

The (in)visible hand: do workers discriminate against employers? *

PHILIPP DOERRENBERG

DENVIL DUNCAN

DANYANG LI

April 13, 2021

Abstract

Although a large literature has studied discrimination in the labor market, there is no evidence on sex- and race-based discrimination of employees against (potential) employers. We implement a randomized experiment in an online labor market to contribute to this gap in the literature. In our experiment, workers make labor-supply decisions after we randomly expose them to signals about the race and sex of the employer. We find evidence of discrimination on the quantity and quality of work, but not on the general willingness to work in our labor task. Our results also suggest that discriminating employees try to conceal their behavior ex-post. An additional survey with randomized components suggests that our results are not driven by statistical discrimination.

JEL Classification: J7, J22, C93

Keywords: labor market; employee-to-employer discrimination; gender discrimination; racial discrimination; online labor market

***Doerrenberg:** University of Mannheim Business School, CESifo, ZEW and IZA. Email: doerrenberg@uni-mannheim.de. **Duncan** (Corresponding Author): O'Neill School of Public and Environmental Affairs, Indiana University, IZA and ZEW. Email: duncande@indiana.edu. Postal Address: SPEA 375F, 1315 East 10th Street, Bloomington, Indiana 47403, USA. **Li:** Department of Economics, Hofstra University. Email: Danyang.Li@hofstra.edu. Felipe Rojas provided excellent research assistance. We thank Christoph Feldhaus, Brad Heim, David Jaeger, Andreas Peichl, Justin Ross, Jan Schmitz, Johannes Voget as well as various seminar participants for helpful comments and suggestions. IRB approvals were obtained from Indiana University and Hofstra University.

1 Introduction

The existing literature on discrimination in labor markets tends to focus on discrimination among employers against employees and among employees against other employees. Studies on employer-to-employee discrimination cover cases where employers make job-related decisions (e.g., hiring decisions, promotion, and salary increases) based on workers' characteristics such as sex and race rather than workers' productivity. Numerous audit and experimental studies have confirmed discrimination with respect to race and sex in this branch of the literature (see the surveys by [Bertrand and Duflo 2017](#) and [Neumark 2018](#)). Studies on employee-to-employee discrimination focus on cases where employees discriminate against each other. This could arise among employees of similar rank (e.g., [Hedegaard and Tyran 2018](#)) or employees of different ranks (e.g., [Glover et al. 2017](#); [Abel 2019](#)). While both branches have highlighted the disparate treatment of minority groups in the labor market, another dimension of discrimination in labor markets remains largely understudied: discrimination by employees against (potential) employers. We contribute to this strand of the literature by studying the following question: do workers discriminate against potential employers on the basis of the employer's race or sex?

Conceptually, employee-to-employer discrimination is different than employer-to-employee discrimination, regardless of the motivation for discrimination. For example, consider taste-based discrimination: paying money to someone I do not like possibly elicits a different psychological response than receiving money from someone I do not like. A similar difference might arise when we think about exerting effort for someone rather than having someone exert effort for us. Statistical discrimination may also be different across these dimensions of discrimination: while employees (especially those in a gig-market environment) are primarily interested in being paid by their employer, employers seeking an employee (gig worker) attach importance to a wider set of attributes. For example, an employer is interested in an employee's effort, diligence, productivity, reliability and punctuality.

Employee-to-employer discrimination is also conceptually different from employee-to-employee discrimination (even against employees of higher ranks). First, group-identity possibly affects how workers view their co-workers versus their employer. Second, workers generally have greater interaction with their co-workers (incl. those of higher rank) than with the employer. Third, with some exceptions, shirking by the employee will harm the employer but not the co-workers of higher rank. Fourth, in many smaller operations like the one we study, hiring/firing, wage setting, evaluation, and promotion decisions are handled by the employer and not co-workers of higher rank. As such, it is not clear that a worker's discriminatory response toward co-workers will mimic her response toward her employer.

Employee-to-employer discrimination is potentially relevant in different types of work settings. Its relevance has particularly increased significantly given the dramatic increase in gig-economy markets where workers are able to choose among many employers. Employee-to-employer discrimination is also relevant in traditional labor markets during times of low labor-market thickness and in contexts where individuals with specialized skills are able to select their employer. Finally, the inability of employers to perfectly monitor employees' labor effort

implies that employee-to-employer discrimination can manifest in all kinds of labor markets in the form of shirking.

We address our research question using data generated in a randomized experiment with 2174 subjects in an online labor market (Amazon Mechanical Turk, mTurk).¹ We choose this online labor market for at least two reasons. First, it is pro-typical for a set-up where employees have the opportunity to discriminate against employers. Second, it is characterized by a high degree of anonymity, which allows us to randomly inject one non-anonymous component (to signal race and sex; see below) to an anonymous setting and to isolate the discrimination effect in the absence of any confounding factors.

The subjects in our experiment are invited to complete a survey about a topic unrelated to our discrimination study. Upon completion of this survey, subjects are randomly assigned to one of five groups and offered an opportunity to complete an additional bonus task where they are paid a piece rate to transcribe information from gasoline receipts. The announcement of the additional bonus task features a photograph of a hand holding a gasoline receipt. We signal the employer’s race and sex by randomly varying the presence and characteristics (sex and race) of the hand in this photograph across treatment groups (as in [Doleac and Stein 2013](#)).² After being exposed to the treatment photograph, subjects are asked if they wish to complete the bonus task. Subjects who respond *yes* are allowed to transcribe up to 40 gasoline receipts. Signaling sex and race through the hand in a photograph (and not through, say, pictures of a face) is an important design feature that allows us to study discrimination in the absence of confounders such as demographic information, interactions, sympathy or trustworthiness.

The experimental design also allows us to observe subjects who are treated, but decide not to complete the bonus task. As a result, we are able to study discrimination against employers on two margins: an *extensive margin*, the decision to work for the employer in the bonus task or not,³ and an *intensive margin*, the amount and quality of work conditional on accepting the task.

We find the following average effects in our data. First, on average across all workers, we find no evidence of discrimination against employers for neither sex nor race on the extensive-margin decision to work in the bonus task. Second, we find evidence of discrimination on the intensive margin across the full sample of workers who decide to work in the bonus task. In particular, we find that workers are less accurate and transcribe fewer receipts for black employers, relative to white employers (though the latter is not significant). Interestingly, workers are significantly more accurate and tend to transcribe more receipts for female employers, relative to male employers.

¹mTurk is an established online labor-market platform and a big player in the gig-economy.

²The photograph used in the control group does not feature a hand while the photograph used in the treatment groups features either a black or white hand (to signal race) with or without nail polish (to signal sex). The title of our paper is a play on the title of [Doleac and Stein \(2013\)](#) who also used pictures of hands to signal race.

³We use the term ‘extensive margin’ when we examine the decision to work in our bonus task or not after the exposure to the randomized treatments. It is of course possible that subjects who decline our bonus task work on another Human Intelligence Task (HIT) on mTurk. In this regard, our usage of the term ‘extensive margin’ should not be understood as describing the decision to work at all or not.

We split the sample by white and non-white workers to study within-group bias in the effect of the black-hand treatment. While we do not see any heterogeneity on the extensive margin, we do find that white workers discriminate against black employers on the intensive margin while non-white workers do not discriminate against black employers. In fact, the average discrimination against black employers that we reported above for the intensive margin is almost entirely driven by white workers. We also study the discrimination behavior separately for female and male employees. This exercise shows that males transcribe more receipts and do so more accurately when exposed to a female hand (both relative to male hand) and females work more accurately for other females.

The average extensive-margin result described above could either be a true zero effect or the result of a non-salient treatment. Results from our pilot study demonstrate that the treatment is salient, which suggests that the effect is most likely a true zero.⁴ To further test this conclusion, we distinguish between the two explanations by relying on responses to a post-experimental survey that asks subjects about the sex and race of the person holding the receipt in the treatment photograph. The responses to these questions are used to define variables that indicate the self-reported salience of the treatment. For example, a subject is classified *sex-salient* if her self-reported sex-treatment matches her actual sex-treatment. We find that self-reported treatment salience is much higher for the male and white treatments compared to the female and black treatments. This pattern of salience raises a third possible explanation for our average null effect; subjects who discriminated against a minority employer indicate in the self-reported post-experimental survey that they did not discern the sex or race of the employer to hide their discriminating behavior. In other words, the null results could also reflect offsetting heterogeneous responses across treatment salience.⁵

We explore this potential source of heterogeneity by estimating the treatment effects separately for salient and non-salient subjects. The extensive margin results confirm that sex-salient workers are significantly more likely to work for female employers, and sex-nonsalient workers are significantly less likely to work for females (both relative to male employers). Similarly, we find that race-salient workers are significantly more likely to work for black employers and race-nonsalient workers are less likely (though not significant) to work for a black employer (both relative to white employer).⁶ These opposing effects cancel each other, thus producing an average null effect on the extensive margin. We would have expected to see zero effects among nonsalient subjects because these subjects seemingly did not discern the sex/race of their potential employer and therefore should not exhibit any discriminatory behavior. That we find evidence of discrimination among nonsalient workers suggests that these workers misreported their responses to questions about race and sex of the employer in the post-experimental survey,

⁴We conducted a pilot study to test the salience of the photos used in the experiment. Subjects had no trouble identifying the race or sex of the hands used in the experiment. See Section 2.1 for details on the pilot study.

⁵For example, the treatment effect can be zero if treatment-salient subjects are more likely to work for employer of type X while treatment non-salient subjects are less likely to work for group X.

⁶In other words, we find that workers who reported to have discerned the sex/race of the employer were significantly more likely to work for female/black employers and those workers who reported to have not discerned the sex/race were less likely to work for a female/black employer (both relative to white employer).

either deliberately or unconsciously. This finding is consistent with the literature on dishonesty which generally finds that dishonest people are interested in appearing honest (Mazar et al. 2008).

Overall, our results suggest that workers consider potential employers' race and sex when making labor supply decisions. But, do workers discriminate against employers because of taste (Becker 1957) or statistics (Arrow 1972; Phelps 1972)? The primary channel through which statistical discrimination could arise in our setting is through workers' beliefs about the likelihood that employers of a certain type would honor the labor contract. Results from a survey of mTurkers who did not participate in our real-effort experiment (as described in the paragraph below) reveal that non-payment is quite frequent on mTurk; more than 50% of mTurkers have experienced an employer's refusal to pay for work already completed. Statistical discrimination potentially arises if workers believe that such non-payment is more likely for employers of different types. For example, a worker might be less likely to work for black or female employers if they believe that these employers are less likely to approve and pay the bonus.

We check for the source of discrimination by surveying a sample of mTurkers (N=955) who did not participate in our main study; referred to as 'follow-up survey' throughout. The primary goal of the survey is to collect information about workers' interactions with employers including workers' perceptions about employers of different types and non-payments. The survey includes several questions and also features randomized components in which we expose survey respondents to the same treatment pictures as in our main experiment. Results from this survey suggest that the sex and race gaps we estimate are *not* driven by statistical discrimination. For example, we find that mTurkers believe male and female employers are equally likely to pay a bonus that was specified in a Human Intelligence Task (HIT). The evidence against statistical discrimination is even stronger in the case of the race gap. On the one hand, results from our real-effort labor task study show that race-salient workers were more likely to work for a black employer. On the other hand, race-salient mTurkers in our follow-up survey generally believe that black employers are less likely to pay a bonus. These two findings are inconsistent with statistical discrimination since they suggest that workers are more likely to work for employers who are less likely to pay for the task.

1.1 Contribution to the Literature

Our paper contributes to the literature on labor-market discrimination by exploring the extent to which employees discriminate against employers. There have been numerous empirical studies of discrimination in labor markets (see Riach and Rich 2002; Bertrand and Duflo 2017; Neumark 2018 for reviews). While some studies focus on employees discriminating against their bosses or other employees (Glover et al. 2017; Hedegaard and Tyran 2018; Abel 2019; Benson et al. 2019), the existing literature mostly studies discrimination by employers toward (potential) workers (see the reviews and recent examples such as Agan and Starr 2017; Chan and Wang

2018; Phillips 2020; Jaeger et al. 2020; Acquisti and Fong 2020; Coffman et al. 2021).⁷ Such strong emphasis on discrimination by employers against employees is not surprising considering that traditional labor markets are often characterized by contexts that provide employers the flexibility to discriminate against employees. In particular, workers are generally competing with many other workers for a limited number of job openings. Additionally, in many cases the firm is a neutral entity with respect to race and sex because the owner of the firm is not a single person, but rather many shareholders of varying types including other firms. As a result, the worker might neither be willing nor able to discriminate against the employer.⁸

However, even traditional labor markets include many arrangements where workers are able to discriminate against employers. For example, consider small family-owned firms (where the identity of the owner is known) or settings where workers are more accurately described as short-term independent contractors. Workers in these settings work directly for an employer and are therefore able to consider the employer’s characteristics when deciding whether to take on a given task and the effort to exert conditional on accepting the task. These types of employment arrangements have increased significantly in the last decade with the rise of the ‘gig’ economy (e.g., see Farrell and Greig 2016, Farrell et al. 2018, and Katz and Krueger 2019). As we elaborate above, discrimination against employers is conceptually different than discrimination against employees (including employees of higher rank). Therefore, it is important to understand the extent of discrimination in these settings. A related working paper is Asad et al. (2020) who study altruism and reciprocity as motivations for *white* employees discriminating against their employers.

We further add to an extensive literature that examines discrimination in platform and online markets. On the one hand, researchers suggest that the growth of the gig economy favors minorities, compared to the traditional labor market. The gig economy generates worker flexibility that can possibly narrow the gender wage gap (Cook et al. 2018). In fact, Hyperwallet (2017) reports that “86% of female gig workers believe gig work offers the opportunity to make equal pay to their male counterparts”. Brown (2018) found that ridehail services, such as Uber and Lyft, discriminate less than taxi services and extend reliable car access to neighborhoods underserved by taxis. On the other hand, many studies argue that discrimination is still rife in the gig economy, including markets such as housing rental on AirBnB (Edelman et al. 2017), ride-share (Ge et al. 2016), and consumer markets (Pope and Sydnor 2011; Nunley et al. 2011; Doleac and Stein 2013; Zussman 2013; Ayres et al. 2015).

Our experimental context is an online labor market where workers complete micro tasks for pay on a contractual basis. Unlike many of the other market platforms that have been studied so far, race and sex are not particularly salient in the market that we use for our experiment. With the exception of name, participants in the mTurk labor market know very

⁷Our finding that discriminating subjects try to conceal their behavior ex-post has, to the best of our knowledge, not been documented in the literature on discrimination in labor markets, and therefore constitutes a contribution in itself to the large literature.

⁸Of course, a worker might still express a preference for working for a supervisor of the particular race and sex (Glover et al. 2017). However, this dimension of discrimination is conceptually different than discrimination against employers; see further above in the Introduction.

little about each other. Additionally, there is no face-to-face interaction, and communication is primarily by email when needed. This is a particularly interesting case to study for at least two reasons. First, the high degree of anonymity is advantageous because it increases our confidence that our findings are driven by our treatment rather than some unobserved factors related to the interaction between employer and employee.

