nosek (2015). Statistically small effects of the implicit association test can have societally large effects.

Greenwald, A. G., B. A. Nosek, and M. R. Banaji (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology 85*(2), 197.

Grossman, P. J., C. Eckel, M. Komai, and W. Zhan (2016). It pays to be a man: Rewards for leaders in a coordination game. Monash Economics Working Papers 38-16, Monash University, Department of Economics.

Hannan, R. L., G. P. McPhee, A. H. Newman, and I. D. Tafkov (2012). The effect of relative performance information on performance and effort allocation in a multi-task environment. *The Accounting Review 88*(2), 553–575.

Hoffman, M., L. B. Kahn, and D. Li (2017). Discretion in hiring. *The Quarterly Journal of Economics 133*(2), 765–800.

Holmstrom, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org. 7*, 24.

Horton, J. J., D. G. Rand, and R. J. Zeckhauser (2010). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics 14*(3), 399–425.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (forthcoming). The allocation of talent and us economic growth. *Econometrica*.

Jamieson, K. H. et al. (1995). *Beyond the double bind: Women and leadership.* Oxford University Press on Demand.

Judge, T. A., C. J. Thoresen, J. E. Bono, and G. K. Patton (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological bulletin 127*(3), 376.

Katz, L. F. and A. B. Krueger (2018). Understanding trends in alternative work arrangements in the united states. Working paper, National Bureau of Economic Research.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental analysis of neighborhood effects. *Econometrica 75*(1), 83–119.

Kuhnen, C. M. and A. Tymula (2012). Feedback, self-esteem, and performance in organizations. *Management Science 58*(1), 94–113.

List, J. A. (2017). When corporate social responsibility backfires: Theory and evidence from a natural field experiment. Working Paper 24169, National Bureau of Economic Research.

Loury, G. C. (2009). *The anatomy of racial inequality.* Harvard University Press.

Ma, D. S., J. Correll, B. Wittenbrink, Y. Bar-Anan, N. Sriram, and B. A. Nosek (2013). When fatigue turns deadly: The association between fatigue and racial bias in the decision to shoot. *Basic and applied social psychology 35*(6), 515–524.

Mueller, G. and E. Plug (2006). Estimating the effect of personality on male and female earnings. *ILR Review 60*(1), 3–22.

Muse, L. A., S. G. Harris, and H. S. Feild (2003). Has the inverted-u theory of stress and job performance had a fair test? *Human Performance 16*(4), 349–364.

Netchaeva, E., M. Kouchaki, and L. D. Sheppard (2015). A mans (precarious) place: Mens experienced threat and self-assertive reactions to female superiors. *Personality and Social Psychology Bulletin 41*(9), 1247–1259.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature 56*(3), 799–866.

Newcombe, M. J. and N. M. Ashkanasy (2002). The role of affect and affective congruence in perceptions of leaders: An experimental study. *The Leadership Quarterly 13*(5), 601–614.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics 122*(3), 1067–1101.

Nosek, B. A., M. R. Banaji, and A. G. Greenwald (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice 6*(1), 101.

Nosek, B. A., A. G. Greenwald, and M. R. Banaji (2005). Understanding and using the implicit association test: Ii. method variables and construct validity. *Personality and Social Psychology Bulletin 31*(2), 166–180.

Oswald, F. L., G. Mitchell, H. Blanton, J. Jaccard, and P. E. Tetlock (2013). Predicting ethnic and racial discrimination: A meta-analysis of iat criterion studies. *Journal of personality and social psychology 105*(2), 171.

Paolacci, G., J. Chandler, and P. G. Ipeirotis (2010). Running experiments on amazon mechanical turk. *Judgement and Decision Making 5*(5), 411–419.

Pope, D. G., J. Price, and J. Wolfers (2018). Awareness reduces racial bias. *Management Science 64*(11), 4988–4995.

Reuben, E., P. Sapienza, and L. Zingales (2014). How stereotypes impair womens careers in science. *Proceedings of the National Academy of Sciences 111*(12), 4403–4408.

Rudman, L. A. and S. E. Kilianski (2000). Implicit and explicit attitudes toward female authority. *Personality and social psychology bulletin 26*(11), 1315–1328.

Sandberg, S. (2015). Lean in-women, work and the will to lead.

Sarsons, H. (2017a). Gender differences in recognition for group work. *American Economic Review: Papers and Proceedings 107*(5), 141–145.

Sarsons, H. (2017b). Interpreting signals in the labor market: evidence from medical referrals. *Job Market Paper*.

Sinclair, L. and Z. Kunda (2000). Motivated stereotyping of women: Shes fine if she praised me but incompetent if she criticized me. *Personality and social psychology bulletin 26*(11), 1329–1342.

Straub, T., H. Gimpel, F. Teschner, and C. Weinhardt (2014). Feedback and performance in crowd work: A real effort experiment. Tel Aviv, Israel.

