The (In)visible Hand: Do Workers Discriminate Against Employers? *

Philipp Doerrenberg Denvil Duncan Danyang Li

August 30, 2022

Abstract

Although a large literature has studied discrimination in the labor market, there is little 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 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. This study is registered in the AEA RCT Registry and the unique identifying number is: AEARCTR-0005792. IRB approvals were obtained from Indiana University and Hofstra University. The authors do not have any competing interests to declare.

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 2022; Abel and Buchman 2020). 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 relevant in various 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. But employee-to-employer discrimination is also relevant in traditional labor markets and in contexts where individuals

with specialized skills are able to select their employer.[1] 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 2345 subjects in an online labor market (Amazon Mechanical Turk, mTurk).[2] We choose this online labor market for a couple of reasons. First, it is characterized by a high degree of anonymity, which allows us to inject one non-anonymous component (race/sex) and thus isolate the discrimination effect in the absence of any confounding factors. Second, it is an environment with very little repeated interactions. These two features are different from more traditional labor markets, but should work in our favor. For example, we would expect larger effects in settings where workers and employers know each other and have repeated interactions. The reason is that workers know of the repeated interactions before they apply for a job. Consequently, we suspect that workers for whom an employer's race or sex is a problem are less likely to take a job under repeated interactions compared to one-time interactions. This difference in the likelihood of taking the job should result in larger effects.

The subjects in our experiment are invited to complete a survey about mileage taxes. 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).[3] 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 interactions, sympathy, or trustworthiness.[4]

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

[1] Data from various sources including the Gig Economy Data Hub, the International Labor Organization's (ILO), the Trades Union Congress, and the iLabor project at Oxford all point to significant growth in the size and scope of the gig-economy. Further, the U.S. Small Business Administration office of Advocacy estimates that small business in the US that account for 40% of private sector payroll. We provide more details on the size of the gig economy and the traditional labor market in Section 1.1 below.

[2] mTurk is an established online labor-market platform and a big player in the gig-economy. For example, Robinson et al. (2019) estimates that there are over 225,000 mTurkers in the US over the 2016-2019 period with over 50,000 new workers joining each year. Importantly, Robinson's estimates represent a lower bound on the number of workers on the mTurk platform used in our study.

[3] The 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.

[4] We 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.

or not,[5] and an *intensive margin*, the amount and quality of work conditional on accepting the task.

We find the following main 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 for black employers, relative to white employers, and more accurate for female employers, relative to male employers. We also find that workers transcribe more receipts for female employers and fewer receipts for black employers. However, these latter 'quantity' results are not robust to multiple hypothesis adjustments.

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.

To advance the understanding of discrimination behavior on the extensive margin, we analyze 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. Considering that our pilot study suggests that our treatment pictures were salient, this pattern raises the possibility that subjects misreport self-reported saliency in the post-experimental survey to hide their discriminating behavior.

We explore this potential explanation of our findings by examining the relationship between the transcription decision and self-reported saliency in two ways. First, we estimate the correlation between self-reported saliency and the transcription decision in each treatment separately. Our results indicate that saliency is strongly correlated with the transcription decision in the minority treatments only. Second, we estimate the extensive-margin sex and race gaps separately for the salient and nonsalient samples. In the absence of any strategic misreporting, we should observe zero effects for nonsalient subjects. However, this is not what we find: the results show 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

[5] We 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.

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]

Taken together, the results from our analysis of the self-reported saliency measure suggest that some workers strategically misreported their responses to questions about race and sex of the employer in the post-experimental survey. This finding is consistent with the literature on dishonesty which generally finds that dishonest people are interested in appearing honest (Mazar et al. 2008). Note that we do not view this exercise as a conventional heterogeneity analysis where one studies whether treatment effects vary depending on pre-defined characteristics. Instead, we embrace the endogeneity of self-reported treatment salience to improve our understanding of discrimination in our context.

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. 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 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.

[6]In 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).

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 2022; 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, there are at least two segments of the labor market where workers are able to discriminate against employers. First, the traditional labor market includes many small businesses where workers can easily identify their 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. This segment of the traditional labor market is quite large and has a non-trivial share of female and nonwhite owners. According to the US Small Business Administration Office of Advocacy, there was an estimated 32 million small businesses in the US in 2020 and 19% of them have paid employees which accounts for 40% of private sector payroll.[9] Most relevant for us is the fact that about 25% of small businesses that employ workers are home-based, which suggests they are small enough for employees to easily identify their employer. Additionally, over 19% and 17% of small businesses that have at least one employee have a female or nonwhite owner, respectively, in 2017.

Second, the opportunity for workers to discriminate against employers has increased significantly with the rise of the 'gig' economy in the last decade. While difficult to measure due to variation in definition, there is a lot of very strong suggestive evidence of a large and rapidly growing gig economy (Farrell and Greig 2016; Farrell et al. 2018; Katz and Krueger 2019). For example, survey-based estimates of the size of the gig-economy range from 0.5% - 1.1% and 5% - 15% of the adult population for the USA and Europe, respectively, depending on the reference period used in the survey (International Labour Office 2021). Using data directly from

[7] Our 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.

[8] Of 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.

[9] Source: U.S. Small Business Administration, [Last Accessed March 06, 2022].

online labor platforms, International Labour Office (2021) reports the number of registered and active workers across five platforms to be 42,781 (99designs), 95,600 (workana), 95,813 (Freelancer), 126,475 (PeoplePerHour), and 1,048,575 (Guru). A similar report published by the Trades Union Congress estimates that approximately 12 percent of the working population in England and Wales carried out remote online digital tasks such as ours in 2021; this is up from 4.9 percent in 2016 (Trade Union Congress 2021). Both of these segments of the labor market provide significant opportunities for workers to discriminate against employers and our work is among the first to explore the extent to which these opportunities are exploited.

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. 2021). 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). Also using a gig market environment, a related working paper by Asad et al. (2020) explores altruism and reciprocity as motivations for why *white* employees do not discriminate against their black employers.

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 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 tax survey for a flat fee of \$0.65.[10] 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 tax 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 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 tax 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).[11] See Figure 2 for a screen shot of the details shown to subjects at the time they receive treatment.

[10] The results of this survey are used in an unrelated paper about road mileage tax (Duncan et al. 2020).

[11] Our set-up does not allow us to measure the time per picture directly. However, we estimate the time per picture by fitting the following regression: $d_i = \alpha + \beta pics_i + \delta X_i + \epsilon_i$, d is total time spent in the experiment by subjects who transcribed at least 1 picture, pics is number of pictures transcribed, and X is a vector of covariates. The estimated time per pic is 39 seconds when we restrict the sample to those subjects in the 5th to 95th percentile of the duration distribution.

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 subjects 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.[12] 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. Accuracy 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 fist 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.

Post-Treatment Survey and Treatment Saliency. We run a post-experimental survey in which we collect data on whether subjects discern the race and sex of the person in the picture 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?

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

[12] Extensive-margin in our context refers solely to the decision to transcribe gasoline receipts or not; see footnote 5 above in which we acknowledge this point.

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 and race non-salient otherwise; i.e., she is race salient if her perceived race-treatment matches her actual race-treatment.

Subjects did not know that they would be asked the saliency questions at the time when they made their labor supply decisions. Therefore, it is possible that having decided against working for a specific type of employer, a subject might try to conceal the role of race or sex in this labor supply decision by misreporting her treatment. Such misreporting could potentially affect the observed race and sex gaps. Consequently, the treatment-salient measures are used to investigate the extent to which subjects intentionally misreport their perceived treatment. 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, 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's 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. Alex is a very common diminutive for the male name Alexander, and female names such as

Alexandria, Alexa, Alexia and other variations of these names. All of these variants are highly ranked names in the social security name database. For example, we find that there were more female than male versions of names for which Alex is the diminutive in each of the years 1980, 1990, 2000, and 2010, in the social security's Popular Names database. Additionally, the total number of babies with these Alex-type names is very similar between sex.[13] It is more difficult to find hard data on names by race. However, a google image search for the name Alex uncovers males and females of color. Therefore, we argue that the name 'Alex Wright' should minimize the likelihood that subjects select the HIT based on the requesters race or sex. Importantly, 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, which suggests that our pool of subjects is not influenced by the name of the requester.[14]

Second, we want to make sure subjects receive the signal we intended to send via the hands in the pictures. Recall that we signal race by skin color and sex by presence of nail polish. One of the underlying assumptions here is that nail polish signals sex and nothing else (e.g., a woman's social status). We argue that the nail polish color we use contains very little information beyond the sex of the hand with the nail polish. The reason is that $85 - 90\%$ of women use nail-care products across the world (Goldstein Market Intelligence 2020). Furthermore, there were 395,658 licensed nail technicians, over 54,000 nail salons, and the average price of a basic manicure was approximately $21 in the US in 2018 (Nail Magazine 2020). In other words, nail salons are so pervasive, basic manicure so inexpensive, and nail polish use so widespread that it is plausible to assume that a hand with a common nail polish color signals sex and nothing more.[15]

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 C.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, using the self-reported measure of saliency, we find that the race and sex of the white-male hand used in the actual experiment was identified by 79% and 69% of the subjects,

[13]Source: Social Security Popular Names Database[Accessed November 11, 2021].

[14]On 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).

[15]Of course, there are types of nail-care services that would arguably signal more than sex. For example, nail extensions, acrylic, french tips, and other forms of nail art most likely signals more than sex. However, these kinds of more elaborate and expensive nail care were not used in our setting.

respectively (see further below). 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.[16]

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. 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 (employer). We attempt to make it clear to the subjects that the hand in the picture is that of the employer 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 in the picture to the employer; 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"*.[17]

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 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 an 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

[16]Note 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.

[17]While we believe this framing of the treatment makes a clear connection between the hand subjects see in the treatment and the employer, we cannot rule out the possibility that some subjects fail to make that connection. Even so, the percentage of subjects who fail to make the connection should be balanced across treatments.

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.[18] 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).[19]

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).

[18] For example, a simple search for transcription tasks on the platform at 7PM on September 14, 2021 yielded 475 text-transcription tasks like ours and 55 video/audio transcription tasks.