Second, this setting allows us to comment on the likely effects of increasing the saliency of employer characteristics. Interestingly, we find that minority groups might benefit on some margins but not others. This is unlike the existing literature, which finds almost unanimous evidence that minority groups face discrimination when race and sex are salient. These results are suggestive of the possibility that discrimination by employees toward employers might be different than the traditional setting where employers discriminate against employees (see above for the conceptual differences between these dimensions of discrimination).

One implication of our findings is that crowd-source labor markets that are designed with strong contract-based arrangements like mTurk can minimize their employer’s exposure to discrimination on the basis of race and sex by reducing the saliency of these characteristics. While mTurk maintains a strong sense of anonymity, this is not true of all similar labor markets. For example, some of these labor markets require both employers and workers to establish user-profiles with names and pictures.

The remainder of the study proceeds as follows. We describe the experimental design and data in section 2. We present the empirical strategy and main results (including heterogeneity) in section 3. Section 4 presents results of a follow-up study to distinguish statistical and taste-based discrimination. We conclude the paper in section 5.

2 Design and implementation

Our objective is to determine whether workers consider employers’ race or sex when making labor supply decisions. We isolate the effect of employer’s race and sex on workers’ labor supply via a randomized experiment on a crowd-sourcing labor platform. The remainder of this section provides a detailed description of the experimental design.

2.1 Design

Recruitment. We recruit subjects based in the U.S. from Amazon’s Mechanical Turk (mTurk) using a Human Intelligence Task (HIT) that invites subjects to complete a road mileage user-fee survey for a flat fee of \$0.65.⁹ All subjects who decide to complete this survey are directed to the external website of a survey provider (Qualtrics). The invitation to the road mileage user-fee survey and the entire survey itself do not include any signals regarding the race or sex of the employer (see below for how this was achieved). Upon completion of this initial survey, subjects complete a brief demographic-questionnaire and are then randomly assigned to one of five treatment groups (between-subjects design) that differ only in the race/sex signal of the

⁹The results of this survey are used in an unrelated paper about road mileage user fees.

employer that we send to subjects. Figure 1 provides an illustrative diagram of the flow of the experiment. The interaction between employers and employees in our experiment is a one-time event. This design feature removes any reputational motives among employers and employees.

Treatment. Once the software has assigned subjects to treatment groups, we thank them for completing the mileage user-fee survey, and inform them that there is an opportunity to earn additional income by completing a transcription task for the same employer (called requester in mTurk language). The additional task and its description to subjects are identical across treatment groups; subjects are asked to transcribe gas station name, date of purchase, gallons of gasoline purchased, price per gallon, and total sale value from gasoline receipts hoarded by one of the authors. We also tell them the approximate time it will take to transcribe the information from one receipt (approximately 30 seconds) and the wage per receipt (\$0.06). See Figure 2 for a screen shot of the details shown to subjects at the time they receive treatment.

When we inform subjects about the additional working opportunity, we show them an example of the gasoline receipts which are to be transcribed in the additional task (see Figure 3). Subjects perceive these pictures to be an illustration of the transcription task. The picture that we show subjects at this stage is identical across treatment groups except for the signal of race and sex (see below). We start with a stock of gasoline receipts that one of the authors collected over a four year period for tax reasons. From this stock of receipts, we selected approximately 100 receipts that were in good condition; all of the information we wanted subjects to transcribe was visible.

Following Doleac and Stein (2013), we signal race and sex of the potential employer by showing subjects a picture of a hand holding a gasoline receipt. For this purpose, we selected four hand models; one black and one white female, and one black and one white male. In order to make race and sex salient, we selected black hand models with dark skin, and asked the females to wear nail polish. We then conducted a photo-shoot where we took a picture of each person holding each of the receipts; approximately 100 pictures per person. The pictures included only the receipt and the models' hands. Finally, we selected the most clear pictures for each hand model. From this list, we identified the set of clear receipts that were common across models. This left us with 40 receipts, which were used in the experiment.

Our treatment groups differ with respect to the picture that subjects are shown when we illustrate the additional working opportunity to them. We have five treatment groups:

- i) **Black-Female** (BF): Subjects see receipts held by a female hand with black skin and nail polish.
- ii) **Black-Male** (BM): Subjects see receipts held by a male hand with black skin (no nail polish).
- iii) **White-Female** (WF): Subjects see receipts held by a female hand with white skin and nail polish.
- iv) **White-Male** (WM): Subjects see receipts held by a male hand with white skin (no nail polish).

v) **Control:** Subjects see receipts that do not include a hand.

Note, again, that the receipts available for transcription are identical across treatments. Receipts are presented in the same order across subjects and groups.

After we expose subject to one example of these receipts, we ask them if they would like to work in the additional task and transcribe the gasoline receipts (see Figure A.1). Because the transcription was not included in the initial recruitment HIT, we make it clear to the subjects that the transcription task is optional and that there is no penalty for opting out. Subjects who respond *yes*, transcribe receipts sequentially and are allowed to exit the task after each receipt (see Figure A.2 for details).

Outcome Variables. The experimental design allows us to measure three outcomes; one extensive-margin outcome and two intensive-margin outcomes. First, we measure an extensive-margin response based on a subject’s decision to accept the additional labor task or not.¹⁰ Second, among all subjects who say *yes* and proceed to transcribing pictures (i.e., conditional on deciding to work on the additional task), we measure the number of receipts transcribed. Recall that subjects could exit after each transcribed picture and that the maximum number of receipts that could be transcribed was 40. The majority of workers did not transcribe 40 pictures and we observe sufficient variation in this variable. Third, we measure transcription accuracy among transcribers by comparing each subject’s transcriptions with the actual information on each receipt.¹¹

Post-treatment survey and treatment saliency. One of the features of our experimental design is that we run a post-experimental survey in which we collect data on whether subjects discern the race and sex of the person in the pic that was randomly exposed to them. For this purpose, we ask the following two post-experiment survey questions of all subjects, including those who decided not to transcribe any of the receipts:

1. What is the race of the person holding the receipt in the picture?
2. What is the sex of the person holding the receipt in the picture?

¹⁰Extensive-margin in our context refers solely to the decision to transcribe gasoline receipts or not; see footnote 3 above in which we acknowledge this point.

¹¹Accuracy was calculated as follows: Each receipt had seven items for subjects to transcribe. Let n_i be the number of receipts transcribed by subject i . Then the total number of items transcribed by subject i is $T_i = 7 * n_i$. If we define c_i as the number of correct items for subject i , then the accuracy rate for subject i is $a_i = c_i / T_i$. The first step in creating this variable is to transcribe the receipts ourselves. Second, we compare our transcription of each item to the corresponding transcription for each subject and adopt two separate rules to identify correct entries. The first rule is strong in the sense that an entry is correct if it is an exact match to the corresponding entry on the receipt. This decision rule does not allow for rounding of dollar figures. However, the receipts report dollar figures to three decimal places and we do observe that some subjects round these entries. Further, there is a lot of variation across subjects in the rounding rule; some round to two decimals place, others to one decimal place and so on. Because we did not include any instructions about rounding in the experiment, we adopt a weaker definition of accuracy, where an entry is labeled accurate if it matches the corresponding entry on the receipt or any of its possible rounded representations. So, if the receipt lists price at \$2.476, then \$2.476, \$2.48, \$2.5, and \$2 would all be coded as accurate under weak accuracy, while only \$2.476 would be coded as accurate under strong accuracy. Following this procedure, we calculate $a_i = c_i / T_i$ for each subject.

Possible responses are black/female, white/male, 'I don't know', and 'The picture did not include a person'. We randomized the order of the questions and the answer possibilities to control for any order effects. We use responses to these questions to create a measure of (self-reported) treatment-salience. The salience measure indicates if subjects' perceived treatment is equal to their actual treatment, where perceived treatment is based on the subjects' responses to the questions above. For example, a subject's perceived race-treatment is black if her response to question 1 is black. Additionally, she is labeled race-salient if her actual treatment is black; i.e., if her perceived race-treatment matches her actual race-treatment.

We use the treatment-salient measures to study whether treatment effects of the randomized pictures are different between subjects who responded correctly to these questions (salience subjects) and subjects whose responses were incorrect (nonsalient subjects). If, for example, we see non-zero effects for nonsalient subjects, this might be an indication for strategic misreporting to conceal discriminating behavior.

Motivation for the flow of the experiment. We choose a design where subjects recruited via a mTurk HIT flow through the experiment as follows (see Figure 1 for a graphical illustration): complete an unrelated survey, be provided the opportunity to work on an additional labor task and get exposed to randomized sex and race signals, decide to accept or reject additional labor task, complete labor task if accepted additional task task, complete post experiment survey (both if accept or decline additional labor task).

The key advantage of this design flow is that we are able to collect information on all subjects who are exposed to treatment whether or not they decide to work in the additional labor task. As a result, we are able to study the effect of the treatment on the decision to say *yes* or *no* to our labor task HIT (i.e., what we label the extensive margin). Studying the extensive margin is plausibly the most relevant response margin and therefore very important. An alternative design flow would have been to signal race and sex directly when the HIT is advertised on the mTurk platform. However, this approach would not have allowed us to study the extensive margin since we would only be able to observe people who eventually accept the HIT. In addition, we are able to relate better to the usual type of discrimination papers which study the effect of randomly provided signals of applicants (e.g., via fake CVs) on what is comparable to our extensive-margin response, the decision to be hired, called back or invited for an interview. Our design also allows us to measure labor effort conditional on accepting the task; we use number and accuracy of transcription to measure effort.

Salience of employers' race and sex. There are three sources of saliency to consider. First, it is important that subjects do not select the initial (survey) mTurk HIT based on their perception of the requester's sex or race (because we only have data for workers who accept the initial HIT). When subjects see a HIT on mTurk and decide on whether to accept the HIT, they do not receive any information about the requester except the requester's name. In order to ensure that selection of the initial HIT is not based on the sex or race of the requester, we select a requester-name 'Alex Wright', which is mostly neutral with respect to race and

sex. This approach should minimize the likelihood that subjects select the HIT based on the requesters race or sex. We can confirm that the demographic characteristics of our sample is very comparable to that of other samples that the authors and other researchers have recruited from mTurk in the past.¹²

Second, we want to make sure subjects receive the signal we intended to send via the hands in the pictures. The saliency of the race and sex signals were tested in a **pilot-experiment**. Subjects were recruited on mTurk (N=120) to view a picture and answer questions about the picture. Subjects were randomly assigned to one of four groups and each subject in each group was shown one picture with a sex/race mix: black female, white female, black male, white male. As shown in Table B.1, we found that the majority of subjects correctly identified the race and sex of the hands. Specifically, 83% of subjects correctly identified the race and sex of the black-female hand; 62% and 75% correctly identified the race and sex of the black-male hand, respectively; and 90% of subjects correctly identified the race and sex of the white-female hand. The race of white-male hand used in the pilot was correctly identified 86% of the time, but the sex was only correctly identified 25% of the time. In response to the latter pilot result, we changed the white-male hand model for the actual experiment, but did not run another pilot. However, the race and sex of the white-male hand used in the actual experiment was correctly identified by 79% and 69% of the subjects, respectively. The treatment is presented in a way that should maximize the salience of the hands. After subjects complete the mileage survey and click submit, they are taken to a new page that has the picture of the receipt at the top of the page. Depending on the size of the subject’s screen, the picture with the receipt and the hand will be the only thing the subject sees before scrolling down the page.¹³

We follow-up this design feature with a survey at the end of the experiment to capture subjects perception of their treatment status. Subjects are asked about the race and sex of the person in the picture they saw. We also ask subjects about the United States president in order to check if subjects were paying attention. As will be important, note that the post-experimental survey is self-reported and subjects might strategically lie in the post-experimental survey.

Finally, we want to make sure subjects make the connection between the hand in the picture and the requester. We attempt to increase the connection between the hand and the requester by writing the mTurk HIT and the treatment in first-person singular ‘I’. For example, the mTurk HIT includes language like “I would like your opinion about the move toward the mileage tax.” Similarly, the instructions subjects see when they receive treatment is written with the intent of connecting the hand to the person making the job request; see Figure 2. For example, we tell subjects “*I want to know how much I would pay in mileage tax compared to what I now pay for gasoline tax*”, “*I would like you to transcribe information from **my** gasoline receipts*”, and “*I have included a sample of one of **my** receipts above*”.

¹²On average, our sample is 78% white, 56% with B.Sc. or higher, 37 years old, and 48% female (see Table 1). This is comparable to the US sample in Bohren et al. (2019) and the samples in Duncan and Li (2018) and Kuziemko et al. (2015).

¹³Note that we measure intent-to-treat effects and any differences across experimental groups can be attributed to the experimental variation even in a situation where some workers did not discern the race and sex of the hand in the picture.

2.2 Implementation

The experiment was conducted on Qualtrics using subjects recruited from mTurk. We first create a human intelligence task (HIT) that is advertised on mTurk. The HIT includes a description of the initial survey and compensation. We deliberately exclude any mention of the transcription task in the HIT. Instead, we recruit a large sample of subjects to complete a survey and then introduce the treatment. In this way we are able to collect data on all of subjects that are randomly assigned to one of our treatments, even if they subsequently refuse to transcribe the receipts. Note that we present the transcription task as an additional working task to subjects and that we give subjects the deliberate and explicit opportunity to quit after the initial survey. This ensures that they do not feel confused when they are presented the additional task which was not initially mentioned in the advertisement of the HIT on the mTurk website.

Subjects are told to accept the HIT and click on the weblink if they are interested in completing the survey. Subjects who click on the link are taken to our Qualtrics site where they complete the survey before being assigned to a treatment group to transcribe images. We selected the mileage user-fee survey and gasoline receipt transcription task because it allowed us to present the whole experiment as one event being implemented by a private citizen who is concerned about her state potentially adopting a road mileage user-fee (thus the first survey part) and who is a frequent driver (thus the gasoline receipts). We view it as advantageous that both parts of our study – the initial survey on road mileage user-fees and the subsequent experiment with gasoline-receipt transcription – are in the context of car driving; this makes it appear like an integrated set-up with related components. This reduces the likelihood that subjects view the HIT as part of an academic study thus preserving the reliability of their decisions and responses. Transcribing text from a scanned or photographed receipt is a common type of task on mTurk. This further reduces the chances that subjects realize they are participating in an experiment.