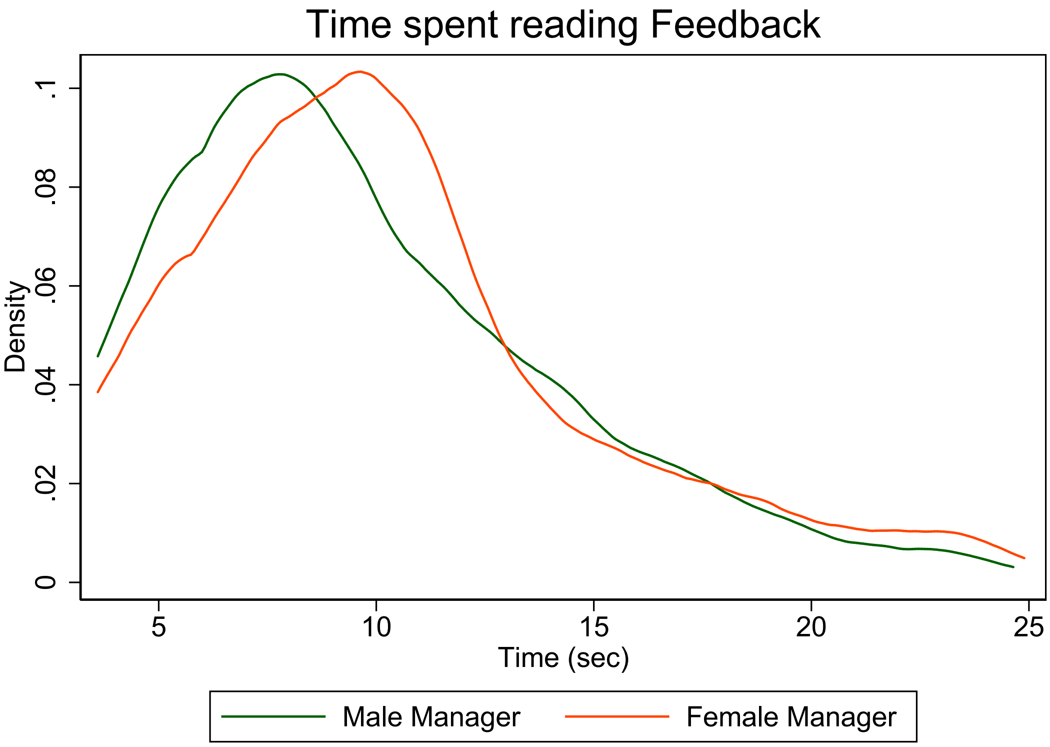
Vandello, J. A. and J. K. Bosson (2013). Hard won and easily lost: A review and synthesis of theory and research on precarious manhood. *Psychology of Men & Masculinity 14*(2), 101.

Zenger, J. and J. Folkman (2019). Research: Women score higher than men in most leadership skills. *Harvard Business Review*.