[19] Kuziemko et al. (2015) and DellaVigna and Pope (2018) are recent examples of economics papers using Amazon's Mechanical Turk.

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.[20] We cleaned the data in the following ways before performing our empirical analysis. First, we drop 76 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. Second, we drop 86 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 2345 total observations; approximately 470 subjects per treatment. Importantly, these adjustments were equally distributed across treatment groups (see Table C.2 in online appendix).

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 C.3), 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 present p-values from a ranksum test of the null hypothesis that the mean for each demographic variable is the same in the last two columns of Table 1.[21] Except for age and education, there is no statistically significant difference between the female and male treatments. We find that, relative to the male treatment, the female treatment is approximately 1 year older and has 2 percentage points more subjects with a Graduate degree; $p - value = 0.011$, respectively. Similarly, subjects' race is the only statistically significant difference between the black and white treatment groups; 4 percentage points more non-white subjects in the white treatment relative to the black treatment. The black treatment also has 4 percentage point fewer subjects with B.Sc. We control for these variables in the empirical analysis and find that they do not change our results.

[20] As 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.

[21] The 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.

3 Empirical Strategy and Results

This section describes our empirical strategy and results. We first describe the empirical strategy. Following that is a discussion of the results on both the extensive-margin (decision to accept bonus task in Section 3.2) and the intensive-margin (number of receipts transcribed and accuracy in Section 3.3).

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:

$$y_i = \alpha + \beta Treatment_i + \delta X_i + \epsilon_i, \tag{1}$$

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 Section 2.1). X is a vector of subject-level covariates including age, sex, race, education and urban, and ϵ_i is a standard error term.

When we estimate the race-gap, $Treatment$ 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 estimate the sex-gap, $Treatment$ 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.[22] 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.

3.2 Labor Supply Gaps on the Extensive Margin

3.2.1 Main Result

Figure 5 reports the overall acceptance rate across race and sex groups. The figure shows that the mean acceptance rate was approximately 36% across treatment groups.[23] 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 results presented in the first column of Panels A and B for race and sex, respectively, are practically zero, which indicates that, on average, subjects were equally likely to work for

[22] We 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 $\beta = \beta_b - \beta_w$.

[23] Figure C.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 C.2 for detailed results across treatment groups.

black (female) employers as white (male) employers.

We also estimate 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 gap estimates among the nonwhite workers. The within-group results presented in Table 2 show that the gaps are statistically zero both within and between groups.[24]

3.2.2 Self-reported Saliency: Do People Try to Conceal their Discrimination Behavior?

To shed more light on discrimination behavior on the extensive margin, we leverage responses to a post-experimental survey that asks subjects about the sex and race of the person holding the receipt in the treatment photograph. We classify a subject as race-salient if her *self-reported* race treatment matches her actual race treatment. Sex salient subjects are defined similarly. Motivated by a literature showing that dishonest people are interested in appearing honest (Mazar et al. 2008), we use these post-experimental measures of self-reported saliency to examine if people stand up to their potential discrimination behavior. If, for example, we see non-zero effects for nonsalient subjects, this might be an indication for strategic misreporting to conceal discriminating behavior.

Self-reported Salience by Treatment Group. 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. The fact that self-reported saliency differs so wildly between the minority and majority groups is indicative of misreporting. This is especially suspicious given the high level of saliency for both groups in the pilot study.[25]

[24] These extensive margin results are unchanged when we adjust for multiple hypothesis testing. See Panels A, D, and E of Table B.1 and the discussion in Appendix B.

[25] Results presented in Figures D.1 to D.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 know the race of the hand in the picture. On the other hand, over 80% of subjects in the

Regression Analysis to Shed Light on Potential Misreporting. Considering that the signals in Figure 3 are clear, as suggested by results from our pilot study, there are at least two possible explanations for the observed difference in subjects' perception of the employer's race and sex. It could be that subjects received and responded to the race/sex signal consciously or subconsciously, but ex-post misreported the race and sex in an effort to conceal their biases. Alternatively, 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.

We explore these two possibilities by estimating equation (1) separately for salient and non salient subjects. Notice that our question is not: do subjects respond one way when treatment is salient and a different way when treatment is non-salient. This type of question would require an exogenous measure of saliency, which we do not have. Instead, we want to know if subjects manipulate their self-reported saliency to mask any discrimination behavior implied by their labor supply choices. We address this question by estimating the correlation between the transcription decision and the self-reported saliency. A correlation different from zero would be highly suggestive of misreporting among subjects.

We find strong indications for this type of manipulation. First, we find that subjects in the minority treatments (female and black) who transcribed images are more likely to be classified as salient (see Table C.6). Now, one might argue that this is driven by the fact that transcribers spend more time with the receipts. However, we do not believe this is an adequate explanation because we find no evidence of a correlation between saliency and transcribing among subjects in the majority treatments (male and white). The fact that transcription is correlated with saliency in the minority treatments and not the majority treatments is further suggestive of strategic misreporting.

Second, we check if the transcription gaps are correlated with saliency. If nonsalient subjects are really manipulating their self-reported saliency then we would expect the transcription gap among nonsalient subjects to be different from zero. This is precisely what we find and report in Table 3. The estimated race gap is approximately 11 percentage points among subjects who correctly reported the race of their treatment group. Comparable estimate for the sex gap is 15 percentage points. These findings are true both for the full sample of workers as well as minority and majority group workers.

Of greater interest is that we find negative and statistically significant gaps among the non-salient subjects. Subjects for whom race was not salient were approximately 5 percentage points more likely to transcribe for white than black employers (although not statistically different from zero). We also find that non-salient subjects are almost 10 percentage points more likely to transcribe for male employers than female employers (statistically significant). Interestingly, this effect is driven by males workers who are 14 percentage points less likely to work for

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. The summary statistics in Table C.5 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.

females. A simple chi-squared test confirms that the gaps among self-reported salient subjects are statistically different from those in the non-salient sample. We further confirm that the gaps are different between these two groups in regression specifications that pool all observations and include an interaction between treatment and salience (results available upon request).[26] Importantly, the correlation between saliency and the sex/race gaps is robust to adjustments for multiple hypothesis testing; see Panels B and C of Table B.1 and discussion in Appendix B.

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. In particular, if 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 either of the saliency 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.

3.3 Labor Supply Gaps on the Intensive Margin

This section explores the intensive-margin labor supply decisions: number of pictures transcribed and accuracy.

Potential Selection Problem. The intensive-margin decisions are only observed for subjects who decide to accept the transcription task. Therefore, any observed race or sex gap is conceptually driven by the randomly assigned treatment *and* selection (i.e., incidental truncation). In particular, it is possible that the subjects who accept the task in the black-hand treatment differ from those who accept the task in the white-hand treatment in ways that also influence effort and accuracy.

While we cannot fully rule out selection as a driver of our intensive-margin results, we have two reasons to believe selection does not pose a problem in our setting. First, we check for the severity of selection by comparing the characteristics of subjects who accepted the task across treatment groups. The results presented in Table C.7 indicate that the subjects who select to transcribe images are similar on observables across treatment groups. Second, while we cannot rule out the existence of differences between groups in unobservables, we suspect the only unobservable that matters is a worker's underlying preference toward race and sex.[27] Furthermore, this particular unobservable would bias our estimates toward zero. Let's consider an extreme example. Suppose there are two types of workers: biased and unbiased. Also suppose that only unbiased workers transcribe for black employers while both biased

[26] Notice that the estimated effects in Table 2 are not equal to the weighted average of the estimated effects in the salient and non-salient samples from Table 3. This is because treatment is correlated with the decision to transcribe and the self-reported saliency.

[27] Overall, the only kinds of unobservables that matter are the ones that influences the transcription decision differentially across treatment. We cannot think of any other factor that would fit this bill except a worker's bias toward race and sex. A person will accept a task after seeing treatment if they have the time, have the skill, need the money, have no outside option, etc. But all of these should be balanced across groups. More importantly, there is no reason to think that these characteristics differentially affect the decision to transcribe for black vs white hand.

and unbiased workers transcribe for white employers. This type of selection would inflate the average number and accuracy of transcribed images in the black treatment relative to the white treatment, which in turn would lower the estimated treatment effect. Therefore, if workers' bias is the primary unobservable characteristic driving the decision to transcribe, then our estimated intensive margin effects would be lower bounds. The saliency results described in Section 3.2.2 are supportive of the assumption that the decision to transcribe is driven by subjects' biases. Consequently, we argue that the selection bias is most likely pushing us toward lower-bound estimates on the intensive margins.

3.3.1 Main Result

Summary Statistics. Approximately 37% (or 857) of subjects transcribed at least 1 receipt, and subjects transcribed an average of 10 receipts. However, the distribution is highly skewed; the median number of transcribed receipts is 4, 75% of subjects transcribed fewer than 13 receipts, 90% transcribed fewer than 40 and only 95 subjects transcribed all 40 receipts. Figure C.3 shows how the mean number of transcribed receipts varies across treatment groups. Subjects in the control group transcribed 13.8 receipts on average. While subjects in the treatment groups transcribed fewer receipts than those in the control group, the reduction appears to be larger for black and male employers compared to white and female employers, respectively. A similar pattern is observed for accuracy in Figure C.4. We find that subjects in the control group got about 86% 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.