We chose to run the experiment on mTurk for several reasons. First, mTurk is one of the largest online labor markets where job offers are posted and workers choose jobs for payment. According to Amazon, there are over 500,000 workers from 190 countries in the mTurk labor market: <https://requester.mturk.com/tour>. Therefore, mTurk has a special place in the digitally-mediated labor markets that have come on the scene in the last decade. Second, experimenter effects are avoided because subjects do not know that they participate in an experiment (Paolacci, Chandler, and Ipeirotis 2010; Horton, Rand, and Zeckhauser 2011; Buhrmester, Kwang, and Gosling 2011; Mason and Suri 2011). Importantly for us, we are able to identify the effect of race and sex in a naturally occurring labor market. In general, experiments on Amazon’s Mechanical Turk therefore combine internal and external validity since it is a “real” labor market with actual workers where randomized trials can be conducted (Horton et al. 2011).¹⁴

¹⁴Kuziemko et al. (2015) and DellaVigna and Pope (2018) are recent examples of economics papers using Amazon’s Mechanical Turk.

Payment. The experiment ends for each subject when she decides to stop or when she transcribes 40 pictures, whichever comes first. In either case, each subject is instructed to copy her personal ID number and paste it in the entry box on the mTurk website. This process is necessary for us to match subjects to their mTurk worker ID and thus process their payments. Subjects receive a participation reward of \$0.65, which is paid as long as a subject accepts the HIT and completes the survey. Additionally, subjects are paid a piece rate of \$0.06 for each transcribed receipt. Given the payment restrictions imposed by the mTurk platform, we frame the piece rate as a bonus in all communications to the subjects. Overall, we paid a total of \$2419 for 2500 subjects who took an average of 7.8 minutes to complete the study; this translates to an hourly effective wage of approximately \$7.4, which is above the Federal minimum wage (\$7.25 per hour since 2009).

2.3 Data Summary

Data Cleaning. We fielded the experiment in two waves collecting 1250 responses each time for a total of 2500 subjects; approximately 500 observations per treatment group.¹⁵ We cleaned the data in the following ways before performing our empirical analysis. First, we calculate the total time taken to complete the experiment in minutes and trimmed the top 5% and bottom 5% of the sample. This removed subjects who took fewer than 2 minutes or more than 38 minutes; 267 subjects uniformly distributed across our treatment groups. Second, we drop 66 subjects who stated in check questions that the president of the US is Michael Jordan since this is an indication that subjects were simply clicking through the study. Finally, we drop 68 cases where subjects had the same IP-address because this might be an indication that the same subject is taking the experiment multiple times or it could be that turkers from other countries are taking the experiment when they should not. These adjustments leaves us with 2174 total observations; a bit over 430 subjects per treatment. Importantly, these adjustments were equally distributed across treatment groups.

Demographic Characteristics. Because we are interested in race and gender discrimination and the groups are very similar to each other on observables (see appendix Table B.2), we combine the treatment groups in the following ways for our analyses: black, white, control, male, female. Summary statistics for these race and sex combinations are presented in Table 1. Overall, our sample is typical of other mTurk samples; average age of 37, 78% white, 48% female, 51% urban, and highly educated with approximately two-thirds of subjects having at least a two-year college degree. Data from the 2018 American Community Survey suggests that our sample is fairly comparable to the U.S. population on age, race and sex. However, the mTurk sample is less urban and more highly educated than the U.S. population.

For the purpose of comparing demographic characteristics across treatment groups, we take the difference in means between treatment and control groups for each demographic vari-

¹⁵As indicated above, the HIT included two parts: a mileage user-fee survey and a transcription task. The current paper analyses the data from the transcription task. The mileage user-fee data are used to write a separate paper on public opinion of mileage userfees.

able and present these results in Table B.3 along with p-values from a ranksum test of the null hypothesis that the means are the same.¹⁶ Of particular interest is the difference between race and sex groups. Except for education, there is no statistically significant difference between the female and male treatments. We find that, relative to the male treatment, the female treatment has 4 percentage points fewer subjects with a B.Sc. degree and four percentage points more subjects with a Graduate degree; p - value = 0.089 & 0.011, respectively. Similarly, subjects' race is the only statistically significant difference between the black and white treatment groups; four percentage points more white subjects in the black treatment relative to the white treatment. We control for these variables in the empirical analysis and find that they do not change our results.

Salience of Treatment. Figure 4 shows the share of subjects whose perceived treatment matches their actual treatment, for race and sex (i.e., share of subjects who are treatment-salient). Note, again, that the numbers in this figure are based on the self-reported post-experiment survey, where subjects potentially lie for strategic reasons (for example, in order to conceal their preceding discrimination behavior). We find that just under 40% of subjects in the black treatment correctly perceived their treatment compared to over 80% of subjects in the control and white treatment groups. Similarly, subjects in the male treatment were more likely to correctly perceive their treatment status than subjects in the female treatment; 65% for males versus 47% for females.¹⁷ The summary statistics in Table B.4 show that the demographic profile of subjects is mostly similar across race and sex salience. Subjects for whom race was salient tended to be modestly younger and from urban areas, while sex-salient subjects tended to be modestly younger.

3 Empirical Strategy and Results

This section describes our empirical strategy and results. We present extensive-margin results followed by results for number of receipts transcribed and accuracy.

3.1 Empirical Strategy

We estimate Equation 1 to determine if subjects consider requesters' race and sex when making their labor supply decision in the transcription task.

¹⁶The majority of the differences between treatment and control groups are statistically indistinguishable from zero. Notable exceptions are race, sex, and education where we observe small differences between the treatment and control groups in some cases.

¹⁷Results presented in Figures C.1 to C.6 show subjects' responses across the possible responses on the post-experiment race and sex questions. We find that just under 40% of subjects in the black treatment correctly perceived their treatment, while 30% reported being in the white treatment, 7% reported being in the control group, and 24% did not know the race of the hand in the picture. On the other hand, over 80% of subjects in the control and white treatment groups correctly perceived their treatment status with the remaining subjects mostly saying they don't know the race of the hand. The findings are somewhat similar when we look at the salience of the sex treatments. Subjects in the women treatments are more likely to misperceive their true treatment status.

$$y_i = \alpha + \beta Treatment_i + \delta X_i + \epsilon_i, \quad (1)$$

where y_i is one of three outcome variables for subject i ; bonus-task acceptance, number of receipts transcribed, and accuracy. Bonus-task acceptance is an indicator variable that takes a value of 1 if the subject accepted the transcription task and zero otherwise. Number of receipts transcribed is the number of receipts that a subject transcribes. Accuracy is measured by the share of accurate entries (see footnote 11). X is a vector of subject-level covariates including age, sex, race, education and urban, and ϵ_i is a standard error term.

We estimate equation 1 using a subject’s randomly determined treatment assignment. When we estimate discrimination with respect to race, *Treatment* is equal to 1 if the subject was assigned to a black-hand treatment and zero if the subject was assigned to a white-hand treatment. When we study sex-based discrimination, *Treatment* is equal to 1 if the subject was assigned to the female-hand treatment and zero if the subject was assigned to a male-hand treatment.¹⁸ Therefore, β is the estimated race or sex gap depending on the specification; positive values indicate that discrimination benefits the minority group (black or female employers). The model provides intention-to-treat (ITT) effects of the randomly assigned race and sex signals. In these specifications, Equation 1 is estimated using the full sample and separately using the treatment-salient and treatment-nonsalient samples. We also explore inter-group responses by estimating the model separately for white and non-white workers, and female and male workers.

The intensive-margin responses (number of pictures transcribed and accuracy) are only observed for subjects who decide to accept the transcription task. Therefore, any observed race or sex effect is conceptually driven by the randomly assigned treatment and selection (i.e., incidental truncation). In particular, it is possible that the subjects who respond to the black-hand treatment differ from those who respond to the white-hand treatment in ways that also influence effort and accuracy. We check for the severity of this problem by comparing the characteristics of the subjects who accepted the task across treatment groups. The results presented in Table B.5 indicate that the groups are similar on age, sex, race, education level, and urbanization.

3.2 Results Across Full Sample

Extensive Margin: Bonus-task Acceptance. Figure B.2 reports the overall acceptance rate across race and sex groups. The figure shows that the mean acceptance rate was approximately 36% across treatment groups.¹⁹ Importantly, there does not appear to be much difference in subjects’ willingness to transcribe receipts across employer characteristics. We estimate equation 1 to check whether workers decision to transcribe receipts was influenced by the employers race or sex and present the results in Table 2. The result for task acceptance is

¹⁸We exclude the control group in these specifications. However, the result we obtain is the same as if we estimated $transcribed = \alpha + \beta_b black + \beta_w white + \epsilon$ and then calculate $\beta = \beta_b - \beta_w$.

¹⁹Figure B.1 shows that including a hand did not affect the extensive margin decision to transcribe receipts. Receipts with a hand had an acceptance rate of approximately 37% compared to 35% for the control group. See Figure B.3 for detailed results across treatment groups.

presented in the first column of Panels A and B for race and sex, respectively. The estimated coefficients are practically zero for both race and sex, which indicates that, on average, subjects were equally likely to work for black/female employers as white/male employers.

Intensive Margin I: Number of Pictures. Approximately 36% (or 753) of subjects transcribed at least 1 receipt, and subjects transcribed an average of 7.6 receipts. However, the distribution is highly skewed; the median number of transcribed receipts is 3, 75% of subjects transcribed fewer than 9 receipts, 90% transcribed fewer than 21 and only 37 subjects transcribed all 40 receipts. Figure B.4 shows how the mean number of transcribed receipts varies across treatment groups. Subjects in the control group transcribed 10 receipts on average. While subjects in the treatment groups transcribed fewer receipts than those in the control group, the reduction appears to be uniform across the race treatments but slightly larger for male employers compared to female employers.

We estimate intensive margin effects using equation 1 with ‘number of transcribed receipts’ as the outcome variable. The results in the second column of Panels A and B of Table 2 show the treatment effect on number of receipts transcribed for race and sex, respectively. We find that workers transcribed more pictures for female employers relative to male employers (p -value = 0.059) and fewer receipts for black employers relative to white employers (p -value = 0.2). The estimated effects are approximately 1.5 additional picture for females and 1 fewer picture for blacks, which is approximately 15% and 10% of the standard deviation of number of pictures transcribed for sex and race, respectively.

Intensive Margin II: Accuracy. We define accuracy as the share of correct entries across all transcribed receipts for each subject and test whether accuracy differs depending on employer’s race and sex (see footnote 11 for a more detailed description of the measurement of accuracy). Figure B.5 summarizes the accuracy of transcription across race groups. We find that subjects in the control group got about 87% of their entries correct. The corresponding rate for the treatment groups is 76% and 80% for black and white groups, respectively, and 80% and 75% for female and male employers, respectively.

The third column of Panels A and B of Table 2 shows that on average, workers are more accurate for female employers and less accurate for black employers. Workers were 4.4 percentage points more accurate for female employers and -6.3 percentage points less accurate for black employers; both estimates are precisely estimated.

3.3 Heterogeneity w.r.t. Worker’s Race and Sex

This section explores the within group dynamics by cutting the sample by the race of workers when estimating the race gap, and by the sex of workers when estimating the sex-gap. We classify workers as either white or non-white based on their responses to the survey. The non-white group is a fairly small share of the total sample (only 22%). Therefore, we are careful when interpreting the race estimates.

Panels A and B of Table 3 show the estimated race gap among non-white and white

workers, respectively. The results are suggestive of in-group bias. On the one hand, we find no evidence of a race gap for task acceptance and accuracy among non-white workers. However, nonwhite workers transcribe about two extra pictures for black employers compared to white employers. On the other hand, while white workers were equally likely to accept the task for white and black employers, we find that white workers transcribed fewer pictures for black employers and did so less accurately.

Panels C and D report results for the sex gap by workers' sex. Female workers are about 5% more accurate when working for a female employer but otherwise treat both male and female employers equally with respect to task acceptance and number of transcriptions. Although male workers are equally likely to accept the transcription task for male and female employers, they transcribe 2.6 additional pictures and are about 4% more accurate when working for female employers.

3.4 Salient vs Non-salient Subjects

The null extensive-margin effect described above could be due to either a lack of treatment saliency or a lack of discrimination among workers. Results from our pilot study described in Section 2.1 demonstrate that the treatment is salient, which suggests that lack of saliency is an unlikely explanation of the finding. This section of the paper further explores the saliency angle by using responses to a post-experimental survey that asks subjects about the sex and race of the person holding the receipt in the treatment photograph to define treatment saliency. We classify a subject as race-salient if her self-reported race treatment matches her actual race treatment. Sex salient subjects are defined similarly. We estimate equation 1 separately for salient and non-salient subjects and present the results in Panel A and B of Table 4 for race and sex, respectively. We then split the sample again to explore the effect of saliency within worker type; these results are presented in panels C to F.

3.4.1 Treatment-Saliency Across All Workers

Extensive Margin: Bonus-task Acceptance. Interestingly, we find large and statistically significant effects when we split the sample based on the saliency of the treatment. Subjects for whom the race treatment was salient were 10.8 percentage points more likely to work for a black employer than a white employer. Similarly, sex-salient subjects were 13.5 percentage points more likely to work for a female employer than a male employer. Both estimates are significantly different from zero.

Our null result in the full sample combined with large treatment effects among the 'salient' subjects suggests large negative treatment effects in the non-salient sub-samples. This is precisely what we find. Subjects who "misperceived" their treatment were less likely to work for black or female employers. The estimated gap among the non-salient sub-sample is -9 percentage points (p-value: 0.007) for sex and -4.3 percentage points (p-value: 0.312) for race. This finding is puzzling since we would have expected to find a null result among subjects who paid no attention to the race or sex of the employer. We explore this finding later.

Intensive Margin I: Number of Pictures Transcribed. We find that transcriptions among the non-salient group are in line with expectations; the estimated gap is practically zero for both race and sex. However, we do find evidence that workers for whom the sex-treatment was salient transcribed more pictures for female employers compared to male employers. The comparable gap for race is practically zero.

Intensive Margin II: Accuracy. The accuracy results in Panels A and B show that treatment-salient workers are significantly more accurate for female employers and less accurate for black employers. The gaps point in the same direction among non-salient workers, but the coefficients are smaller and not statistically different from zero. As with the other intensive-margin response (transcribed picture), we thus find no evidence of treatment effects among treatment-nonsalient workers.

3.4.2 Within-group Analysis

Extensive Margin: Bonus-task Acceptance. Panels C and E of Table 4 show that both white and non-white workers for whom the employer’s race was salient were more likely to work for the black employer. This finding is stronger among non-white workers. Interestingly, we find no statistically significant evidence of a racial gap among workers for whom the employer’s race was not salient. However, there is an economically substantive gap among non-white workers for whom race was not salient; the gap among white workers is both economically small and statistically indistinguishable from zero. Panels D and F report similar within-group results for the sex-gap. In particular, we find a large positive and statistically significant gap among both male and female workers for whom the employer’s sex was salient. Interestingly, we find a strong negative sex-gap among male workers for whom the employers sex was not salient.