A Appendix

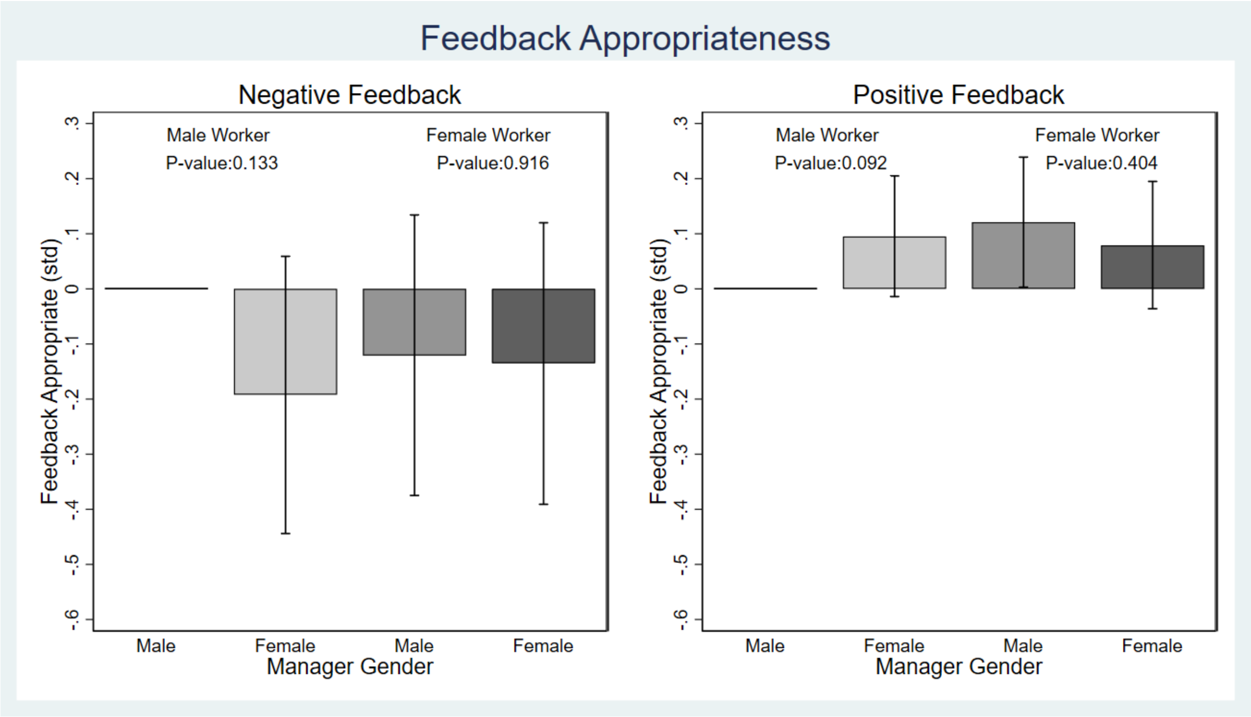
A.1 Figures

Figure A1: Attention to Feedback



Notes: The graph shows the distribution of time spent on feedback of female and male managers. Both the difference in mean and distribution are significant at the 1 percent level.

Figure A2: Feedback Appropriateness



Notes: The Figure shows how workers perceive the appropriateness of positive (left panel) and negative (right panel) feedback. Bars show the difference of each combination relative to the assessment of a male worker paired with a male manager. Effects are reported for a standardized measure of feedback accuracy (mean=0, s.d.=1).

Table A1: Name Matching

MEN	**Education**	unknown	**Age**	unknown	**White**	unknown
Ethan	2.45	0.21	31.6	0.17	0.9	0.12
Doug	2.13	0.21	45.2	0.18	0.89	0.13
Josh	2.32	0.26	31.4	0.22	0.9	0.13
Robert	2.61	0.25	43.4	0.17	0.93	0.23
Darius	2.21	0.17	36.7	0.17	0.06	0.12
Tyrone	1.98	0.06	34.5	0.06	0.04	0.1
Justin	2.57	0.15	32.9	0.1	0.82	0.17
Average	**2.32**	**0.19**	**36.5**	**0.15**	**0.65**	**0.14**
WOMEN	**Education**	unknown	**Age**	unknown	**White**	unknown
Chloe	2.45	0.25	29.9	0.17	0.86	0.196
Lynn	2.36	0.24	45.7	0.13	0.89	0.17
Ebony	2.02	0.17	33.4	0.12	0.03	0.03
Brittany	2.18	0.15	32.4	0.08	0.75	0.08
Shanice	1.9	0.2	32.7	0.2	0.06	0.08
Emily	2.71	0.1	35.3	0.1	0.86	0.1
Jennifer	2.49	0.24	37.2	0.2	0.95	0.17
Dana	2.45	0.25	41.3	0.18	0.85	0.17
Average	**2.32**	**0.20**	**36.0**	**0.15**	**0.66**	**0.12**
Diff: men-women	**-0.004**	**0.013**	**-0.535**	**-0.005**	**0.010**	**-0.018**

Notes: Name associations were collected through an out-of-sample survey with 161 participants recruited trough MTurk. Names were presented in random order. *Education* is coded as 1=less than highschool, 2=highschool, 3=college. The large row presents the difference in average values between male and female names. Unknown measures the share of people who report not having any association with a given name and characteristic.

A.2 Tables

Table A2: Correlation between Outcomes and Worker Characteristics

	Effort			Attitude			
	Legible (1)	Adding (2)	Score (3)	Work (4)	Satisf (5)	Import (6)	Index (7)
Female worker	0.006	0.004	0.167	-0.004	0.197***	0.351***	0.123***
	(0.009)	(0.017)	(0.174)	(0.008)	(0.046)	(0.067)	(0.031)
Age	0.001	0.003***	-0.003	-0.001*	-0.003	0.003	-0.001
	(0.000)	(0.001)	(0.008)	(0.000)	(0.002)	(0.003)	(0.002)
1=college degree	-0.008	-0.054***	-0.221	0.003	-0.100**	-0.130*	-0.053*
	(0.009)	(0.017)	(0.174)	(0.008)	(0.046)	(0.068)	(0.031)
1=worker black	-0.050***	0.003	-1.322***	-0.000	0.103*	0.535***	0.081**
	(0.012)	(0.021)	(0.243)	(0.010)	(0.058)	(0.080)	(0.039)
1=worker Asian	0.002	-0.047	0.374	0.003	-0.112	0.264*	0.017
	(0.019)	(0.040)	(0.355)	(0.016)	(0.106)	(0.140)	(0.065)
Observations	2461	2461	2461	2461	2461	2461	2461
Sample Mean	0.81	0.60	14.97	0.95	4.94	4.15	-0.00
Std Dev	0.25	0.42	4.61	0.21	1.14	1.65	0.77