Main Effects. We estimate intensive margin effects using equation 1 with 'number of transcribed receipts' and accuracy as the outcome variables. The results are presented in Table 4. We find that workers transcribed about 1.8 ($p-value = 0.057$) fewer pictures and were 6 percentage points ($p-value < 0.001$) less accurate when working for a black employer. At the same time, we find that subjects exerted more effort for female employers relative to male employers: 1.6 ($p-value = 0.094$) additional pictures with 4 percentage points ($p-value < 0.001$) greater accuracy. While the accuracy results survive multiple hypothesis tests, the 'num of transcribed pics' results do not. Therefore, we caution the reader when interpreting the 'num of transcribed pics' results (see Appendix B). Table 4 also shows some interesting within group dynamics on the intensive margin. First, while white workers transcribed 3 fewer images and were 8 percentage points less accurate when working for black employers relative to white employers, nonwhite workers treat both types of employers the same.[28] This is suggestive of out-group bias for the intensive-margin race gaps, a finding that survives multiple hypothesis testing. Second, we find that both male and female workers exert greater effort when working for a female employer, though the estimates are not all statistically different from zero. Female workers are about 5 percentage points more accurate when working for a female employer but treat both male and

[28] We would like to caution the reader on the interpretation of the non-white group estimates since the sample of non-white workers is small.

female employers equally with respect to number of transcriptions. Male workers transcribe 2.7 additional pictures and are about 3 percentage points more accurate when working for female employers. Though economically important, the within-worker sex results are not robust to multiple hypothesis testing.

The intensive-margin results appear to be much stronger and consistent for accuracy compared to number of images transcribed. This is not entirely surprising given the reward structure of our experiment. The piece rate design used in our study compensates workers for the quantity rather than quality of transcriptions. Consequently, conditional on working for an employer, it is optimal for a worker to discriminate against the employer on the quality dimension relative to the quantity dimension (Holmstrom 1991). This prediction is largely borne out in our results. Although we find evidence of race and sex gaps in quantity of images transcribed, we find much stronger effects on the accuracy (quality) margin.

Self-Reported Salience. Similar to our extensive-margin analysis presented above, we explore the correlation between self-reported salience and the intensive-margin outcomes in Tables 5 and 6 for number of transcriptions and accuracy, respectively. We find that transcriptions among the non-salient group are in line with expectations; the estimated gap among all workers is practically zero for both race and sex (see first column of Table 5). However, we do find that workers for whom the sex-treatment was salient transcribed approximately 2 more pictures for female employers compared to male employers. Although this estimate is economically meaningful, it is not statistically different from zero. The comparable gap for race is practically zero.

The accuracy results in Table 6 show that treatment-salient workers are significantly more accurate for female employers and less accurate for black employers. The gaps among non-salient workers are both smaller and not statistically different from zero. As with the other intensive-margin response (transcribed pictures), we thus find no evidence of treatment effects among treatment-nonsalient workers. These saliency results are robust to multiple hypothesis testing; see Panels B and C of Table B.1 and discussed in Appendix B.

3.3.2 Further Indications for Misreporting

The extensive margin results described in Section 3.2.1 suggest that subjects misreported their treatment assignment to conceal their transcription decisions. 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 and is further evidence that some subjects misreported the treatment saliency to conceal their extensive margin responses.

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.

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 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.[29] 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.

[29] Although 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.

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 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 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 E.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 self-reported salience of treatment is identical between the real-effort experiment and follow-up survey samples (see Table E.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 E.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 E.1).

Experience with Employers/Requesters and Non-payments. Figure 6 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 6 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% of subjects report that that they check the names of requester, and 71% 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 7. 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 was about 10 percentage points higher than what mTurkers actually did in the real-effort experiment, 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 7 shows that the sex of the requester has no bearing on workers' uncertainty about payment; estimates are both small and statistically indistinguishable from zero. Similarly, there is no statistical or economic evidence that the race of the requester affects the uncertainty of being paid. Table 7 also shows no evidence of within-group differences in the perceived likelihood of being paid. Although the gaps are larger among non-white and female workers relative to white and male workers, respectively, they are all statistically indistinguishable from zero.

The fact that mTurkers' perceptions of employers likelihood to pay are uncorrelated with race or sex of the employer suggest that the gaps observed in our real-effort experiment are not driven by statistical discrimination.

One concern with this approach to uncovering the source of the bias is social desirability bias. In other words, it might be that subjects misstate their responses to avoid being viewed negatively. However, we argue that this is unlikely to be present in our survey for a couple of reasons. First, subjects are presented with a hypothetical scenario so subjects know their responses are not directly impacting anyone. Second, we asked subjects to comment on what they believe other mTurkers would do in the hypothetical scenario. While a subject might lie about her own behavior, we suspect that are more honest when commenting on what other people are likely to do. Third, we observe that subjects reported a higher acceptance rate for other mTurkers than themselves. If subjects were lying then we would expect the opposite result; lower acceptance rate for other mTurkers. Fourth, we find no heterogeneity across respondents' race and sex. In other words, both male and female respondents expressed similar perceptions on likelihood of being paid; same for both white and nonwhite respondents. This lack of heterogeneity is indicative of low levels of desirability bias.

Overall, the results from our mTurk 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. Using post-experimental measures of self-reported treatment saliency, we find suggestive evidence that subjects who discriminate against minority groups try to conceal their behavior ex-post. Our results further 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 for female employers. Though the latter effect on number of transcribed receipts is economically meaningful, it is not robust to adjustments for multiple hypothesis testing.

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.[30] 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. (2022). Do workers discriminate against female bosses? *Journal of Human Resources*. forthcoming.

[30]The survey results can be found online at: https://news.gallup.com/poll/1720/work-work-place.aspx.

Abel, M. and D. Buchman (2020). The effect of manager gender and performance feedback: Experimental evidence from india. IZA Discussion Paper No. 13871.

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.

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. mimeo, available online: https://site.stanford.edu/sites/g/files/sbiybj8706/f/5141-discrimination_in_hiring_evidence_from_retail.pdf.

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. NBER Working Paper 25935.

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 67*(6).

Cook, C., R. Diamond, J. V. Hall, J. A. List, and P. Oyer (2021). The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. *The Review of Economic Studies 88*(5), 2210–2238.

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.

Duncan, D., D. Li, and J. D. Graham (2020). Tax rate design and support for mileage user-fees. *Transport Policy 93*, 17–26.

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.

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.

Goldstein Market Intelligence (2020). Global nail care industry analysis: By gender, by products, by distribution channel, based on geography with covid-19 impact forecast period 2017-2030. Report available online: https://www.nailsmag.com/page/598291/market-research.

Hedegaard, M. S. and J.-R. Tyran (2018). The Price of Prejudice. *American Economic Journal: Applied Economics 10*(1), 40–63.

Holmstrom, Bengt Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics & Organization 7*, 24.

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.

International Labour Office (2021). World employment and social outlook 2021: The role of digital labour platforms in transforming the world of work. Report available online: https://www.ilo.org/global/research/global-reports/weso/2021/lang--en/index.htm.

Jaeger, D. A., J. M. Nunley, A. Seals, and E. J. Wilbrandt (2020). The demand for interns. NBER working paper no 26729.

Jones, D., D. Molitor, and J. Reif (2019). What do workplace wellness programs do? evidence from the illinois workplace wellness study. *The Quarterly Journal of Economics 134*(4), 1747–1791.

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.

Nail Magazine (2020). The big book: Industry statistics highlights. Report available online: https://www.nailsmag.com/page/598291/market-research.

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*. fortcoming.

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.

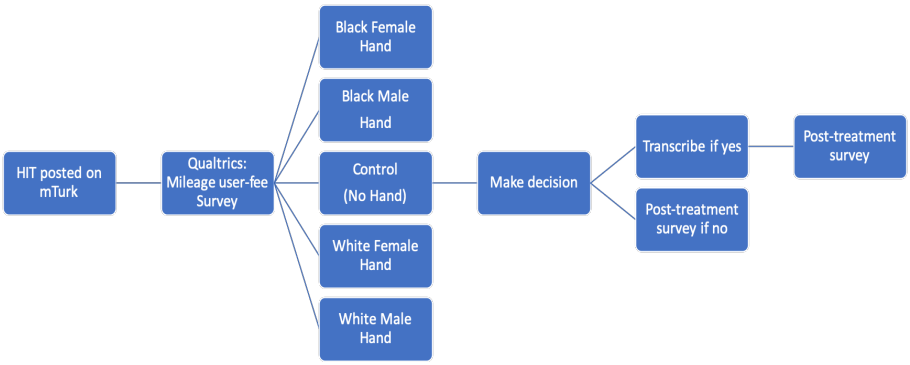
Robinson, J., C. Rosenzweig, A. J. Moss, and L. Litman (2019). Tapped out or barely tapped? recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. *PloS one 14*(12), e0226394.

Trade Union Congress (2021). Seven ways platform workers are fighting back. Report available online: https://www.tuc.org.uk/sites/default/files/2021-11/Platform%20essays%20with%20polling%20data.pdf.

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.

Figure 2: Treatment Instructions

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.

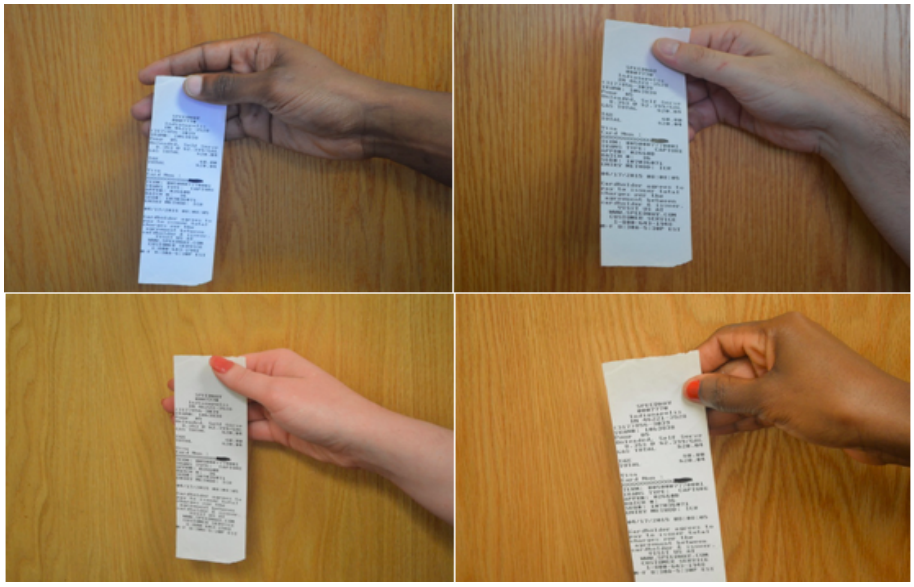
Each receipt should take approximately 30 seconds to transcribe, and I will pay you a bonus of $0.06 for every receipt that you transcribe. You can stop at anytime.