Intensive Margin I: Number of Transcriptions. There is suggestive evidence of within-group bias for the race-gap and out-group bias for the sex-gap; see second column of Panels C-F of Table 4. In particular, we find that non-white workers complete more transcriptions for black employers than white employers, while white workers complete more pictures for white employers than black employers. On the sex margin, we find strong evidence that sex-salient male workers complete more transcriptions for female employers than male employers. However, female workers treat male and female employers equally when deciding how many receipts to transcribe. We find no evidence of a sex or race gap among non-salient workers, which is what we would expect if subjects did not perceived the treatment signal.

Intensive Margin II: Accuracy. Table 4 shows evidence of within-group bias among race-salient white workers; they were more accurate when working for white employers compared to black employers. There is no evidence that non-white subjects’ accuracy level varied with the race of the employer. The results in Panels D and F show that both male and female sex-salient workers were more accurate when transcribing for female employers compared to male employers.

3.4.3 Implications of Saliency Results

We find that the extensive-margin response to the treatment differs between salient and non-salient subjects. This suggests that our salience variable identifies two types of subjects. The treatment-salient workers are biased toward the minority employers (black and female), whereas the treatment-nonsalient workers are biased against the minority workers. The fact that we find significant evidence of discrimination among subjects in the non-salient group on this margin is particularly interesting and open to multiple explanations. We explore this pattern in the data and discuss possible explanations in this subsection.

Figure 3 shows the race and sex signals that we sent to subjects (actual treatment) and Figure 4 shows the *self-reported* accuracy with which those signals were perceived. Recall that the perceptions are surveyed after the experiment and that they represent self-reported measures which are potentially subject to intentionally false answers. Approximately 80% of subjects in white treatments correctly perceived treatment status compared to only 37% in the black treatments. Interestingly, 31% of subjects in the black treatments reported that the employer is white, while 25% stated they did not know the race of the employer. Only 0.71% of subjects in the white treatments stated that the employer was black and 11.6% responded that they did not know the race of the employer. A similar pattern is observed for sex; 65% of subjects in the male treatments correctly identified the sex of the employer compared to only 43% in the female treatments. Additionally, 17% of subjects in the female treatments stated that the hand belonged to a male while only 1.8% of subjects in the male treatments said the hand belonged to a woman.

Considering that the signals in Figure 3 are clear, as suggested by results from our pilot study, then there are at least three possible explanations for the observed difference in subjects' perception of the employer's race and sex. First, it could be that subjects paid attention to the pictures and correctly identified the race and sex, but ex-post misreported the race and sex in an effort to conceal their biases. Second, it could be that subjects received and responded to the race/sex signal subconsciously in a manner that favors majority groups, but ex-post misreported the race and sex in an effort to conceal their biases. This type of channel is referred to as implicit bias in the literature (Bertrand, Chugh, and Mullainathan 2005). Finally, subjects could have ignored the treatment signal and thus failed to respond to the signal, and then guessed at treatment in the post-experiment survey.

The pattern of results presented in Table 4 is consistent with the first and second explanations. In particular, if non-salient subjects were inattentive to the treatment and simply guessed a response to the race and sex questions, then there should be no significant treatment effect in the non-salient samples. This is especially true since attentive and inattentive samples have very similar demographic profiles. Our results are not consistent with this predicted null-treatment effect.

Subjects for whom race was not salient were approximately 4 percentage points more likely to transcribe for white than black employers (although not statistically different from zero). The effect is even more dramatic if we exclude subjects who responded 'I don't know'; in

this case, non-salient subjects are 13 percentage points more likely to work for white employers than black employers ($p - value = 0.03$).²⁰

We also find that non-salient subjects are almost 10 percentage points more likely to transcribe for male employers than female employers (statistically significant). Unlike the race treatments, we also find a large difference among subjects who responded ‘I don’t know’; 25% ($N = 270$) transcribed in the female treatment compared to 34% ($N = 214$) in the male treatment. Furthermore, of the 842 subjects assigned to the female treatments, 147 reported that the hand belonged to a male employer and only 14% of these subjects agreed to transcribe images. Only 15 of the 849 subjects in the male treatment reported that the employer is a female. Overall, these patterns in the data are highly suggestive of strategic misreporting in the post-experiment survey among subjects who discriminate against minority groups on the extensive margin.

Additional evidence suggestive of intentional misreporting among non-salient groups on the extensive margin is the fact that we do not find any statistically significant effects among the non-salient sample on the intensive margin. One possible explanation for this null intensive-margin result among nonsalient subjects is the following. Assume our subject pool includes two types of subjects; those who notice our treatment signal and those who do not. Those who did not notice treatment can guess their treatment status correctly or incorrectly and are likely equally distributed in the salient and nonsalient groups. For those who noticed our treatment signal, the ones who are honest are in the salient group, while those who “misreported” for any reason but mostly due to the guilt of discrimination are in the non-salient group. Notice that subjects who misreported treatment status to cover up discriminatory reasons for their extensive margin response are not part of the sample used to study the intensive margin responses. Therefore, it is possible that the nonsalient sample used to study the intensive-margin responses is predominantly comprised of subjects who were inattentive to the treatment signal. This would explain the null intensive-margin results among nonsalient subjects.

Overall, these results are very suggestive of a split sample. Some subjects are biased toward minority groups and are not afraid to expose their bias. Based on the evidence that the non-salient workers are more likely to work for the male or white employer, we observe that other subjects are biased against minority groups (either consciously or sub-consciously). Those who discriminated against minority groups consciously then presumably try to hide their discrimination behavior in that they report ex-post that they did not notice the race or sex of the person holding the receipt. We argue that this explanation is supported by the fact that subjects identified the race and sex on the hands with high levels of accuracy in our pilot studies.

²⁰Subjects who responded ‘I don’t know’ transcribe at the same rate for both white and black employers; approximately 27% ($N = 213$ in the black treatment and $N = 98$ in the white treatment). Subjects in the black treatment who responded that the employer was white or that there was no hand in the picture transcribe at rates of 26% ($N=262$) and 32% ($N = 56$) for white and ‘no hand’, respectively. Subjects in the white treatment who responded that the employer was black or that there was no hand in the picture transcribe at rates of 33% ($N = 6$) and 43% ($N = 54$) for black and ‘no hand’, respectively.

4 Taste-based or Statistical Discrimination?

Our results suggest that mTurk workers consider an employer’s race and sex when making labor supply decisions. We explore the extent to which these results reflect an underlying preference for certain types of employers (Becker 1957) versus statistical discrimination (Arrow 1972; Phelps 1972). In particular, we report the results of a follow-up survey that we run on mTurk to disentangle taste-based and statistical discrimination as an explanation of the race and sex gaps.

In general, workers are interested in their working environment broadly defined to include wages being paid on time and in full, health risks, collegiality, and other amenities. However, the mTurk labor market is largely anonymous in that workers never meet employers. Additionally, workers complete their tasks in their own environment and on their own schedule. This implies that mTurkers generally do not have to worry about the working environment provided by the employer except for payment. Therefore, the likely source of statistical discrimination in our context is a worker’s expectation of being paid by the employer.²¹ Taste-based discrimination is another possible explanation for our findings. This mechanism of course requires that workers are able to identify the employer’s race or sex. Although most HITs are accompanied by the employer’s name, which may signal the employer’s race and sex, it is not clear if workers generally pay attention to the requester’s name when selecting HITs.

We designed a survey of mTurkers to obtain information on workers’ past experiences with employers and to shed additional light on the role of employer sex and race for mTurkers. The goal is to determine if mTurkers’ perception of likely non-payment is correlated with the race or sex of the employer. We are also interested in identifying the extent to which workers contact employers, pay attention to employer’s name, and whether workers consider employer’s characteristics when selecting a HIT.

4.1 Survey Design

The follow-up survey has four sections. First, we ask subjects to report their age, sex, race, and education. Second, we ask about their usage of mTurk; year they joined mTurk, whether mTurk is their primary job, and number of HITs completed per month. The third section asks about their experiences as an mTurker. Here we ask about frequency with which work is accepted by requesters, experience with requesters’ refusal to pay for work completed, communication with requesters, attentiveness to requesters’ name, and whether knowledge of a requester’s characteristics would affect likelihood of accepting a HIT. We also ask subjects about the likelihood of a requester paying a bonus for a completed task.

For the final section, we randomly assign subjects to one of five groups that correspond to the five treatment groups of the original experiment. That is, we randomly show survey

²¹Although requesters (i.e., employers) are required to hold funds in an Amazon account prior to publishing a HIT, a worker is only paid after the requester approves that worker’s work. Additionally, approval of a task does not guarantee that a bonus will be paid because employers are not required to hold bonus payments in an Amazon account prior to posting a job. In other words, the requester must first approve the work and then process each worker’s bonus separately.

respondents the same treatment pictures that we showed to workers in the original experiment and then asked them a set of questions. Recall that these pictures differed with respect to the race (black or white) or sex (male or female) of the person whose hand holds a gasoline-station receipt. The treatment is presented as a hypothetical scenario. Specifically, the subjects saw the following text: *“In the next set of questions, I am going to ask you about your perception of what Turkers like you are likely to do when faced with a transcription task.”*

Subjects are then asked three questions; i) what percent of mTurkers would accept the HIT?, ii) would you accept the HIT?, iii) how likely is it that the requester would pay the bonus accompanying the HIT?. We also ask subjects to identify the race and sex of the person represented by the hand in the picture (to measure treatment-salience). The full set of questions is available upon request.

Sample. The survey was fielded to 1012 mTurkers who did not participate in the original study. The data are cleaned as follows; we drop all duplicated ipaddresses ($N = 41$) and everyone who identified Michael Jordan as president of the US ($N = 16$). This leaves us with a sample of 955 subjects. Subjects took an average of 4.3 minutes to complete the survey and were paid a flat fee of \$1.

Balance across survey-experiment groups. We find no meaningful nor statistical difference in the observable characteristics between the race treatment groups (See Table D.1). There is a statistically significant difference in the two youngest age groups between the control and male treatments, but these differences are small. We also find that the survey sample is similar to the original real-effort sample in age, sex, race, and education. Importantly, the salience of treatment is identical between the real-effort experiment and follow-up survey samples (see Table D.2).

4.2 Descriptive Results

Prior mTurk experience. Approximately 25% of the subjects report joining mTurk as a worker before 2016, and 13%, 21% and 41% report joining mTurk in 2016, 2017 and 2018, respectively. Subjects report completing an average of 474 HITs per month in the full sample. Panel A of Figure D.1 shows that there is heterogeneity in HITs completed across subjects’ race but not sex; white subjects complete 106 more HITs per month than non-white subjects (Ranksum $p - value = 0.001$), while female subjects complete only 14 more HITs than male subjects (Ranksum $p - value = 0.29$). The mean monthly completed HITs is 488 in the control group, 479 and 462 for black-hand and white-hand treatments, respectively, and 500 and 440 for female-hand and male-hand treatments, respectively. These differences are not statistically distinguishable from zero (see Panel B of Figure D.1).

Experience with Employers/Requesters and Non-payments. Figure 5 presents the results of subjects responses about their experiences as mTurkers. Approximately 25% of subjects report that mTurk is their primary source of employment. The remaining summary information

in Figure 5 describes subjects’ experiences with requesters and are suggestive of both statistical discrimination and taste-based discrimination.

First, 45% and 53% of subjects report that their HITs are accepted ‘all the time’ and ‘most of the time’, respectively. Approximately 54% report being in a situation where the requester refused to pay for a completed HIT, and approximately 80% have contacted a requester in the past. Additionally, subjects reported being 67% confident that a requester who offers a bonus task on an external website, such as was the case in our real-effort task, would pay the bonus upon completion of the task. This suggests that there is a significant amount of doubt about payment in the subjects minds as they make their HIT selection decisions.

Role of employer characteristics for decision to accept a HIT. To the extent that concerns about payment is correlated with perceived race or sex, subjects could use this prior experience to form expectations about the honesty of the requester. Therefore, rather than selecting on the basis of taste, subjects could instead be selecting HITs on the basis of expected payment by the requester.

72% report that that they check the names of requester, and 71% of subjects report that they would consider a requester’s characteristics when making HIT selection decisions on mTurk. These responses could support both taste-based as well as statistical discrimination. It could be that subjects use requesters’ name and characteristics in order to identify probabilistically-honest requesters. Alternatively, subjects could possibly use this information to identify groups of requesters they have a deep-seated bias against.

Interestingly, we do not find any meaningful differences in these self-reported experiences across subjects’ race or sex. This suggests that differential experiences across race and sex is not a strong explanation for the group-dynamics we observe in our real-effort experiment.

4.3 Randomized Survey Experiment

We explore the possibility of separating taste-based from statistical discrimination by presenting subjects with the same treatments as in the original real-effort experiment and then asking about hypothetical acceptance and perceived likelihood of being paid. The question about likelihood of being paid allows us to identify the effect of a requester’s race or sex on workers’ perception of the requester’s honesty, which further allows us to comment on the source of the bias uncovered in our real-effort experiment.

Subjects’ responses to the three post-treatment questions across all treatment groups are summarized in Figure 6. Subjects reported that approximately 46% of other mTurkers would accept the HIT, but only 34% of subjects reported that they themselves would accept the HIT. So, while subjects thought the acceptance rate among other mTurkers is about 10 percentage points higher than what mTurkers actually did, the subjects’ personal acceptance rate is identical to the acceptance rate observed in the experiment.

Importantly for our analysis, subjects reported a 69% likelihood that the requester would pay the bonus. Again, this suggests that subjects have some amount of uncertainty about being paid at the time they make their HIT-acceptance decisions and this is suggestive of statistical

discrimination as a possible explanation for our results. However, Table 5 shows that the sex of the requester has no bearing on workers' uncertainty about payment; estimates are both small and statistically indistinguishable from zero across all samples (full sample, treatment-salient sample and treatment non-salient sample). Similarly, there is no statistical or economic evidence that the race of the requester affects the uncertainty of being paid in the full and race-nonsalient samples. Although, race-salient subjects reported that they thought black requesters were about 4 percentage points less likely to pay than white requesters, the gap is not statistically different from zero (p-value 0.13).

The fact that subjects thought both male and female employers are equally likely to pay suggests that the sex-gaps observed in our experiment are not driven by statistical discrimination. Sex-Salient subjects in the real-effort experiment were more likely to work for females while sex-nonsalient subjects were more likely to work for males despite the fact that both sexes are equally likely to pay for work. The results for race are similarly suggestive of an even stronger rejection of statistical discrimination. Notice that subjects in our salient-samples are more likely to work for the black employer despite the general impression among mTurkers that black employers are less likely to pay. Results for the non-salient sample are also suggestive of a taste-based effect. Subjects in this group were less likely to work for the black employer although the general impression among mTurkers is that there is no difference in likelihood of being paid.

We extend the analysis to account for the possibility of within-group bias by estimating the treatment effect separately for sex and race of the survey respondents, and reporting the results in Table 5. While there does not appear to be any statistical evidence for a difference in perceived likelihood of the employer paying among non-white subjects, we find that white subjects in the salient sample tend to believe that black employers are less likely to pay than white employers. These results further support the case for a strong taste-based source of bias; white workers in the salient sample preferred working for black employers despite the perception among their peers that black employers are less likely to pay. Similarly, non-white workers prefer black employers despite no apparent difference in perception of being paid.