Notes: The dependent variable in Column (1) and (2) is whether the worker says the receipt is legible. The dependent variable in Column (3) and (4) measure if the worker is willing to add up the amounts. The dependent variable in Column (5) and (6) captures the accuracy of transcribing receipts All estimations are OLS. Robust standard errors are in parentheses. The mean of the dependent variable for the control group is reported in the last row. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Previous Experience with Female Manager

	Transcrip. Score				Attitude Index			
	Aggreg.	Feedb.	Pos FB	Neg FB	Aggreg.	Feedb.	Pos FB	Neg FB
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female manager	0.300	-0.635	-0.612	-0.910*	0.042	-0.108	-0.291**	0.079
	(0.303)	(0.449)	(0.658)	(0.470)	(0.096)	(0.144)	(0.141)	(0.225)
Prior Fem	0.235	-0.228	0.102	-0.557	-0.064	-0.189*	-0.527***	0.127
	(0.234)	(0.339)	(0.445)	(0.363)	(0.077)	(0.113)	(0.112)	(0.176)
Fem Mgr x Prior Fem	-0.161	0.710	0.826	0.892*	0.006	0.233	0.542***	-0.072
	(0.325)	(0.482)	(0.711)	(0.509)	(0.106)	(0.156)	(0.170)	(0.238)
Fem Mgr x Prior F x FB		-1.401**	-0.687	-1.359**		-0.355*	-0.525*	-0.066
		(0.643)	(0.970)	(0.658)		(0.210)	(0.296)	(0.290)
Observations	1315	1315	633	682	1315	1315	633	682
Sample Mean	13.94	13.94	13.94	13.94	-0.00	-0.00	-0.00	-0.00
Std Dev	2.112	2.112	2.112	2.112	0.768	0.768	0.768	0.768

Notes: The dependent variable in Column (1) and (2) is whether the worker says the receipt is legible. The dependent variable in Column (3) and (4) measure if the worker is willing to add up the amounts. The dependent variable in Column (5) and (6) captures the accuracy of transcribing receipts Column (7) and (8) use the score as the dependent variable and divide the sample by worker gender. All estimations are OLS. Robust standard errors are in parentheses. The p-value reported in the last row tests if the sum of the two feedback coefficients are different from zero. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4: RDD Estimates: Effect of Positive vs. Negative Feedback

	Legible	Adding	Score	Work	Satisf.	Import	Index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
OLS Estimate	0.010	0.019	0.209	0.083***	0.545***	0.323**	0.353***
	(0.008)	(0.016)	(0.168)	(0.015)	(0.085)	(0.128)	(0.057)
RD Estimate	0.021	0.127	0.131	0.091**	0.388**	0.397	0.240*
	(0.031)	(0.078)	(0.526)	(0.036)	(0.165)	(0.290)	(0.130)
Degree Polyn	1	1	1	1	1	1	1
Bandwidth	1.58	0.81	1.54	1.41	1.38	1.10	1.07
N Left Cutoff	396	181	393	374	336	288	240
N Right Cutoff	525	237	522	516	516	416	338

Notes: ADD. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Effect of Feedback Content on Attitudes by Worker Age (median)

	Work Future		Job Satisf.		Task Import		Index (std)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Young workers								
Feedback	-0.019	-0.026	-0.195**	-0.486***	-0.321**	-0.443**	-0.142**	-0.250**
	(0.013)	(0.025)	(0.089)	(0.149)	(0.132)	(0.210)	(0.059)	(0.103)
Fem Mgr x FB	-0.022	-0.044	-0.019	-0.043	0.116	-0.170	-0.032	-0.159
	(0.017)	(0.033)	(0.122)	(0.205)	(0.185)	(0.290)	(0.079)	(0.136)
Above x FB		0.013		0.525***		0.221		0.194
		(0.028)		(0.181)		(0.269)		(0.122)
Fem x FB x Above		0.042		0.066		0.536		0.243
		(0.036)		(0.245)		(0.376)		(0.161)
Observations	1262	1262	1262	1262	1262	1262	1262	1262
Panel B: Old Workers								
Feedback	-0.005	-0.032	-0.176**	-0.276*	-0.108	-0.092	-0.085	-0.152
	(0.017)	(0.027)	(0.086)	(0.142)	(0.137)	(0.214)	(0.059)	(0.094)
Fem Mgr x FB	-0.041*	-0.096**	-0.201	-0.627***	-0.209	-0.492*	-0.163*	-0.438***
	(0.024)	(0.040)	(0.125)	(0.201)	(0.186)	(0.286)	(0.084)	(0.134)
Above x FB		0.056*		0.209		-0.030		0.137
		(0.034)		(0.173)		(0.274)		(0.119)
Fem x FB x Above		0.102**		0.798***		0.546		0.517***
		(0.048)		(0.247)		(0.372)		(0.169)
Observations	1226	1226	1226	1226	1226	1226	1226	1226
Sample Mean	0.95	0.95	4.94	4.94	4.15	4.15	-0.00	-0.00
Std Dev	0.21	0.21	1.14	1.14	1.65	1.65	0.77	0.77