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*

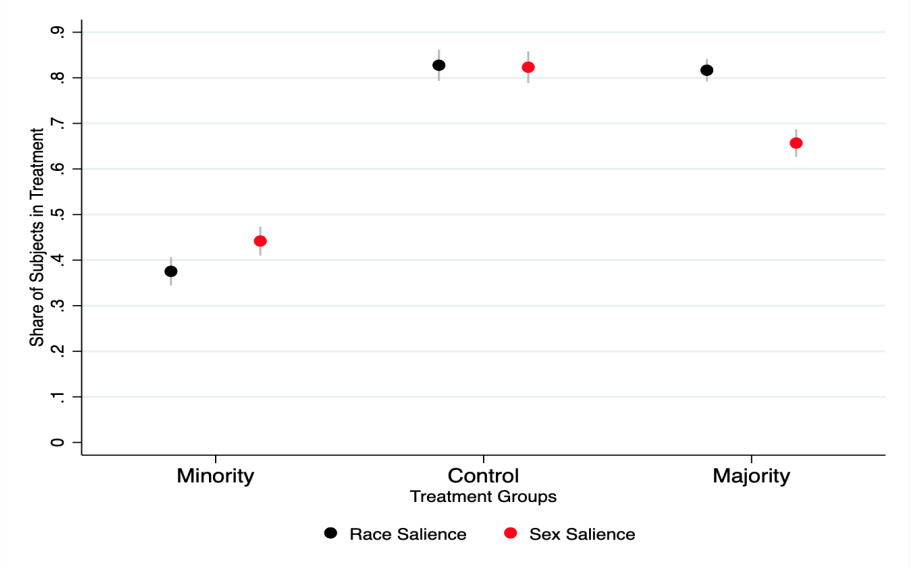
Notes: Reported are the instructions for the bonus transcription task.

Figure 3: Treatment Pictures



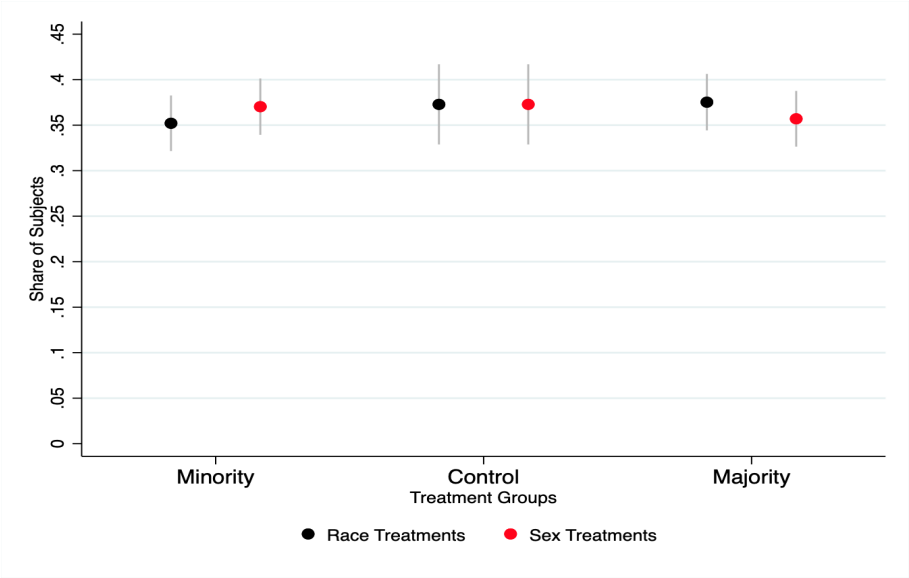
Notes: Reported are the pictures used in the treatment stage of the experiment. The pictures have been compressed significantly to fit side-by-side on one page.

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



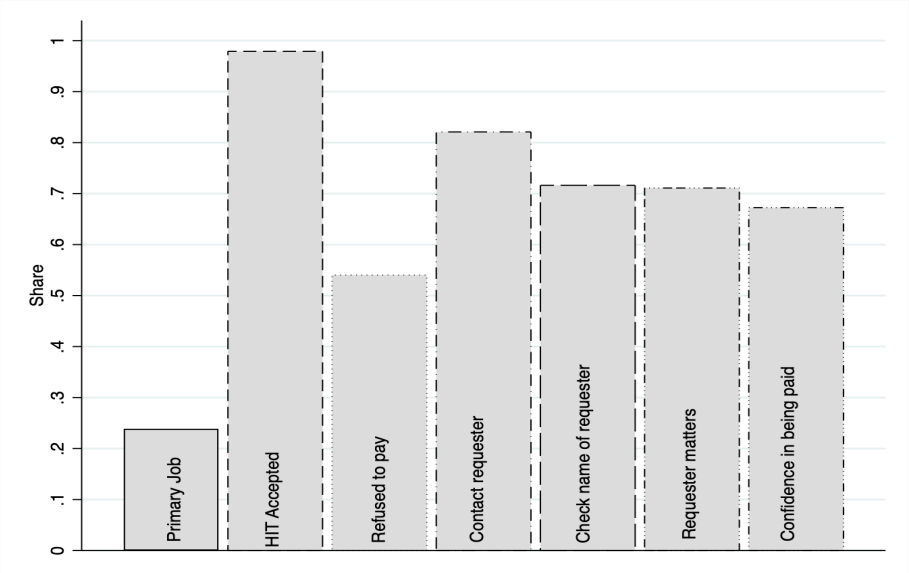
Notes: Reported is the share of subjects whose self-reported perceived treatment matches the actual treatment they are assigned to for race and sex, respectively, along with 95% confidence intervals. Minority refers to Black-hand and female-hand treatments, while majority refers to white-hand and male-hand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

Figure 5: Acceptance Share, by Treatment Group



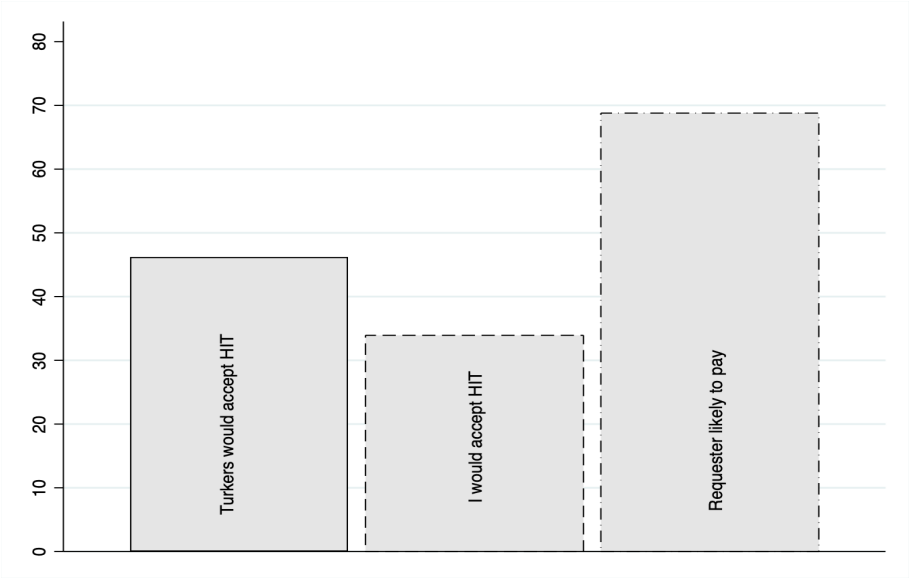
Notes: Reported is the acceptance share by treatment group for race and sex, along with 95% confidence intervals. Acceptance share refers to the share of subjects who agreed to transcribe receipts. Minority refers to Black-hand and female-hand treatments, while majority refers to white-hand and male-hand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

Figure 6: Survey of mTurkers' Labor Supply Experience on mTurk



Notes: Data based on survey of mTurkers who did not participate in the real-effort experiment. 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 7: mTurkers' beliefs about accepting a HIT and being paid



Notes: Data based on survey of mTurkers who did not participate in the real-effort experiment. 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

	Black	White	No Pic	Female	Male	Total	B v W	F v M	ACS(2018)
Age	36.98	37.09	36.31	37.50	36.57	36.89	0.70	0.06	38.2
=1 if female	0.49	0.49	0.42	0.50	0.48	0.48	0.69	0.45	0.51
nonwhite	0.20	0.24	0.26	0.22	0.22	0.23	0.05	0.84	0.25
High School	0.11	0.09	0.10	0.09	0.11	0.10	0.05	0.10	0.27
Some College	0.24	0.23	0.23	0.25	0.23	0.23	0.49	0.34	0.20
2-Year College	0.11	0.12	0.10	0.11	0.12	0.11	0.44	0.82	0.09
B.Sc.	0.38	0.42	0.42	0.38	0.42	0.40	0.06	0.10	0.20
Graduate	0.16	0.15	0.16	0.18	0.13	0.15	0.43	0.01	0.13
Urban	0.51	0.54	0.50	0.53	0.51	0.52	0.19	0.47	0.80

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. The last two columns report he p-vaule of a ranksum test of differences in mean of each covariate for Black v White (B v W) and Female v Male (F v M). 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: Extensive

	Panel A: Race-Gap			Panel B: Sex-Gap		
	Full	Non-White	White	Full	Female	Male
Treatment effect	-0.023	0.000	-0.026	0.010	0.043	-0.015
	(0.022)	(0.049)	(0.025)	(0.022)	(0.033)	(0.031)
Mean	0.374	0.390	0.369	0.358	0.390	0.329
N	1862	412	1450	1862	913	949

Notes: Reported is the race-gap and sex-gap for the extensive-margin based on data from the real-effort experiment. The outcome variable is *Transcribe*, which is equal to 1 if agreed to transcribe pictures and zero otherwise. *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. The gaps are defined such that negative values imply minority groups have worse outcomes. All models include controls for age, sex (= 1 if female), race (= 1 if white), education, and urban (= 1 if live in an urban area). 'Mean' is the average of the outcome variable for the relevant control group; white for the race gap and male for the sex gap. Robust standard errors are reported: $*0.10 * *0.05 * * * 0.01$.