We find no evidence of within-group bias in the sex treatments. Estimates are small and statistically indistinguishable from zero except in the non-salient sample where estimates are fairly large, but imprecisely estimated. Among subjects in the non-salient sample, female subjects report that female employers are about 6 percentage points less likely to pay than male employers, while male workers report that female employers are about 4 percentage points more likely to pay. These results are also inconsistent with statistical discrimination; males in the non-salient sample of the real-effort experiment were less likely to work for female employers despite the fact that their peers in the follow-up survey are of the view that female employers are more likely to pay than male employers.

Overall, the results from our follow-up survey are strongly suggestive that the sex and race gaps identified in our real-effort experiment are not driven by statistical discrimination.

5 Conclusions

We estimate the effect of employers' race and sex on the willingness of workers to persist on a labor task using data generated on Amazon's Mechanical Turk. We find no evidence of discrimination against employers for neither sex nor race on the extensive-margin. However, these null results mask important heterogeneity by salience of treatment, which is suggestive of misreporting by those who do discriminate against minorities. Our results also point to discrimination on the intensive margin. First, workers were less accurate and transcribed fewer receipts for black employers, relative to white employers. Second, workers were significantly more accurate and tended to transcribe more receipts (though the latter is not significant) for female employers.

The fairly strong preference for female employers in our study is consistent with the general trend toward a preference for female bosses in [Gallup polls on Work and Workplace](#).²² Only 5% of participants expressed a preference for a female boss compared to 66% preference for a male boss in the 1953 Gallop Poll. By 2017, the Gallup Poll results showed that the share of participants who preferred a female boss increased to 21% while the share for male bosses fell to 23%. The findings are also in line with [Elsesser and Lever \(2011\)](#) who find that when rating one's own boss, respondents who have female managers do not rate them lower than respondents who have male managers. We acknowledge that working for a female boss is not the same as working for a female employer. However, these results do convey some information about changing attitudes toward female employers.

Results from a follow-up survey suggest that the biases we detect are not driven by statistical discrimination. Although subjects express some uncertainty about the likelihood of being paid for mTurk HITs, this uncertainty is not caused by the sex of the employer. Furthermore, the effect of race on the likelihood of being paid is not consistent with statistical discrimination. On the one hand, we find that subjects who prefer to work for black employers believe black employers are less likely to pay than white employers. On the other hand, those who prefer to work for white employers perceive no difference in likelihood of paying between black and white employers. Therefore, to the extent that the likelihood of being paid is the primary channel through which statistical discrimination would manifest itself in our setting, this finding suggest that the biases we estimate are not driven by statistical discrimination.

References

- Abel, M. (2019). Do workers discriminate against female bosses? IZA Discussion Paper No. 12611.
- Acquisti, A. and C. Fong (2020). An experiment in hiring discrimination via online social networks. *Management Science* 66(3), 1005–1024.
- Agan, A. and S. Starr (2017). Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment*. *The Quarterly Journal of Economics* 133(1), 191–235.

²²The survey results can be found online at: <https://news.gallup.com/poll/1720/work-work-place.aspx>.

- Arrow, K. J. (1972). Some mathematical models of race in the labor market. In A. Pascal (Ed.), *Racial Discrimination in Economic Life*. Lexington, MA: Lexington Books.
- Asad, S. A., R. Banerjee, B. IIM, and J. Bhattacharya (2020). Do workers discriminate against their out-group employers? evidence from the gig economy. Working paper available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3544269.
- Ayres, I., M. Banaji, and C. Jolls (2015). Race effects on ebay. *The RAND Journal of Economics* 46(4), 891–917.
- Becker, G. S. (1957). *The Economics of Discrimination*. Chicago, IL: The University of Chicago Press.
- Benson, A., S. Board, and M. Meyer-ter Vehn (2019). Discrimination in hiring: Evidence from retail sales. Technical report, mimeo, available online: https://site.stanford.edu/sites/g/files/sbiybj8706/f/5141-discrimination_in_hiring_evidence_from_retail.pdf.
- Bertrand, M., D. Chugh, and S. Mullainathan (2005). Implicit discrimination. *American Economic Review* 95(2), 94–98.
- Bertrand, M. and E. Duflo (2017). Field Experiments on Discrimination. *Handbook of Economic Field Experiments* 1, 309–393.
- Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2019). Inaccurate statistical discrimination. Working Paper 25935, National Bureau of Economic Research.
- Brown, A. (2018). Ridehail Revolution: Ridehail Travel and Equity in Los Angeles. *Ph.D. thesis, University of California Los Angeles*.
- Buhrmester, M., T. Kwang, and S. D. Gosling (2011). Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6(1), 3–5.
- Chan, J. and J. Wang (2018). Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science* 64(7), 2973–2994.
- Coffman, K. B., C. L. Exley, and M. Niederle (2021). The role of beliefs in driving gender discrimination. *Management Science*.
- 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 . NBER working paper no. 24732.
- DellaVigna, S. and D. Pope (2018). What Motivates Effort? Evidence and Expert Forecasts. *Review of Economic Studies* 85(2), 1029–1069.
- Doleac, J. L. and L. C. Stein (2013). The Visible Hand: Race and Online Market Outcomes. *The Economic Journal* 123(572), F469–F492.
- Duncan, D. and D. Li (2018). Liar liar: Experimental evidence of the effect of confirmation-reports on dishonesty. *Southern Economic Journal* 84(3), 742–770.

- Edelman, B., M. Luca, and D. Svirsky (2017, apr). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics* 9(2), 1–22.
- Elsesser, K. M. and J. Lever (2011, dec). Does gender bias against female leaders persist? Quantitative and qualitative data from a large-scale survey. *Human Relations* 64(12), 1555–1578.
- Farrell, D. and F. Greig (2016). Paychecks, Paydays and the Online Platform Economy. *JPMorgan Chase and Co. Institute*.
- Farrell, D., F. Greig, and A. Hamoudi (2018). The Online Platform Economy in 2018: Drivers, Workers, Sellers, and Le. *JPMorgan Chase and Co. Institute*.
- Ge, Y., C. Knittel, D. MacKenzie, and S. Zoepf (2016). Racial and Gender Discrimination in Transportation Network Companies. NBER Working Paper No 22776, Cambridge, MA.
- Glover, D., A. Pallais, and W. Pariente (2017). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *The Quarterly Journal of Economics* 132(3), 1219–1260.
- Hedegaard, M. S. and J.-R. Tyran (2018). The Price of Prejudice. *American Economic Journal: Applied Economics* 10(1), 40–63.
- Horton, J. J., D. G. Rand, and R. J. Zeckhauser (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics* 14, 399–425.
- Hyperwallet (2017). The future of gig work is female. Technical report, available online at: <https://www.hyperwallet.com>.
- Jaeger, D. A., J. M. Nunley, A. Seals, and E. J. Wilbrandt (2020). The demand for interns. NBER working paper no 26729.
- Katz, L. F. and A. B. Krueger (2019). Understanding trends in alternative work arrangements in the united states. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 5(5), 132–146.
- Kuziemko, I., M. I. Norton, E. Saez, and S. Stantcheva (2015). How elastic are preferences for redistribution? evidence from randomized survey experiments. *American Economic Review* 105(4), 1478–1508.
- Mason, W. and S. Suri (2011). Conducting behavioral research on Amazon’s Mechanical Turk. *Behavioural Research* 44, 1–23.
- Mazar, N., O. Amir, and D. Ariely (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research* 45(6), 633–644.
- Neumark, D. (2018). Experimental Research on Labor Market Discrimination. *Journal of Economic Literature* 56(3), 799–866.
- Nunley, J. M., M. F. Owens, and R. S. Howard (2011). The effects of information and competition on racial discrimination: Evidence from a field experiment. *Journal of Economic Behavior & Organization* 80(3), 670–679.

- Paolacci, G., J. Chandler, and P. G. Ipeirotis (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5(5).
- Phelps, E. (1972). The statistical theory of racism and sexism. *American Economic Review* 62(4), 659–61.
- Phillips, D. C. (2020). Do low-wage employers discriminate against applicants with long commutes? evidence from a correspondence experiment. *Journal of Human Resources*. forthcoming.
- Pope, D. G. and J. R. Sydnor (2011). What's in a picture? evidence of discrimination from prosper. com. *Journal of Human resources* 46(1), 53–92.
- Riach, P. A. and J. Rich (2002). Field Experiments of Discrimination in the Market Place. *The Economic Journal* 112(483), F480–F518.
- Zussman, A. (2013). Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars. *The Economic Journal* 123(572), F433–F468.

Main Tables and Figures

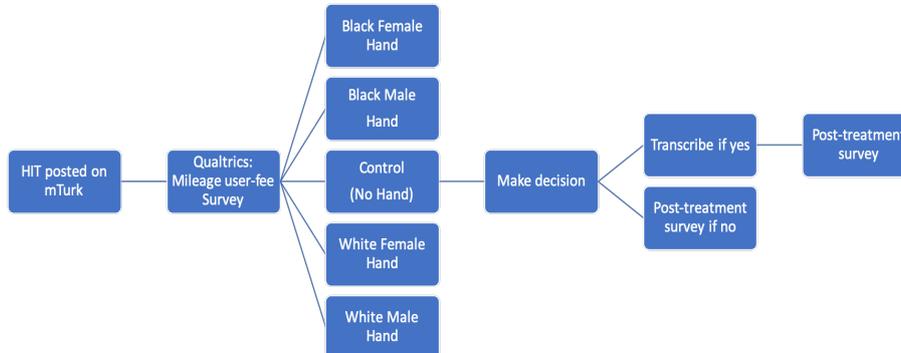


Figure 1: Overview of Experimental Design

Notes: Reported is the flow of the experiment. Subjects are recruited on Amazon's Mechanical Turk (mTurk) to complete a mileage userfee survey on Qualtrics. Subjects are randomly assigned to a treatment group where they are shown a picture of a hand holding a receipt and asked whether they would like to complete a transcription task. Subjects who respond yes transcribe images and then complete a post-experiment survey. Subjects who respond no complete the post-experiment survey.

Thank you! You have earned \$0.65 for completing my survey.

****below is an optional bonus opportunity****

I want to know how much I would pay in mileage tax compared to what I pay now for gasoline tax. To help me, I would like you to transcribe information from my gasoline receipts; this will allow me to estimate my annual gasoline taxes.

Each receipt should take approximately 30 seconds to transcribe, and I will pay you a bonus of \$0.06 for every receipt that you transcribe. You can stop at anytime.

I have included a sample of one of my receipts above. I would like you to transcribe the following information:

- 1.Name of the gas station*
- 2.Date of the purchase*
- 3.Gallons of gasoline purchased*
- 4.Price per gallon*
- 5.Total sale price*

Figure 2: Treatment Instructions

Notes: Reported are the instructions for the bonus transcription task.

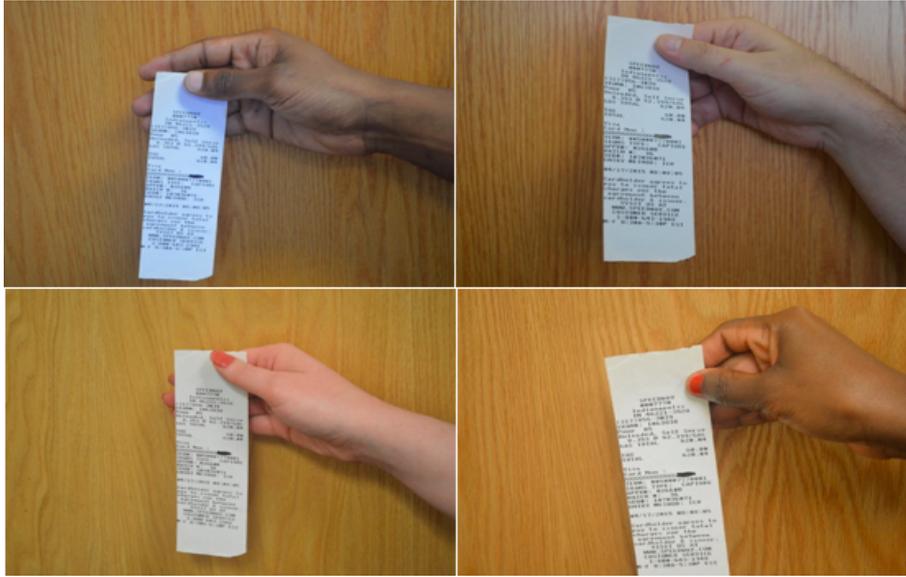


Figure 3: Treatment Pictures

Notes: Reported are the pictures used in the treatment stage of the experiment. The pictures have been compressed significantly to fit side-by-side on one page.

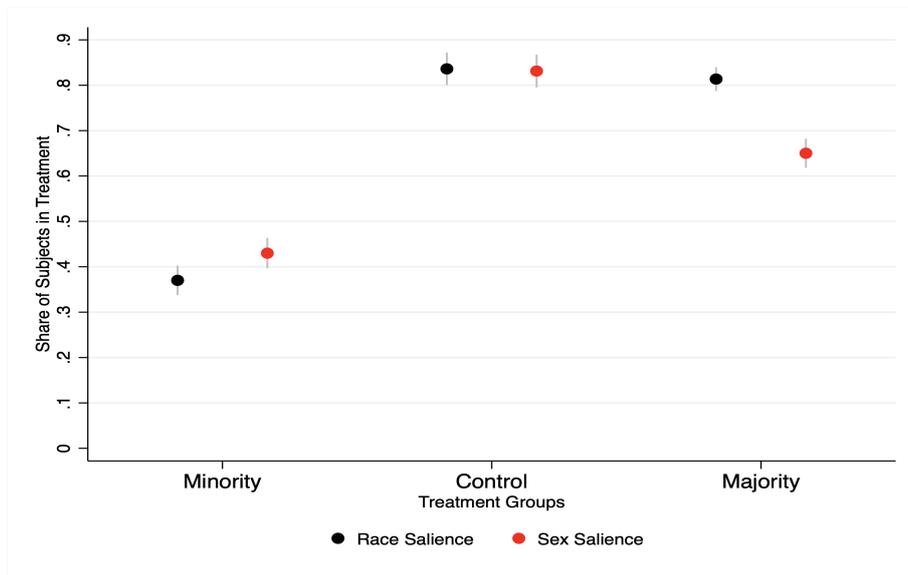


Figure 4: Salience of Race and Sex, by Treatment Group

Notes: Reported is the share of subjects whose self-reported perceived treatment matches the actual treatment they are assigned to for race and sex, respectively, along with 95% confidence intervals. Minority refers to Black-hand and female-hand treatments, while majority refers to white-hand and male-hand treatments.