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

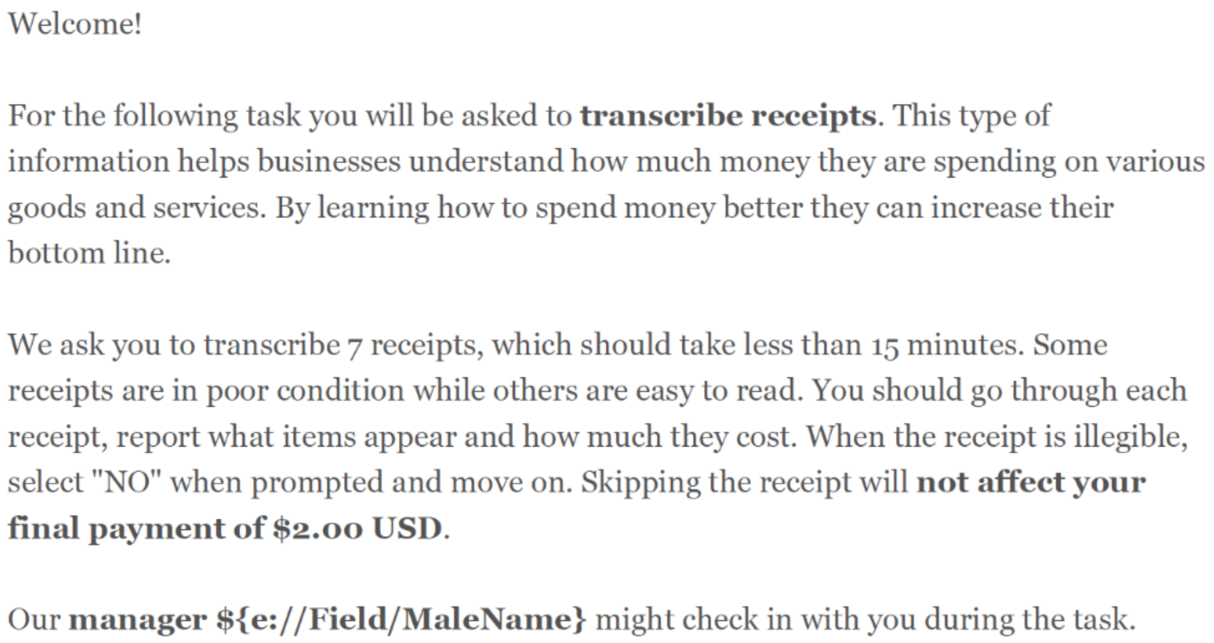
Table A6: Effect of Feedback Content on Attitudes by Worker College Completion (4 yr)

	Work Future		Job Satisf.		Task Import		Index (std)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Completed College								
Feedback	-0.008	-0.018	-0.201**	-0.271*	-0.319**	-0.253	-0.132**	-0.153*
	(0.015)	(0.025)	(0.092)	(0.143)	(0.135)	(0.206)	(0.060)	(0.093)
Fem Mgr x FB	-0.040*	-0.095***	-0.061	-0.305	0.112	-0.198	-0.074	-0.299**
	(0.021)	(0.035)	(0.128)	(0.202)	(0.189)	(0.288)	(0.085)	(0.132)
Above x FB		0.019		0.147		-0.128		0.044
		(0.031)		(0.183)		(0.271)		(0.121)
Fem x FB x Above		0.107**		0.469*		0.602		0.434***
		(0.042)		(0.253)		(0.379)		(0.168)
Observations	1197	1197	1197	1197	1197	1197	1197	1197
Panel B: No College								
Feedback	-0.017	-0.036	-0.160*	-0.465***	-0.089	-0.240	-0.090	-0.227**
	(0.014)	(0.026)	(0.083)	(0.146)	(0.132)	(0.217)	(0.057)	(0.103)
Fem Mgr x FB	-0.017	-0.045	-0.155	-0.378*	-0.194	-0.482*	-0.109	-0.304**
	(0.020)	(0.038)	(0.119)	(0.203)	(0.182)	(0.290)	(0.078)	(0.139)
Above x FB		0.035		0.554***		0.276		0.249**
		(0.029)		(0.171)		(0.270)		(0.118)
Fem x FB x Above		0.050		0.411*		0.534		0.357**
		(0.042)		(0.240)		(0.370)		(0.161)
Observations	1266	1266	1266	1266	1266	1266	1266	1266
Sample Mean	0.95	0.95	4.94	4.94	4.15	4.15	-0.00	-0.00
Std Dev	0.21	0.21	1.14	1.14	1.65	1.65	0.77	0.77