Table 3: Treatment effect of employer race and sex by worker race and sex: Extensive

	Panel A: Race-Gap			Panel B: Sex-Gap		
	Full	Non-White	White	Full	Female	Male
Salient	0.110***	0.151**	0.096***	0.153***	0.158***	0.147***
	(0.032)	(0.070)	(0.036)	(0.032)	(0.045)	(0.044)
Mean	0.385	0.384	0.385	0.369	0.406	0.335
N	1112	246	866	1025	501	524
Non-Salient	-0.054	-0.134*	-0.022	-0.095***	-0.039	-0.143***
	(0.040)	(0.080)	(0.047)	(0.033)	(0.049)	(0.044)
Mean	0.325	0.412	0.288	0.338	0.359	0.317
N	750	166	584	837	412	425

Notes: Reported is the race-gap and sex-gap for the extensive-margin based on data from the real-effort experiment. The outcome variable is *Transcribe*, which is equal to 1 if agreed to transcribe pictures and zero otherwise. *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. All models include controls for age, sex (= 1 if female), race (= 1 if white), education, and urban (= 1 if live in an urban area). 'Mean' is the average of the outcome variable for the relevant control group; white for the race gap and male for the sex gap. Robust standard errors are reported: $*0.10 * *0.05 * * * 0.01$.

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

	Race-Gap			Sex-Gap		
	Full	Non-White	White	Full	Female	Male
Number of Images	-1.794*	1.450	-3.067***	1.579*	0.720	2.691*
	(0.940)	(1.899)	(1.069)	(0.941)	(1.203)	(1.502)
Mean	10.287	8.724	10.808	8.772	8.794	8.748
N	676	159	517	676	374	302
Accuracy	-0.064***	-0.008	-0.080***	0.043***	0.052**	0.034
	(0.016)	(0.036)	(0.018)	(0.016)	(0.022)	(0.023)
Mean	0.816	0.771	0.831	0.766	0.737	0.797
N	676	159	517	676	374	302

Notes: Reported is the race-gap and sex-gap for the intensive-margin based on data from the real-effort experiment. The outcome variables are number of pictures transcribed and share of accurate entries. *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. The gaps are defined such that negative values imply minority groups have worse outcomes. All models include controls for age, sex (= 1 if female), race (= 1 if white), education, and urban (= 1 if live in an urban area). 'Mean' is the average of the outcome variable for the relevant control group; white for the race gap and male for the sex gap. Robust standard errors are reported: $*0.10 * *0.05 * * * 0.01$.

Table 5: Treatment effect of employer race and sex: N. of Pics

	Race-Gap			Sex-Gap		
	Full	Non-White	White	Full	Female	Male
Salient	-0.757	2.590	-2.104	1.967	0.448	4.119**
	(1.265)	(2.692)	(1.430)	(1.217)	(1.594)	(1.921)
Mean	10.966	9.136	11.498	9.537	9.941	9.093
N	470	107	363	443	237	206
Non-Salient	0.196	-0.828	-0.136	0.614	1.671	-0.643
	(1.553)	(3.092)	(1.823)	(1.372)	(1.724)	(2.185)
Mean	6.673	7.429	6.206	7.150	6.357	8.020
N	206	52	154	233	137	96

Notes: Reported is the race-gap and sex-gap for the intensive-margin based on data from the real-effort experiment. The outcome variable is the number of pictures transcribed. *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. All models include controls for age, sex (= 1 if female), race (= 1 if white), education, and urban (= 1 if live in an urban area). 'Mean' is the average of the outcome variable for the relevant control group; white for the race gap and male for the sex gap. Robust standard errors are reported: $*0.10 * *0.05 * * * 0.01$.

Table 6: Treatment effect of employer race and sex: Accuracy

	Race-Gap			Sex-Gap		
	Full	Non-White	White	Full	Female	Male
Salient	-0.061***	0.009	-0.079***	0.053***	0.055**	0.056**
	(0.019)	(0.043)	(0.021)	(0.018)	(0.027)	(0.023)
Mean	0.826	0.788	0.837	0.767	0.732	0.806
N	470	107	363	443	237	206
Non-Salient	-0.025	-0.014	-0.040	0.023	0.035	-0.002
	(0.041)	(0.072)	(0.051)	(0.031)	(0.040)	(0.053)
Mean	0.761	0.719	0.786	0.763	0.749	0.778
N	206	52	154	233	137	96

Notes: Reported is the race-gap and sex-gap for the intensive-margin based on data from the real-effort experiment. The outcome variable is the share of accurate entries. *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. All models include controls for age, sex (= 1 if female), race (= 1 if white), education, and urban (= 1 if live in an urban area). 'Mean' is the average of the outcome variable for the relevant control group; white for the race gap and male for the sex gap. Robust standard errors are reported: $*0.10 * *0.05 * * * 0.01$.

Table 7: Treatment effect of employer race and sex on perceived likelihood of being paid

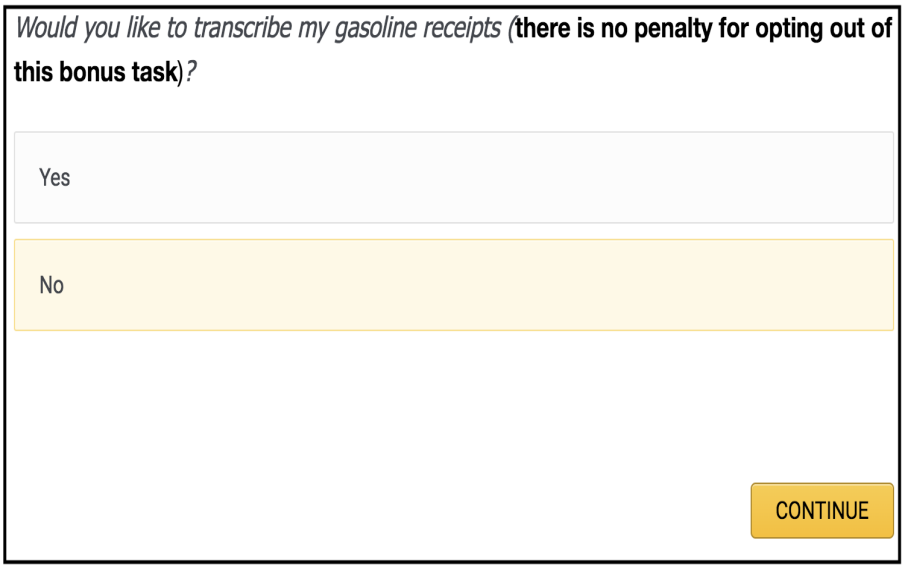
	Full	Non-White	White
Race Gap	-0.890	-2.771	-0.283
	(1.998)	(4.513)	(2.218)
N	765	180	585
	Full	Female	Male
Sex Gap	-0.316	-2.142	0.626
	(2.007)	(3.059)	(2.669)
N	765	348	417

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. All models include controls for age-group dummies, sex (= 1 if female), race (= 1 if white), and education. Robust standard errors are reported: $*0.10 * *0.05 * * * 0.01$.

(Online) Appendix

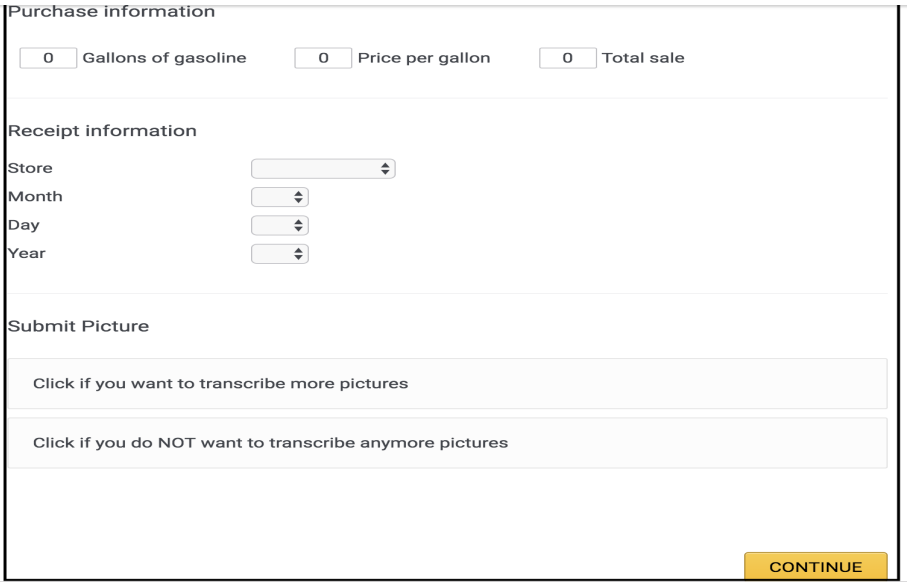
A Additional Details Survey Design

Figure A.1: Treatment Question



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

Figure A.2: Treatment Task



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

B Robustness: Multiple Hypothesis Testing

This section provides a more detailed description of the multiple hypothesis results that are referenced throughout Sections 3.3.1 and 3.2.1 of the paper. Given that we study the effects of two treatments – employer sex and race – on three outcome variables – decision to transcribe, number of transcribed pictures and accuracy – we examine the robustness of our results to multiple hypothesis testing (MHT). To do so, we follow the *wyoung* procedure presented by Jones et al. (2019) with 1000 repetitions to derive standard errors that are adjusted for MHT. The results are presented in Table B.1.

First, we provide MHT for our main analyses in which we estimate the effect of the two treatment variables of interest on our three outcome variables. This MHT jointly tests six hypotheses; 2 treatment variables (race and sex) and 3 outcome variables.[31] The results presented in Panel A of Table B.1 show that the estimates for transcribed and accuracy are robust to MHT. The estimated gaps for number of images transcribed is no longer significant after adjusting for MHT. This finding is consistent with our discussion about the likely impact of wage structure on behavior in Section 3.3.1. Because the wage structure rewards quantity and not quality, we would expect to see much stronger and robust effects for quality compared to quantity.