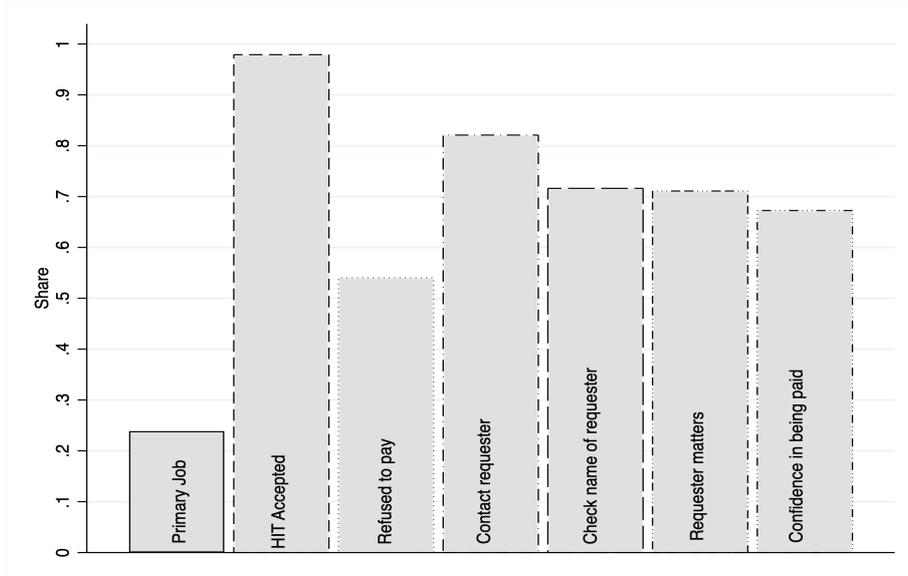


Figure 5: Follow-up Survey: mTurk Experience

Notes: Data based on follow-up survey. Reported is the share of mTurkers who answered “yes” to questions regarding their experiences as mTurk workers, including “*Is mTurk your primary job?*”, “*Is your work often accepted?*”, “*Have you experienced requesters’ refusal to pay for work completed?*”, “*Have you contacted requesters before?*”, “*Do you often pay attention to the requester’s name?*”, and “*Will a requester’s characteristics affect your likelihood of accepting a HIT?*”. the question “*how confident are you that a requester will pay for a bonus task?*” is measured a scale from 0 to 100. The variable was transformed to a 0 to 1 scale and mean confidence level is reported.

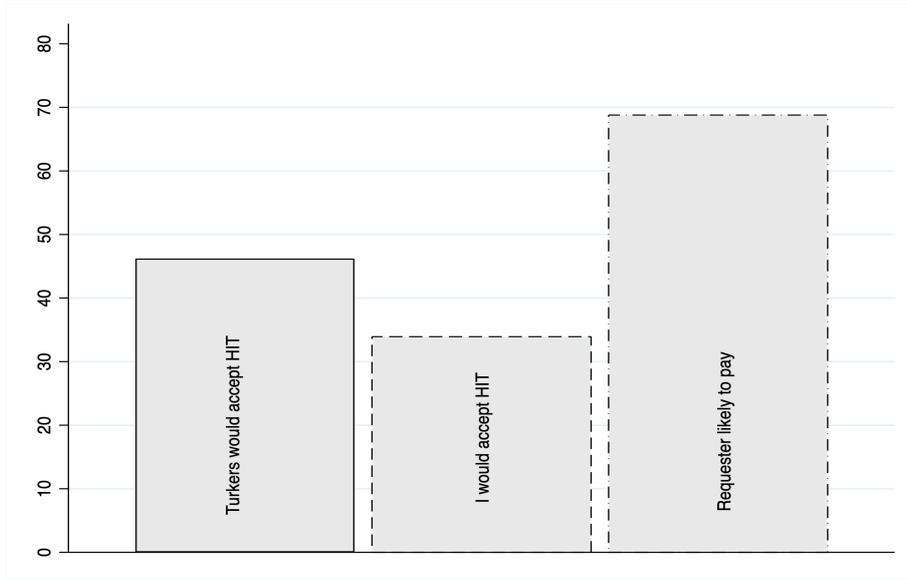


Figure 6: Follow-up Survey: Beliefs about Acceptance

Notes: Data from follow-up survey. Reported is subjects' belief about the percentage of other mTurkers who would accept the hypothetical task, the percentage of subjects who would accept the task themselves, and subjects' confidence level that the requester of the task would pay the stated bonus.

Table 1: Summary Statistics of Covariates

Variables	Black	White	No Pic	Female	Male	Total	ACS (2018)
Age	37.46	37.49	36.64	37.80	37.15	37.31	38.2
White	0.80	0.77	0.76	0.79	0.79	0.78	0.75
Female	0.50	0.49	0.43	0.51	0.48	0.48	0.51
Urban	0.50	0.54	0.49	0.53	0.51	0.52	0.80
Duration (mins)	7.11	7.48	7.68	7.25	7.34	7.37	
High School	0.11	0.09	0.11	0.08	0.11	0.10	0.27
Some College	0.25	0.22	0.23	0.25	0.22	0.23	0.20
2-Year College	0.11	0.12	0.10	0.12	0.12	0.11	0.09
B.Sc.	0.38	0.42	0.41	0.38	0.42	0.40	0.20
Graduate	0.16	0.15	0.16	0.18	0.13	0.16	0.13
N. Obs.	843	848	415	842	849	2106	

Notes: Reported is the mean of each variable by treatment group based on data from the real-effort experiment. We combine data by race and sex. 'No Pic' is the control group. ACS is American Community Survey. The ACS is an annual nationwide survey of approximately 3.5 million households.

Table 2: Treatment effect of employer race and sex

	Transcribe	N. of Pics	Accuracy	Transcribe	N. of Pics	Accuracy
	Panel A: Race			Panel B: Sex		
Constant	0.436*** (0.057)	12.442*** (2.005)	0.874*** (0.040)	0.429*** (0.056)	11.200*** (1.897)	0.815*** (0.039)
Treatment	-0.011 (0.023)	-0.993 (0.775)	-0.063*** (0.017)	0.003 (0.023)	1.461* (0.774)	0.044*** (0.017)
F	4.445	1.550	3.735	4.369	1.934	2.718
R2	0.021	0.022	0.045	0.021	0.025	0.034
N. Obs.	1,674	603	611	1,674	603	611

Notes: Reported is the race-gap and sex-gap for our three outcome variables based on data from the real-effort experiment. *Transcribe* is a binary variable equal to 1 if agreed to transcribe pictures and zero otherwise. *N. of Pics* is the number of pictures transcribed, and *Accuracy* is the share of accurate entries. The gaps are defined such that negative values imply minority groups have worse outcomes.

Table 3: Treatment effect of employer race and sex: within-group

	Transcribe	N. of Pics	Accuracy	Transcribe	N. of Pics	Accuracy
	Panel A: Race gap non-white workers			Panel C: Sex gap for female workers		
Constant	0.375*** (0.136)	15.895*** (4.418)	0.810*** (0.105)	0.488*** (0.083)	11.233*** (2.257)	0.755*** (0.052)
Treatment	0.008 (0.052)	1.947 (1.495)	-0.010 (0.040)	0.031 (0.034)	0.540 (1.001)	0.050** (0.023)
F	1.360	2.113	1.084	1.324	0.675	0.910
R2	0.028	0.108	0.056	0.012	0.016	0.019
N. Obs.	356	135	138	830	338	344
	Panel B: Race gap for white workers			Panel D: Sex gap for male workers		
Constant	0.449*** (0.062)	10.800*** (2.173)	0.867*** (0.042)	0.459*** (0.074)	11.351*** (3.097)	0.825*** (0.057)
Treatment	-0.014 (0.026)	-1.941** (0.891)	-0.078*** (0.019)	-0.019 (0.032)	2.647** (1.228)	0.039 (0.026)
F	4.482	1.441	4.132	2.752	1.959	1.308
R2	0.023	0.029	0.066	0.020	0.056	0.046
N. Obs.	1,318	468	473	844	265	267

Notes: Reported is the race-gap and sex-gap for our three outcome variables based on data from the real-effort experiment. *Transcribe* is a binary variable equal to 1 if agreed to transcribe pictures and zero otherwise. *N. of Pics* is the number of pictures transcribed, and *Accuracy* is the share of accurate entries. The gaps are defined such that negative values imply minority groups have worse outcomes.

Table 4: Treatment effect of employer race and sex: saliency

	Transcribe	N. of Pics	Accuracy	Transcribe	N. of Pics	Accuracy
	Panel A: Race gap all workers			Panel B: Sex gap for all workers		
Full	-0.011 (0.023)	-0.993 (0.775)	-0.063*** (0.017)	0.003 (0.023)	1.461* (0.774)	0.044*** (0.017)
	1,674	603	611	1,674	603	611
Salient	0.108*** (0.034)	-0.570 (1.063)	-0.059*** (0.020)	0.135*** (0.034)	1.468 (1.016)	0.056*** (0.019)
	995	410	418	907	385	392
Non-Salient	-0.043 (0.042)	0.505 (1.261)	-0.021 (0.044)	-0.093*** (0.034)	1.328 (1.112)	0.025 (0.033)
	679	193	193	767	218	219
	Panel C: Race gap Nonwhite workers			Panel D: Sex gap for Female workers		
Salient	0.165** (0.076)	3.552 (2.391)	0.009 (0.047)	0.136*** (0.048)	-0.074 (1.294)	0.053* (0.028)
	209	89	92	449	208	214
Non-Salient	-0.109 (0.084)	0.507 (1.190)	0.003 (0.077)	-0.048 (0.051)	1.594 (1.430)	0.032 (0.042)
	147	46	46	381	130	130
	Panel E: Race gap White workers			Panel F: Sex gap for Male workers		
Salient	0.094** (0.038)	-1.912 (1.168)	-0.077*** (0.022)	0.131*** (0.048)	3.437** (1.646)	0.066** (0.025)
	786	321	326	458	177	178
Non-Salient	-0.017 (0.049)	0.611 (1.672)	-0.031 (0.053)	-0.130*** (0.047)	0.760 (1.662)	0.005 (0.055)
	532	147	147	386	88	89

Notes: Reported is the race-gap and sex-gap for our three outcome variables based on data from the real-effort experiment. *Transcribe* is a binary variable equal to 1 if agreed to transcribe pictures and zero otherwise. *N. of Pics* is the number of pictures transcribed, and *Accuracy* is the share of accurate entries. *Full* is the full sample; *Salient* is the sample of subjects for whom race/sex was salient; *Non-Salient* is the sample of subjects for whom race/sex was not salient. The gaps are defined such that negative values imply minority groups have worse outcomes.

Table 5: Treatment effect of employer race and sex on perceived likelihood of being paid

	Race			Sex		
	Full	Non-White	White	Full	Female	Male
Panel A: Full Sample						
Constant	77.238*** (8.003)	62.216*** (12.176)	66.407*** (6.152)	77.137*** (8.146)	65.888*** (7.605)	63.389*** (12.952)
Treatment	-0.890 (1.998)	-2.771 (4.513)	-0.283 (2.218)	-0.316 (2.007)	-2.142 (3.059)	0.626 (2.669)
N. Obs.	765	180	585	765	348	417
Panel B: Salient Sample						
Constant	76.715*** (9.005)	88.995*** (12.764)	71.575*** (9.609)	83.580*** (7.620)	92.587*** (9.219)	68.359*** (12.531)
Treatment	-4.251 (3.039)	-4.094 (6.579)	-4.146 (3.570)	-0.238 (2.959)	0.944 (4.478)	-1.126 (4.001)
N. Obs.	454	108	346	392	176	216
Panel C: Non-Salient Sample						
Constant	69.575*** (8.140)	49.024** (21.116)	73.933*** (8.527)	62.585*** (7.501)	55.938*** (10.631)	63.498*** (11.115)
Treatment	-0.001 (3.663)	2.485 (9.122)	-1.685 (3.747)	-0.446 (2.903)	-6.032 (4.320)	3.163 (3.833)
N. Obs.	311	72	239	373	172	201

Notes: Reported is the race-gap and sex-gap in the perceived likelihood of being paid for a mTurk task based on data from a randomized survey experiment of subjects who did not participate in the real-effort experiment. Results are presented for the full sample as well as by worker type and saliency. *Full* refers to all workers; *Non-white* indicates sample of non-white workers only; *white* indicates sample of white workers only. *Female* indicates sample of female workers only; and *Male* indicates sample of male workers only. *Salient* is the sample of subjects for whom race/sex was salient; *Non-Salient* is the sample of subjects for whom race/sex was not salient. The gaps are defined such that negative values imply minority groups have worse outcomes.

(Online) Appendix

A Additional Details Survey Design

Would you like to transcribe my gasoline receipts (there is no penalty for opting out of this bonus task)?

Yes

No

CONTINUE

Figure A.1: Treatment Question

Notes: Reported is the question that subjects saw after the treatment instructions.

Purchase information

0 Gallons of gasoline 0 Price per gallon 0 Total sale

Receipt information

Store

Month

Day

Year

Submit Picture

Click if you want to transcribe more pictures

Click if you do NOT want to transcribe anymore pictures

CONTINUE

Figure A.2: Treatment Task

Notes: Reported is the data entry screen that subjects used when transcribing data from the gasoline receipts.

B Pilot Study, Balancedness, Additional (Summary) Statistics and Results

Table B.1: Pilot-experiment Results

Panel A: Reported Race					
Assigned Group	No Response	Black	White	Total	Race Salience
Black Female	2	29	4	35	83%
Black Male	4	18	7	29	62%
White Female	2	1	26	29	90%
White Male	4	0	24	28	86%

Panel B: Reported Sex					
Assigned Group	No Response	Female	Male	Total	Sex Salience
Black Female	6	29	0	35	83%
Black Male	6	1	22	29	75%
White Female	2	26	1	29	90%
White Male	14	7	7	28	25%

Notes: Reported is the race and sex perception of subjects. A subject's group assignment is indicated in the rows and her perceived group assignment is indicated in columns. Panel A reports results for perceived race and and Panel B reports results for perceived sex.

Table B.2: Summary Statistics by Treatment Group

Variables	BF	BM	Control	WF	WM	Total
Age	38.224	36.712	36.639	37.383	37.594	37.311
White	0.809	0.800	0.757	0.761	0.774	0.780
Female	0.516	0.491	0.434	0.496	0.478	0.483
Urban	0.536	0.470	0.491	0.530	0.551	0.516
Duration (mins)	6.793	7.429	7.684	7.712	7.252	7.374
High School	0.081	0.132	0.106	0.083	0.092	0.099
Some College	0.265	0.226	0.229	0.229	0.214	0.233
2-Year College	0.117	0.104	0.101	0.116	0.132	0.114
B.Sc.	0.346	0.408	0.407	0.407	0.426	0.399
Graduate	0.191	0.130	0.157	0.165	0.136	0.156
N. Obs.	419	424	415	423	425	2106

Notes: Reported is the mean of each variable by treatment group based on data from the real-effort experiment. BF is black female hand, BM is black male, WF is white female and WM is white male.