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.3 Study Design

Figure A3: Introduction



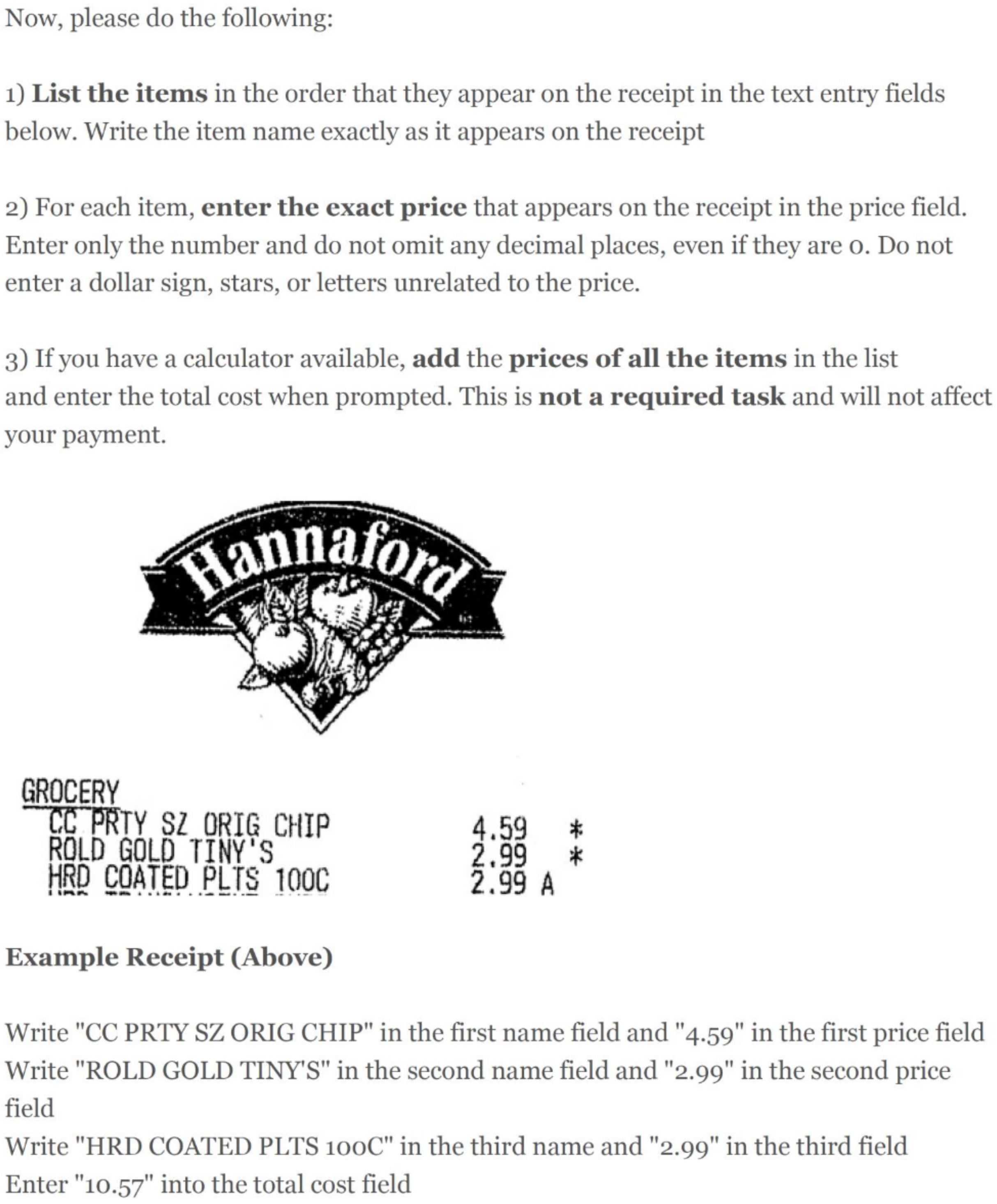
Welcome!

For the following task you will be asked to **transcribe receipts**. This type of information helps businesses understand how much money they are spending on various goods and services. By learning how to spend money better they can increase their bottom line.

We ask you to transcribe 7 receipts, which should take less than 15 minutes. Some receipts are in poor condition while others are easy to read. You should go through each receipt, report what items appear and how much they cost. When the receipt is illegible, select "NO" when prompted and move on. Skipping the receipt will **not affect your final payment of $2.00 USD**.

Our **manager ${e://Field/MaleName}** might check in with you during the task.

Figure A4: Introduction: Task Example



Now, please do the following:

1) **List the items** in the order that they appear on the receipt in the text entry fields below. Write the item name exactly as it appears on the receipt

2) For each item, **enter the exact price** that appears on the receipt in the price field. Enter only the number and do not omit any decimal places, even if they are 0. Do not enter a dollar sign, stars, or letters unrelated to the price.

3) If you have a calculator available, **add** the **prices of all the items** in the list and enter the total cost when prompted. This is **not a required task** and will not affect your payment.