Second, we conduct MHT for the analyses in which we study the treatment effects separately by the self-reported saliency of the race and sex of the hands displayed in the treatment pictures. It is important to recall that our saliency results do not have a causal interpretation. They are included in the paper only as a means of exploring the reason for our null extensive margin results. Even so, we check and can confirm that the correlations we describe in Tables 3, 5, and 6 are robust to MHT. The MHT results for the race gap tests six hypotheses (1 treatment $\times$ 3 outcome variables $\times$ 2 saliency status) and are presented in Panel B Table B.1. We follow the same procedure to check the robustness of the correlations between saliency and the sex gap; results presented in Panels C.[32]

Third, we check the within-group race and sex gaps for robustness to MHT. We specify a MHT that jointly tests 12 hypotheses for the race gap: i.e., 3 outcomes $\times$ 2 treatments $\times$ race dummy (=1 if white and 0 if non-white). Six of these hypotheses – those that check the race gap by race of worker – are presented in Panel D of Table B.1, while the remaining six hypotheses that check the sex gap by race of worker are not shown since these are not part of our study. We repeat this exercise for the sex-gap and report the results in Panels E. We find that the results for the race gap by race of worker are robust to MHT. However, some of the results for sex gap by sex of worker are not robust to MHT.

Overall, the MHT results suggest that we can be confident in our extensive margin results

[31] Note that the structure of the MHT requires that we include both treatment variables, race and sex, in the estimations (i.e., treatment race is a control variable in specifications where we are interested in the effect of sex and vice versa). Therefore, the model and thus the estimated coefficients are slightly different from the corresponding results presented in the Tables 2 and 4, which are based on Equation 1.

[32] Note that we test the effects of race and sex in separate MHTs because the saliency-status is potentially different across the treatment being studied. For example, a potential employee might be sex-salient, but not race-salient (or vice versa). As a result, we cannot define a common saliency subgroup for worker race and sex.

as well as the accuracy results. However, some caution must be exercised in interpreting the results for number of transcriptions.

Table B.1: Multiple Hypothesis Testing for race and sex gaps

	Race-Gaps			Sex-Gaps		
	Effect	Unadjusted	Adjusted	Effect	Unadjusted	Adjusted
	Panel A: full sample					
	Race-Gaps			Sex-Gaps		
Transcribed	-0.023	0.306	0.543	0.010	0.663	0.669
N. Transcribed	-1.802	0.055	0.197	1.588	0.091	0.230
Accuracy	-0.064	0.000	0.000	0.044	0.006	0.031
	Panel B: Race Gap x Saliency					
	Non-Salient Sample			Salient Sample		
Transcribed	-0.054	0.178	0.536	0.110	0.001	0.002
N. Transcribed	0.196	0.899	0.904	-0.757	0.550	0.904
Accuracy	-0.025	0.549	0.906	-0.061	0.001	0.007
	Panel C: Sex Gap x Saliency					
	Non-Salient Sample			Salient Sample		
Transcribed	-0.095	0.004	0.016	0.153	0.000	0.000
N. Transcribed	0.614	0.655	0.650	1.967	0.107	0.301
Accuracy	0.023	0.453	0.704	0.053	0.004	0.019
	Panel D: Race Gap x Race of worker					
	White workers			Nonwhite workers		
Transcribed	-0.026	0.296	0.928	-0.001	0.979	0.996
N. Transcribed	-3.088	0.004	0.034	1.504	0.430	0.942
Accuracy	-0.081	5.006e	0.000	-0.008	0.829	0.982
	Panel E: Sex Gap x Sex of worker					
	Male workers			Female workers		
Transcribed	-0.016	0.605	0.925	0.043	0.192	0.754
N. Transcribed	2.721	0.071	0.506	0.681	0.571	0.925
Accuracy	0.035	0.136	0.689	0.051	0.018	0.161

Notes: Reported is the race-gap and sex-gap with adjustments for multiple hypothesis testing. *Transcribe* is equal to 1 if agreed to transcribe pictures and zero otherwise. Accuracy is the share of accurate entries. 'Effect' is the estimated sex/race gap; 'Unadjusted' is the original p-value, which does not account for multiple hypothesis testing, and 'Adjusted' is pvalues adjusted for multiple hypothesis testing. *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. Panels A, B and C each test 6 hypotheses. Panels D and E tests 12 hypotheses each, however, only the 6 that are relevant for our study are shown here. Panel D excludes results for the sex gap by race of worker and Panel E excludes results for the race gap by sex of worker. All models include controls for age-group dummies, sex (= 1 if female), race (= 1 if white), and education. Robust standard errors are reported: $*0.10 * *0.05 * * * 0.01$.

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

Table C.1: Pilot-experiment Results

	BF	BM	WF	WM
Race-Salient	82.86	62.07	89.66	82.76
Sex-Salient	82.86	75.86	89.66	25.00

Notes: Reported is the percent of subjects who correctly identified the race and sex of the hand in the pictures used in our experiment. Data are from our pilot experiment. A subject's group assignment is indicated in the columns: BF is black female, BM is black male, WF is white female, and WM is white male.

Table C.2: Number of Subjects who are dropped by treatment group

	BF	BM	No Pic	WF	WM	Total
President	13	17	17	13	16	76
IPAddress	16	13	23	19	15	86

Notes: Reported is the number of subjects who are dropped from the dataset across treatment groups. President is an indicator variable (=1 if say Michael Jordan is president of the USA) and IPAddress is an indicator variable (=1 is two or more subjects have the same IP address). BF is black female hand, BM is black male, WF is white female and WM is white male.

Table C.3: Summary Statistics by Treatment Group

	BF	BM	No Pic	WF	WM	Total
Age	37.85	36.11	36.31	37.14	37.03	36.89
=1 if female	0.51	0.48	0.42	0.49	0.48	0.48
nonwhite	0.20	0.21	0.26	0.25	0.23	0.23
High School	0.09	0.13	0.10	0.08	0.09	0.10
Some College	0.26	0.23	0.23	0.23	0.23	0.23
2-Year College	0.12	0.10	0.10	0.11	0.13	0.11
B.Sc.	0.35	0.41	0.42	0.41	0.43	0.40
Graduate	0.19	0.13	0.16	0.16	0.13	0.15
Urban	0.53	0.48	0.50	0.53	0.54	0.52

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 C.4: Balancedness test

	M v C	M v F	F v C	W v C	W v B	B v C
Age	0.90	0.06	0.13	0.38	0.70	0.62
=1 if Female	0.05	0.45	0.01	0.03	0.69	0.01
=1 if Non-white	0.13	0.84	0.18	0.53	0.05	0.02
High School	0.61	0.10	0.40	0.33	0.05	0.53
Some College	0.94	0.34	0.48	0.97	0.49	0.55
2-Year College	0.30	0.82	0.39	0.21	0.44	0.52
B.Sc.	0.98	0.10	0.17	0.94	0.06	0.14
Graduate	0.20	0.01	0.35	0.65	0.43	0.85
Urban	0.56	0.47	0.24	0.16	0.19	0.73

Notes: Reported are the Pvalues from a ranksum test of the differences in means between treatment groups. M v C is male compared to control; M v F is male compare to female; F v C is female compared to control. W vs B is white compared to black, and W vs C is white compared to control and B v C is black compared to control.

Table C.5: Characteristics of Salient and Non-salient samples

	Race		Sex		
	Non-Salient	Salient	Non-Salient	Salient	Full sample
Age	38.50	35.99	38.18	36.04	36.89
=1 if female	0.48	0.47	0.48	0.47	0.48
nonwhite	0.22	0.23	0.23	0.23	0.23
High School	0.11	0.09	0.11	0.10	0.10
Some College	0.20	0.25	0.22	0.24	0.23
2-Year College	0.11	0.11	0.11	0.11	0.11
B.Sc.	0.39	0.41	0.38	0.42	0.40
Graduate	0.18	0.14	0.18	0.13	0.15
Urban	0.50	0.52	0.51	0.52	0.52

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 C.6: Correlation between transcribing and saliency

	Panel A: Race		Panel B: Sex	
	Black	White	Female	Male
Correlation	0.241***	0.038	0.299***	0.029
	(0.032)	(0.026)	(0.032)	(0.033)
pValue	0.000		0.000	
N	932	930	930	932

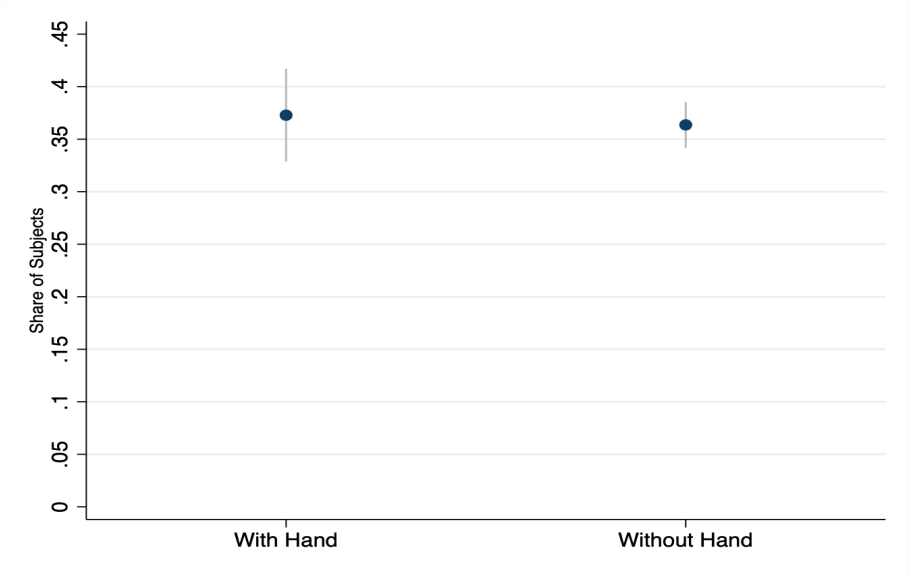
Notes: Reported is the estimated correlation from the model $S_i = \alpha + \beta Transcribed_i + \delta X_i + \epsilon_i$. S_i is either race or sex saliency, X_i is a vector of covariates. The model is estimated separately for black and white employers when the outcome variable is race saliency, and separately for male and female employers when the outcome variable is sex saliency. pValue is from a test of differences in β between the minority and majority group. Robust standard errors are reported: $*0.10 * *0.05 * * * 0.01$.