Table B.3: Balancedness test

Demographic	Black v Control		White vs Control		Female V Control		Male V Control		W vs B		F vs M	
	Difference	Pvalue	Difference	Pvalue	Difference	Pvalue	Difference	Pvalue	Pvalue	Pvalue	Pvalue	Pvalue
Age	0.825	0.378	0.850	0.270	1.163	0.120	0.515	0.664	0.834	0.203		
White	0.048	0.052	0.011	0.664	0.028	0.256	0.030	0.226	0.067	0.929		
Female	0.069	0.021	0.053	0.075	0.072	0.016	0.050	0.092	0.512	0.369		
Urban	0.012	0.702	0.049	0.104	0.041	0.170	0.019	0.522	0.126	0.365		
Duration (Mins)	-0.572	0.943	-0.203	0.920	-0.430	0.704	-0.344	0.685	0.780	0.340		
High School	0.001	0.968	-0.019	0.282	-0.024	0.161	0.006	0.754	0.176	0.038		
Some College	0.017	0.516	-0.007	0.773	0.018	0.480	-0.009	0.729	0.247	0.193		
2-Year College	0.009	0.623	0.023	0.239	0.015	0.421	0.017	0.381	0.388	0.929		
B.Sc.	-0.030	0.304	0.009	0.759	-0.031	0.293	0.010	0.742	0.101	0.089		
Graduate	0.004	0.873	-0.006	0.792	0.022	0.341	-0.024	0.259	0.602	0.011		

Notes: Reported is the difference in mean between treatment and control groups for each variable based on data from the real-effort experiment. Pvalues from a ranksum test of the differences in means between treatment groups are also reported. W vs B is white compared to black, and F vs M is female compared to male.

Table B.4: Characteristics of Salient and Non-salient samples

	Race		Sex		Full sample
	Non-Salient	Salient	Non-Salient	Salient	
Age	39.16	36.28	38.52	36.50	37.31
White	0.78	0.78	0.78	0.78	0.78
Female	0.49	0.48	0.49	0.48	0.48
Urban	0.50	0.53	0.51	0.52	0.52
Duration(mins)	6.42	7.91	6.40	8.03	7.37
High School	0.11	0.09	0.10	0.10	0.10
Some College	0.20	0.25	0.23	0.24	0.23
2-Year College	0.12	0.11	0.12	0.11	0.11
B.Sc.	0.39	0.40	0.37	0.42	0.40
Graduate	0.18	0.14	0.18	0.14	0.16

Notes: Reported is the mean of each variable for the salient and non-salient samples based on data from the real-effort experiment. A subject is in the race-salient sample if her self-reported perceived race treatment matches her assigned race treatment and the non-salient sample otherwise. A subject is in the sex-salient sample if her self-reported perceived sex treatment matches her assigned sex treatment and the non-salient sample otherwise.

Table B.5: Characteristics of Transcribers

	BF	BM	Control	WF	WM
Age	36.06	34.58	35.41	37.35	36.07
Female	0.59	0.58	0.50	0.58	0.47
White	0.80	0.77	0.75	0.78	0.74
High School	0.09	0.11	0.10	0.10	0.07
Some College	0.24	0.28	0.19	0.25	0.23
2-Year College	0.11	0.09	0.12	0.09	0.11
B.Sc.	0.35	0.41	0.42	0.42	0.43
Graduate	0.21	0.11	0.17	0.15	0.15
Urban	0.53	0.49	0.55	0.52	0.54
N. Obs.	148	150	144	158	152

Notes: Reported is the mean of each variable for subjects who transcribed at least 1 picture. BF includes subjects who were assigned to the black female hand treatment; BM includes subjects who were assigned to the black male hand treatment; Control includes subjects who were assigned to the control group; WF includes subjects who were assigned to the white female hand treatment; and WM includes subjects who were assigned to the white male hand treatment.

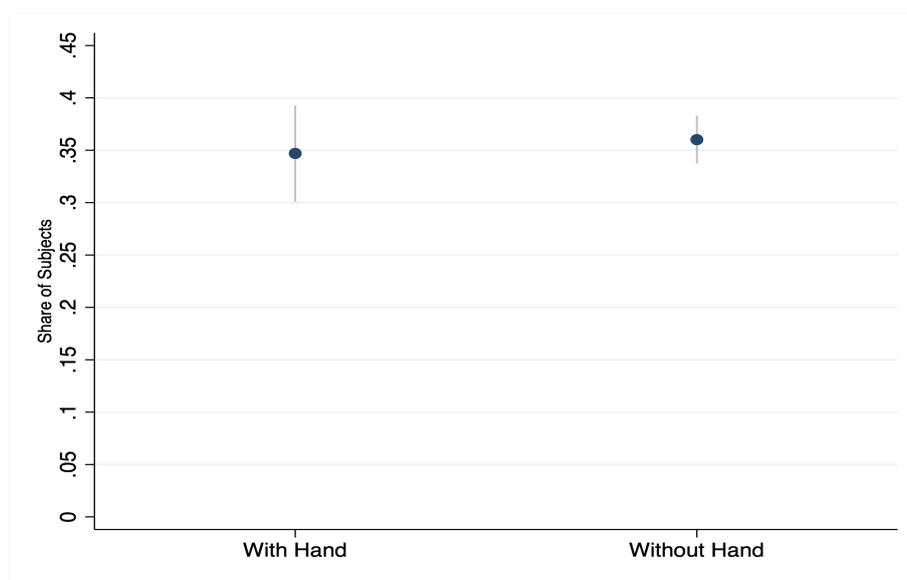


Figure B.1: Transcription share

Notes: Reported is the share of subjects who agreed to transcribe receipts. 'With Hand' indicates that the image of the receipt included a hand; 'Without Hand' indicates subjects in the control group.

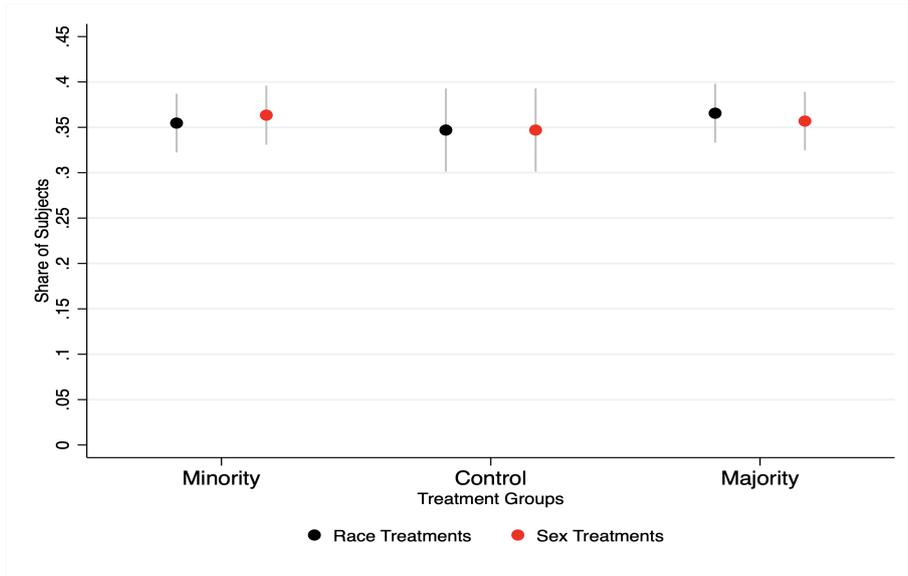


Figure B.2: Acceptance Share, by Treatment Group

Notes: Reported is the acceptance share by treatment group for race and sex, along with 95% confidence intervals. Acceptance share refers to the share of subjects who agreed to transcribe receipts. Minority refers to Black-hand and female-hand treatments, while majority refers to white-hand and male-hand treatments.

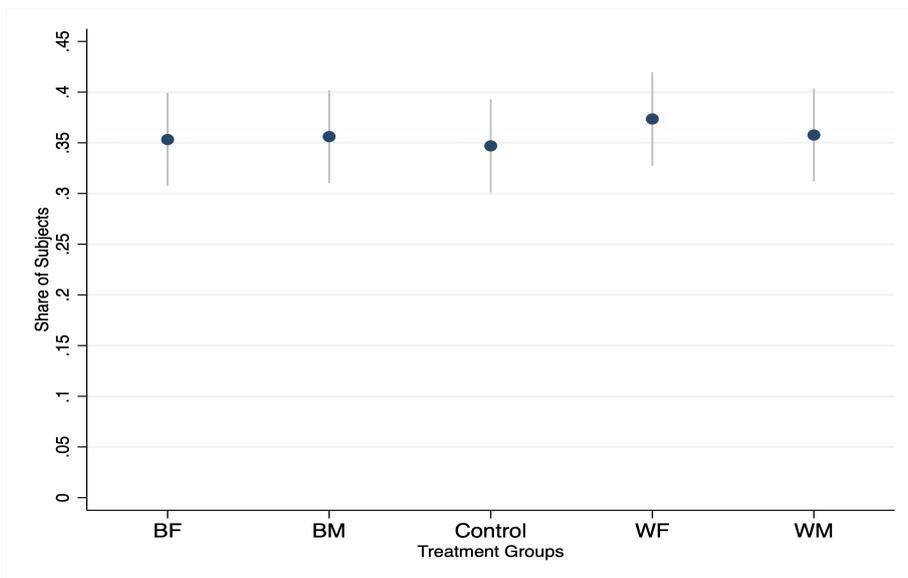


Figure B.3: Acceptance Share across all treatment groups

Notes: Reported is the share of subjects who agreed to transcribe receipts in each treatment group. BF includes subjects who were assigned to the black female hand treatment; BM includes subjects who were assigned to the black male hand treatment; Control includes subjects who were assigned to the control group; WF includes subjects who were assigned to the white female hand treatment; and WM includes subjects who were assigned to the white male hand treatment.

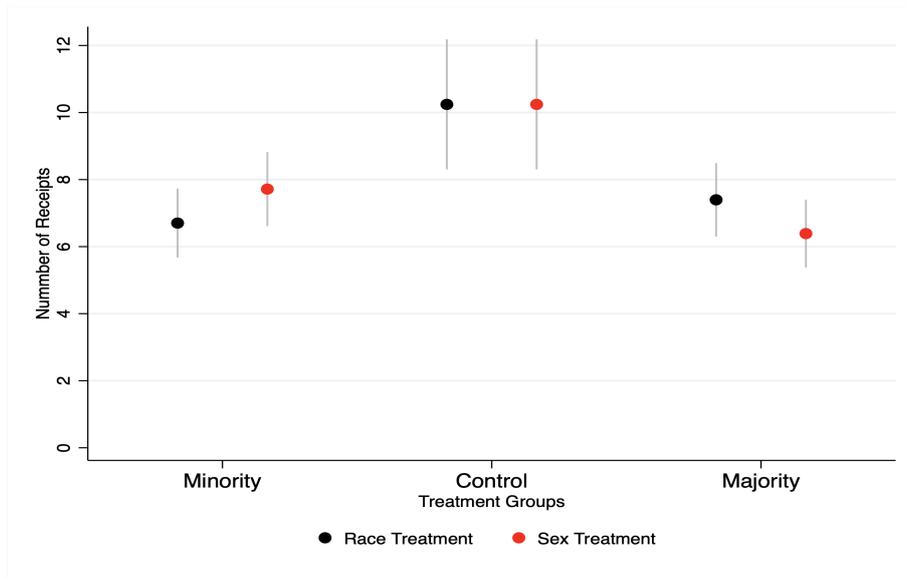


Figure B.4: Mean Number of Transcribed Receipts, by Treatment Group

Notes: Reported is the mean number of receipts transcribed by treatment group for race and sex, along with 95% confidence intervals. Minority refers to Black-hand and female-hand treatments, while majority refers to white-hand and male-hand treatments.

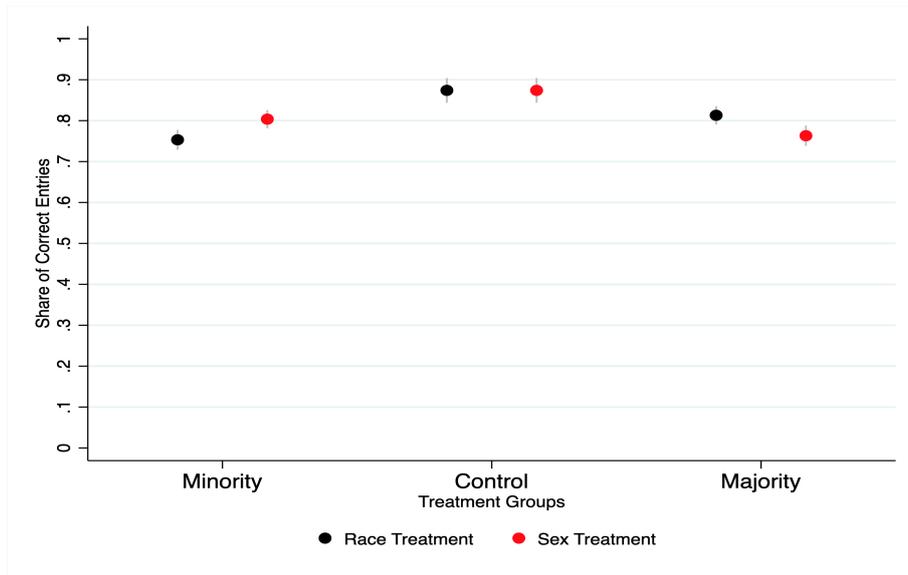


Figure B.5: Accuracy Rate, by Treatment Group

Notes: Reported is the share of accurate transcriptions across treatment groups, along with 95% confidence intervals. Minority refers to Black-hand and female-hand treatments, while majority refers to white-hand and male-hand treatments.

C Salience Across Treatment Groups

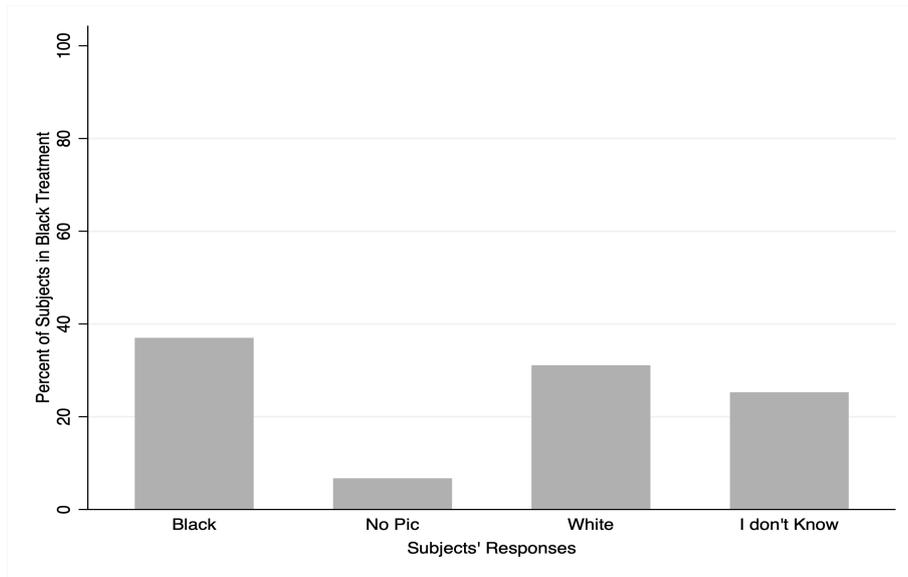


Figure C.1: Race Salience in the Black Treatment Group

Notes: Reported is the percent of subjects in the black treatment who gave each of the possible responses to the question: "What is the race of the person holding the receipt in the picture?".