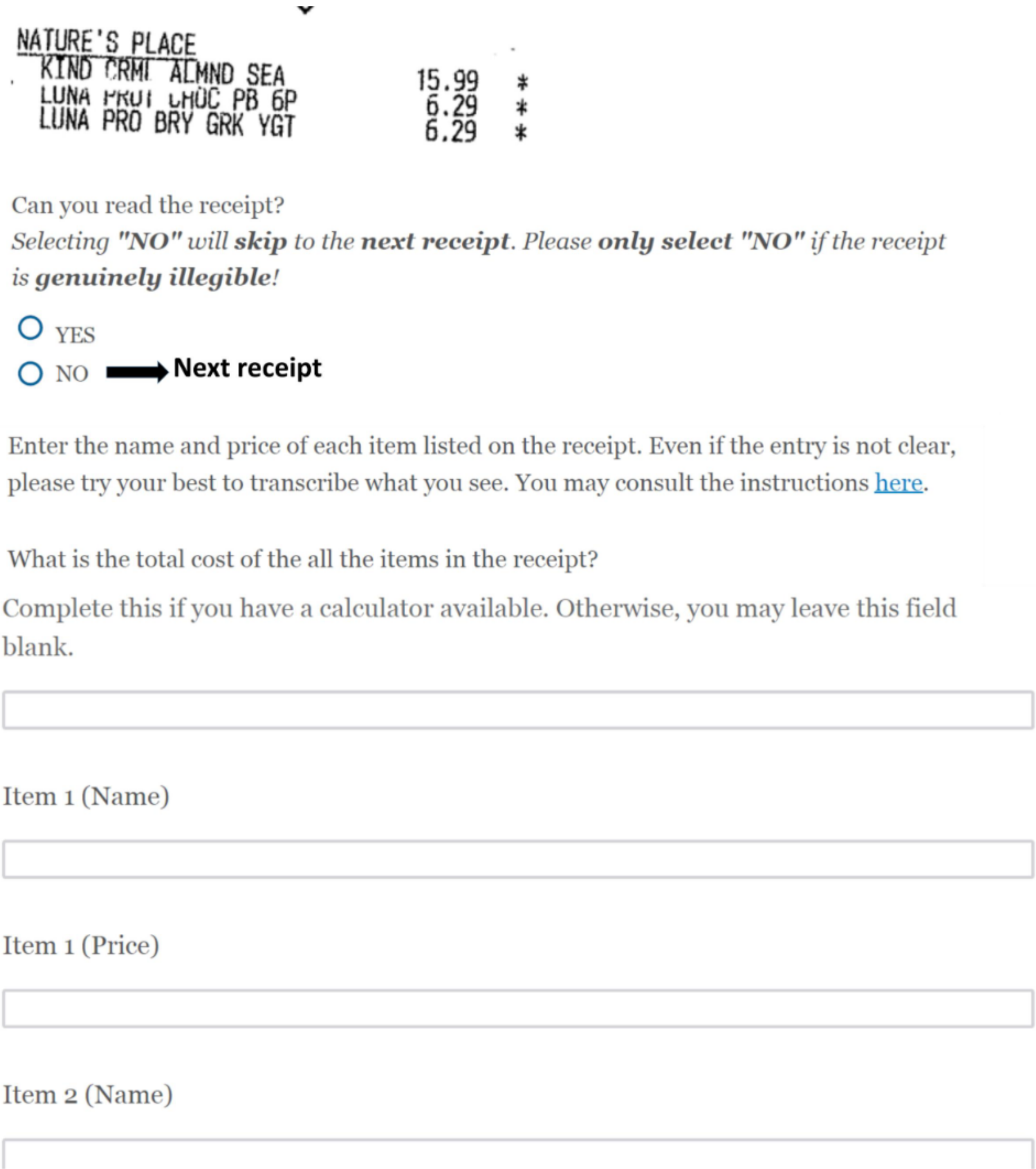
GROCERY

CC PRTY SZ ORIG CHIP	4.59	*
ROLD GOLD TINY'S	2.99	*
HRD COATED PLTS 100C	2.99	A

Example Receipt (Above)

Write "CC PRTY SZ ORIG CHIP" in the first name field and "4.59" in the first price field
Write "ROLD GOLD TINY'S" in the second name field and "2.99" in the second price field
Write "HRD COATED PLTS 100C" in the third name and "2.99" in the third field
Enter "10.57" into the total cost field

Figure A5: Task



NATURE'S PLACE

KIND CRMI ALMND SEA	15.99	*
LUNA PROT CHOC PB 6P	6.29	*
LUNA PRO BRY GRK YGT	6.29	*

Can you read the receipt?

*Selecting **"NO"** will **skip** to the **next receipt**. Please **only select "NO"** if the receipt is **genuinely illegible**!*

○ YES

○ NO → **Next receipt**

Enter the name and price of each item listed on the receipt. Even if the entry is not clear, please try your best to transcribe what you see. You may consult the instructions here.