Table C.7: Characteristics of Transcribers

	BF	BM	No Pic	WF	WM	Total
Age	35.84	34.72	35.43	37.43	35.83	35.87
=1 if female	0.59	0.56	0.49	0.57	0.49	0.54
nonwhite	0.21	0.24	0.28	0.24	0.26	0.25
High School	0.11	0.10	0.09	0.10	0.07	0.10
Some College	0.24	0.27	0.21	0.25	0.25	0.24
2-Year College	0.12	0.09	0.12	0.08	0.11	0.10
B.Sc.	0.35	0.41	0.42	0.42	0.43	0.41
Graduate	0.19	0.12	0.16	0.14	0.14	0.15
Urban	0.51	0.50	0.54	0.52	0.50	0.52

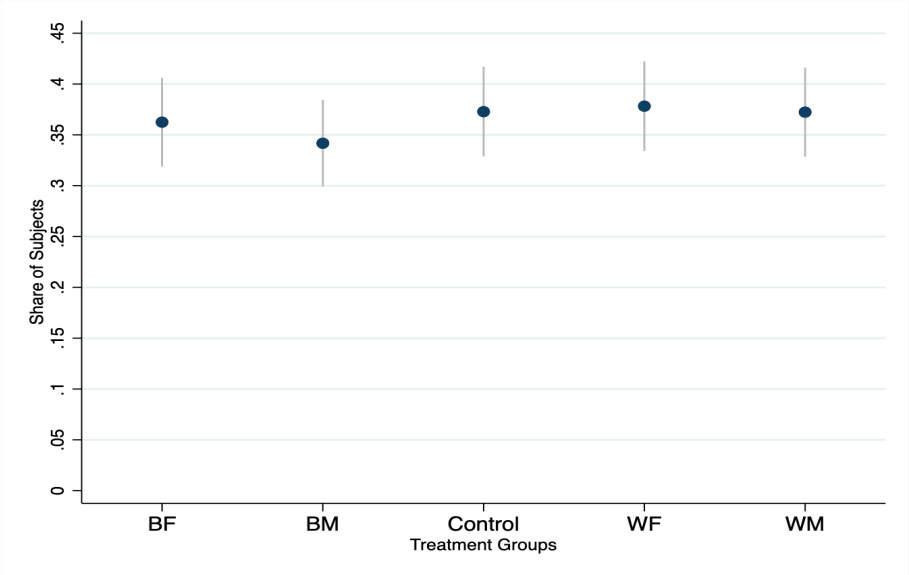
Notes: Reported is the mean of each variable for subjects who transcribed at least 1 picture. BF includes subjects who were assigned to the black female hand treatment; BM includes subjects who were assigned to the black male hand treatment; Control includes subjects who were assigned to the control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



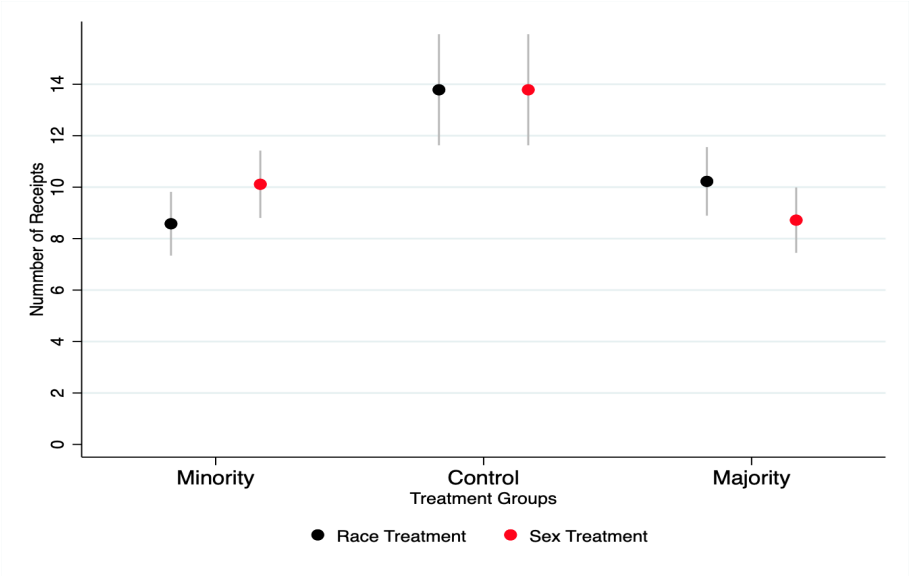
Notes: Reported is the share of subjects who agreed to transcribe receipts. 'With Hand' indicates that the image of the receipt included a hand; 'Without Hand' indicates subjects in the control group.

Figure C.2: Acceptance Share across all treatment groups



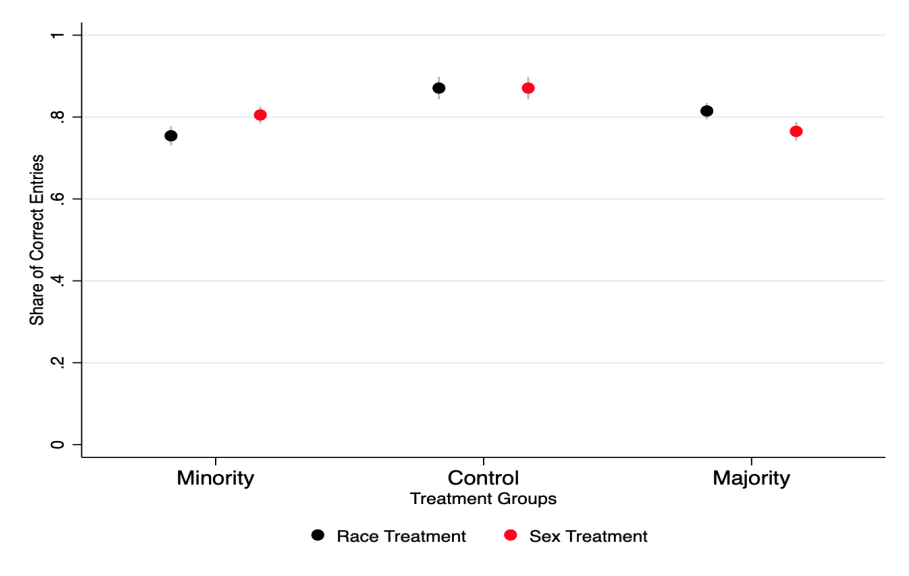
Notes: Reported is the share of subjects who agreed to transcribe receipts in each treatment group. BF includes subjects who were assigned to the black female hand treatment; BM includes subjects who were assigned to the black male hand treatment; Control includes subjects who were assigned to the control 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 C.3: 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.

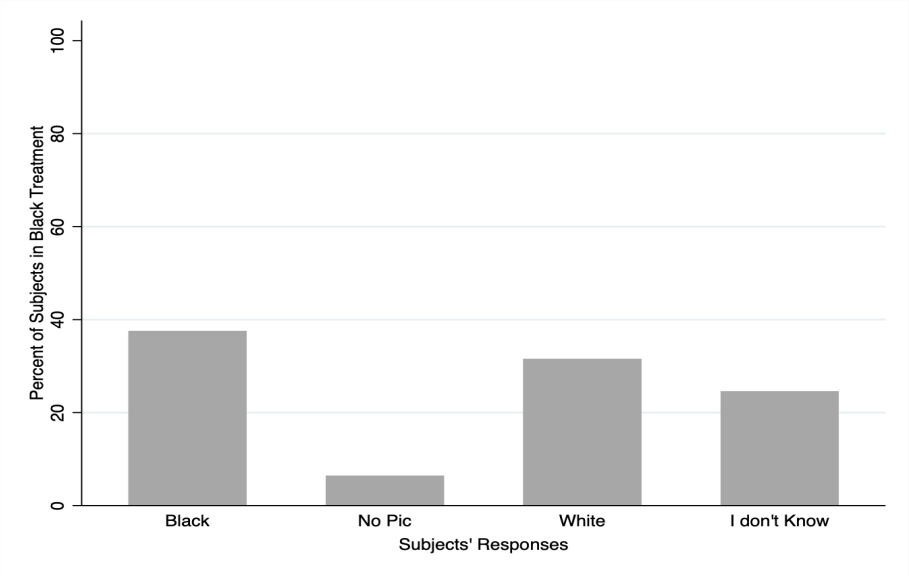
Figure C.4: Accuracy Rate, by Treatment Group



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.