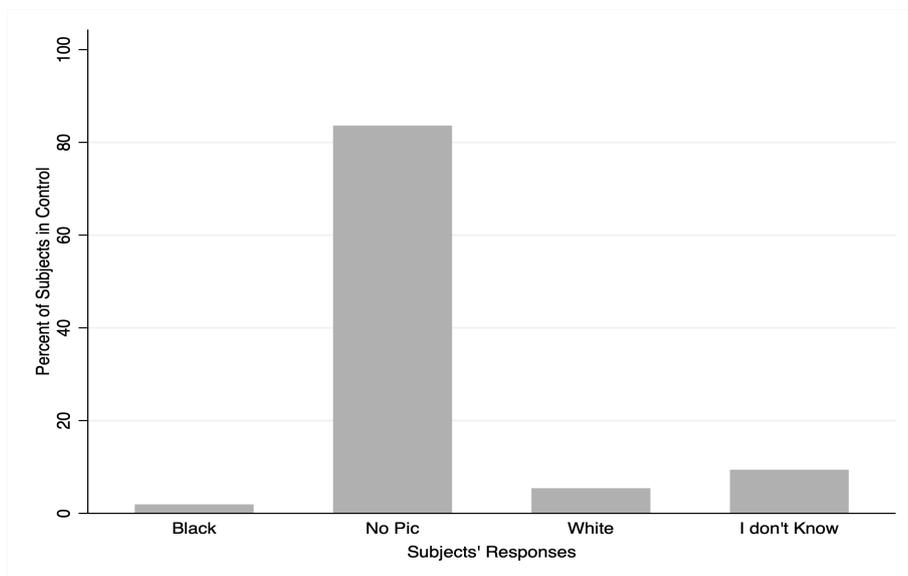


Figure C.2: Race Salience in the Control Group

Notes: Reported is the percent of subjects in the control group who gave each of the possible responses to the question: "What is the race of the person holding the receipt in the picture?".

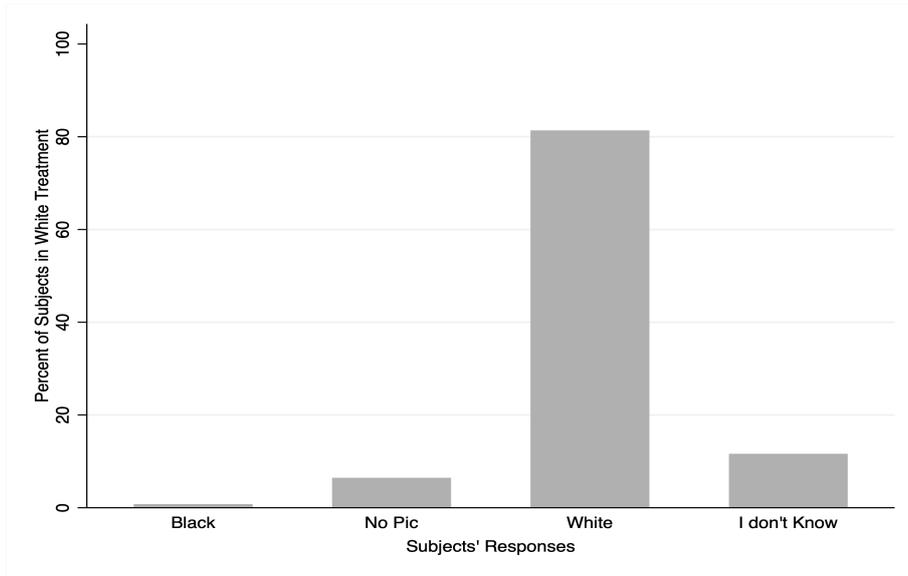


Figure C.3: Race Salience in the White Treatment Group

Notes: Reported is the percent of subjects in the white treatment who gave each of the possible responses to the question: "What is the race of the person holding the receipt in the picture?".

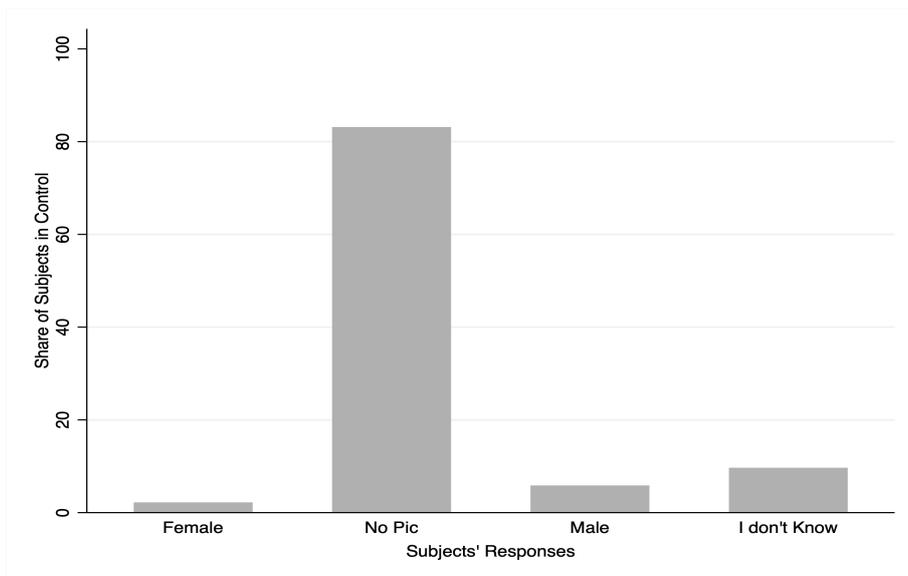


Figure C.4: Sex Salience in the Control Group

Notes: Reported is the percent of subjects in the control group who gave each of the possible responses to the question: "What is the sex of the person holding the receipt in the picture?".

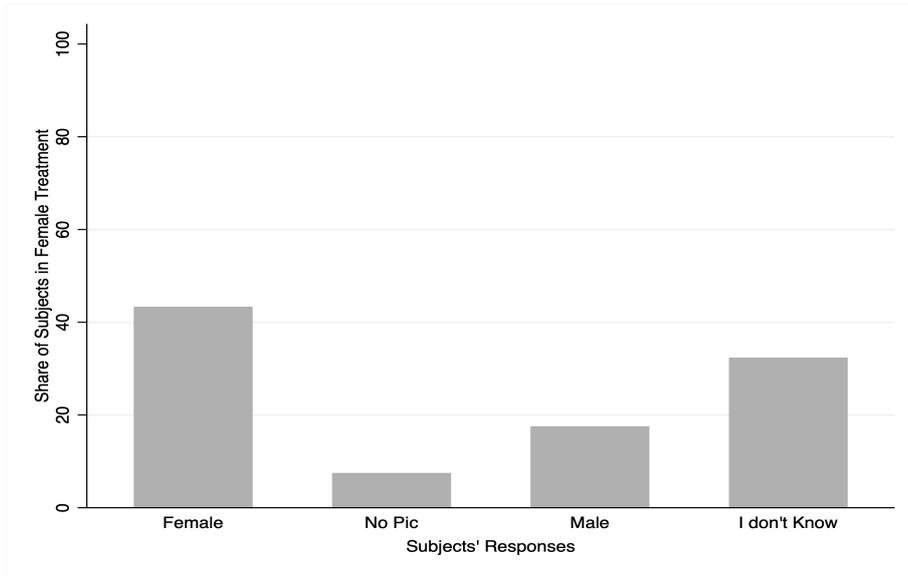


Figure C.5: Sex Salience in the Female Treatment Group

Notes: Reported is the percent of subjects in the female treatment group who gave each of the possible responses to the question: “What is the sex of the person holding the receipt in the picture?”.

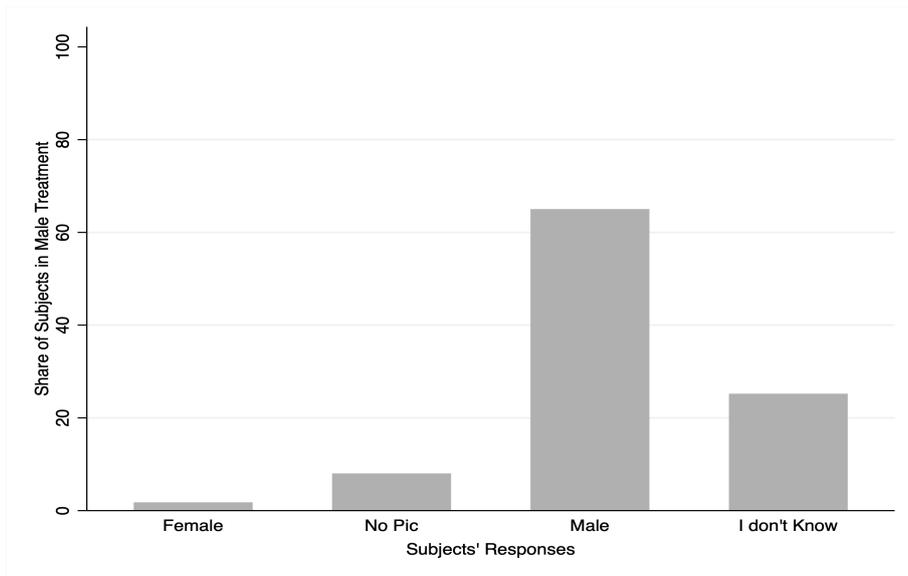


Figure C.6: Sex Salience in the Male Treatment Group

Notes: Reported is the percent of subjects in the male treatment group who gave each of the possible responses to the question: “What is the sex of the person holding the receipt in the picture?”.

D Follow-up Survey: Additional Statistics

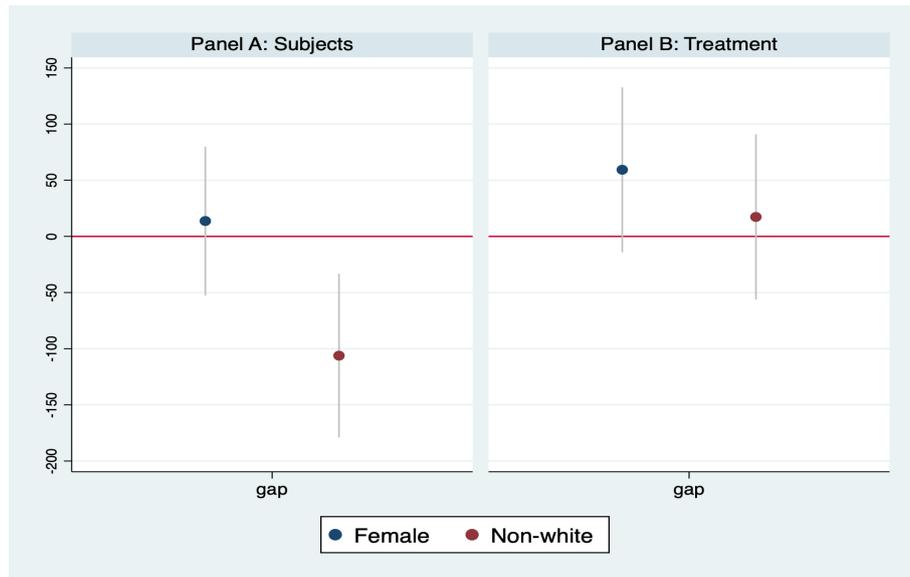


Figure D.1: Follow-up Survey: mTurker Work Experience Survey: Monthly HITs

Notes: Reported is the gap in the mean number of monthly HITs completed by subjects along with 95% confidence intervals. Differences are calculated between female and male subjects, and between non-white and white subjects in Panel A. Differences are calculated between female and male treatments and between black and white treatments in Panel B.

Table D.1: Follow-up Survey: Balancedness test

Variables	Male v Control	Female control	White vs Control	Black vs Control	Female v Male	White vs Black
18 to 24	0.411	0.598	0.807	0.993	0.095	0.753
25 to 34	0.425	0.386	0.750	0.787	0.038	0.463
35 to 44	0.426	0.513	0.748	0.251	0.862	0.308
45 to 54	0.343	0.281	0.395	0.240	0.864	0.670
55 to 64	0.569	0.896	0.973	0.697	0.390	0.661
>65	0.762	0.228	0.313	0.553	0.197	0.542
White	0.246	0.425	0.302	0.355	0.660	0.899
Female	0.592	0.468	0.701	0.378	0.812	0.534
Duration(mins)	0.712	0.885	0.949	0.870	0.537	0.830
High School	0.994	0.416	0.502	0.874	0.306	0.516
Some College	0.816	0.956	0.564	0.464	0.720	0.102
2-Year College	0.603	0.912	0.674	0.830	0.606	0.796
B.Sc.	0.525	0.468	0.509	0.483	0.910	0.957
Graduate	0.915	0.809	0.553	0.782	0.865	0.276

Notes: Reported is the difference in mean between treatment and control groups for each variable based on data from the follow-up survey. P-values from a ranksum test of the differences in means between treatment groups are also reported. W vs B is white compared to black, and F vs M is female compared to male.

Table D.2: Salience of Treatment in Experiment and Survey

Treatment	Race Salient		Sex Salient	
	Experiment	Survey	Experiment	Survey
BF	0.43	0.41	0.39	0.41
	419	185	419	185
BM	0.31	0.32	0.62	0.55
	424	192	424	192
Control	0.84	0.88	0.83	0.88
	415	190	415	190
WF	0.84	0.84	0.47	0.47
	423	200	423	200
WM	0.79	0.78	0.68	0.62
	425	188	425	188
Total	0.64	0.65	0.60	0.59
	2106	955	2106	955

Notes: Reported is the share of subjects whose self-reported perceived treatment matches the actual treatment they are assigned for race and sex, respectively, in the real-effort experiment and the follow-up survey. A subject is in the race-salient sample if her self-reported perceived race treatment matches her assigned race treatment and the non-salient sample otherwise. A subject is in the sex-salient sample if her self-reported perceived sex treatment matches her assigned sex treatment and the non-salient sample otherwise. BF includes subjects who were assigned to the black female hand treatment; BM includes subjects who were assigned to the black male hand treatment; Control includes subjects who were assigned to the control group; WF includes subjects who were assigned to the white female hand treatment; and WM includes subjects who were assigned to the white male hand treatment.