What is the total cost of the all the items in the receipt?

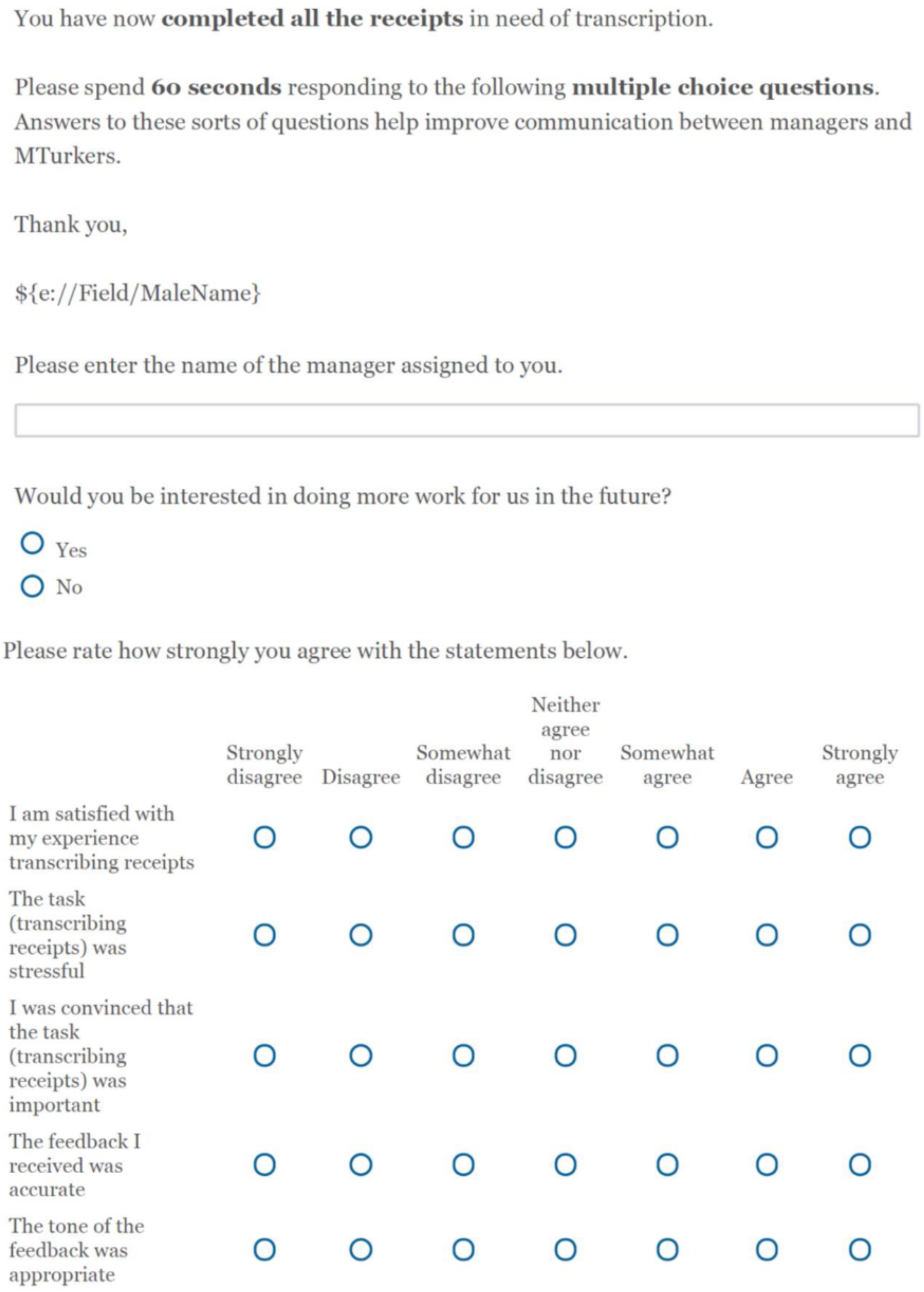
Complete this if you have a calculator available. Otherwise, you may leave this field blank.

Item 1 (Name)

Item 1 (Price)

Item 2 (Name)

Figure A6: Survey



You have now **completed all the receipts** in need of transcription.

Please spend **60 seconds** responding to the following **multiple choice questions**. Answers to these sorts of questions help improve communication between managers and MTurkers.

Thank you,

${e://Field/MaleName}

Please enter the name of the manager assigned to you.

Would you be interested in doing more work for us in the future?

○ Yes

○ No

Please rate how strongly you agree with the statements below.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I am satisfied with my experience transcribing receipts	○	○	○	○	○	○	○
The task (transcribing receipts) was stressful	○	○	○	○	○	○	○
I was convinced that the task (transcribing receipts) was important	○	○	○	○	○	○	○
The feedback I received was accurate	○	○	○	○	○	○	○
The tone of the feedback was appropriate	○	○	○	○	○	○	○