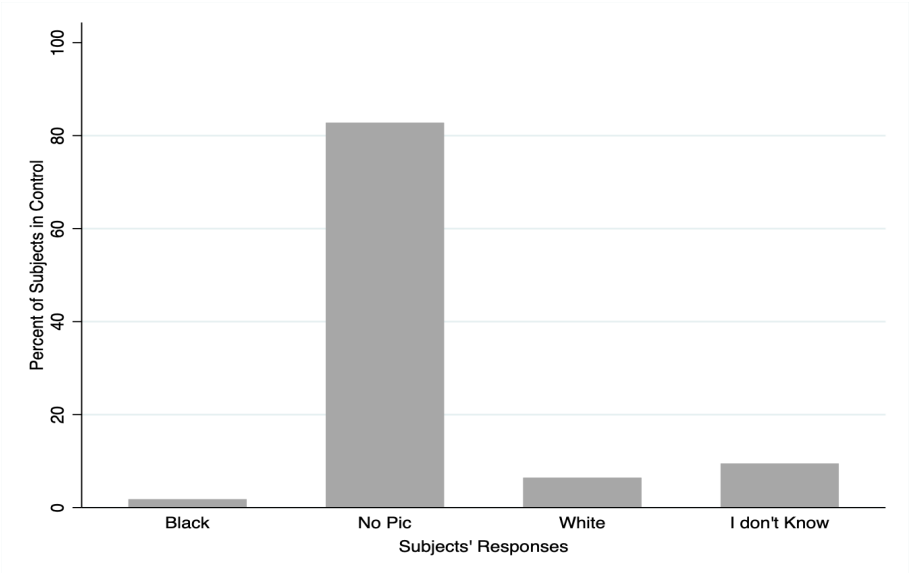
D Salience Across Treatment Groups

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



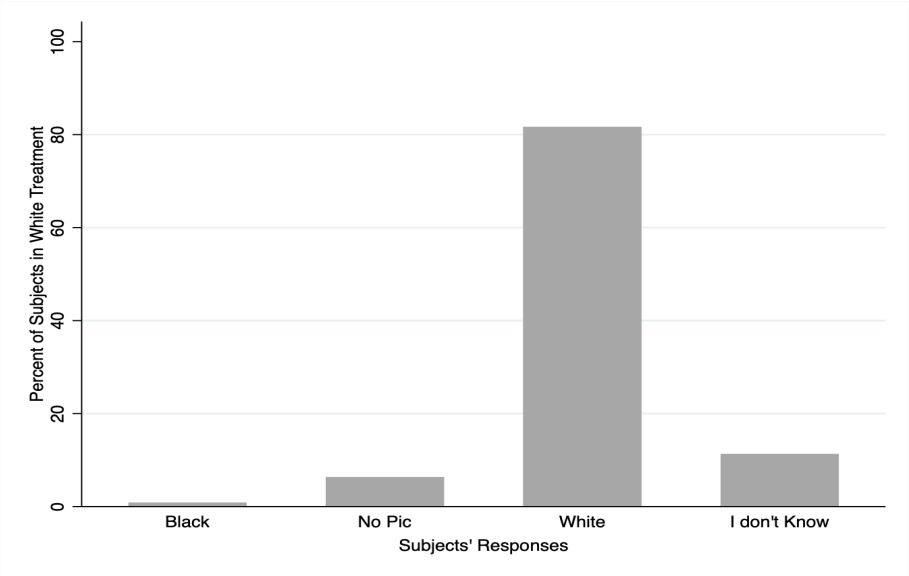
Notes: Reported is the percent of subjects in the black treatment who gave each of the possible responses to the question: "What is the race of the person holding the receipt in the picture?".

Figure D.2: Race Salience in the Control Group



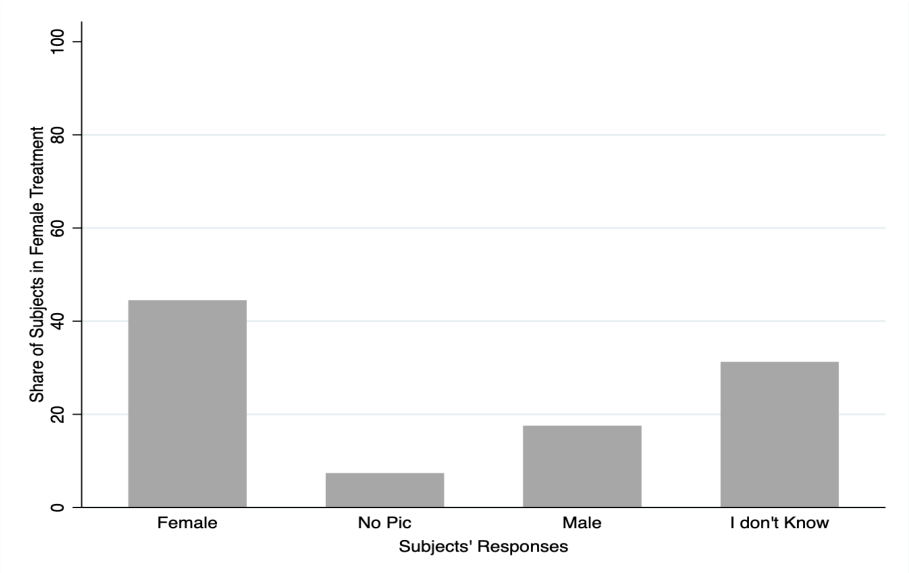
Notes: Reported is the percent of subjects in the control group who gave each of the possible responses to the question: "What is the race of the person holding the receipt in the picture?".

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



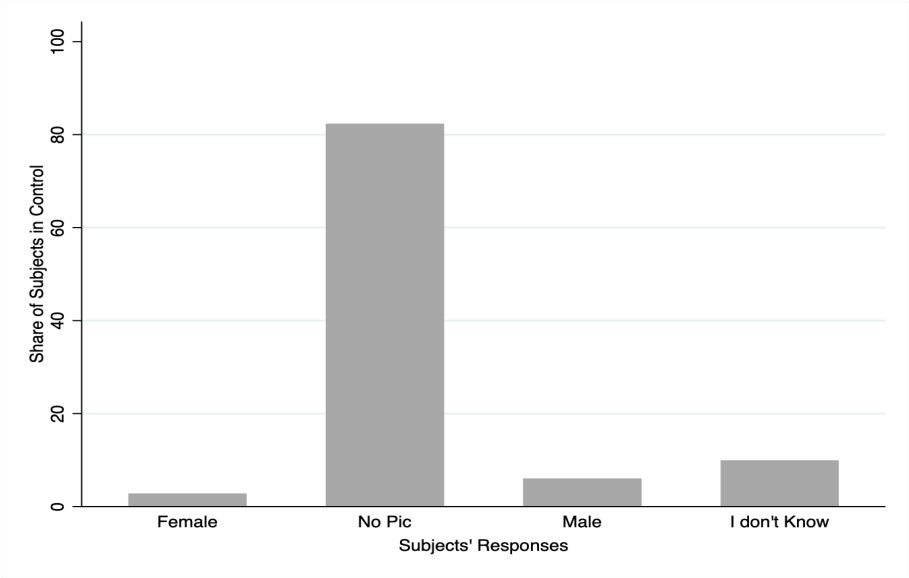
Notes: Reported is the percent of subjects in the white treatment who gave each of the possible responses to the question: "What is the race of the person holding the receipt in the picture?".

Figure D.4: Sex Salience in the Control Group



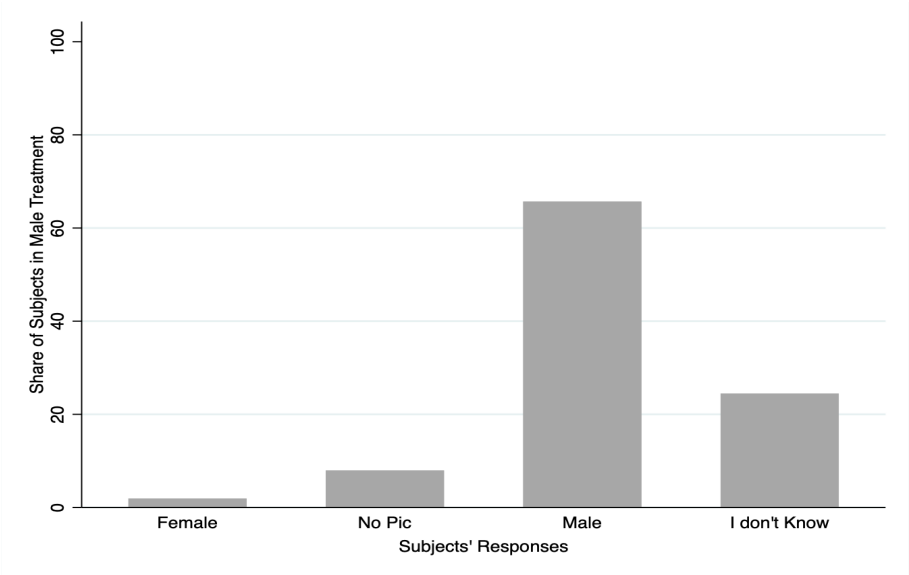
Notes: Reported is the percent of subjects in the control group who gave each of the possible responses to the question: "What is the sex of the person holding the receipt in the picture?".

Figure D.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?".

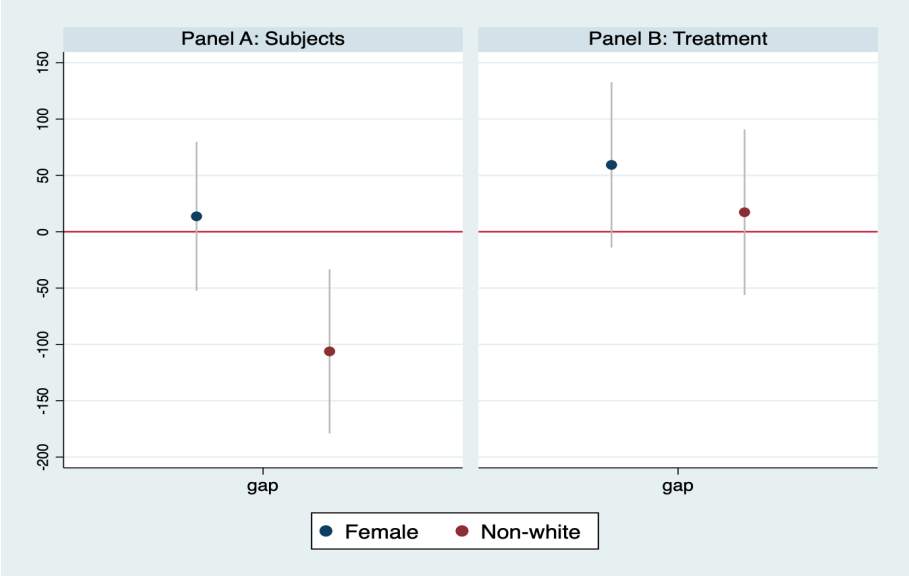
Figure D.6: Sex Salience in the Male Treatment Group



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?".

E Follow-up Survey: Additional Statistics

Figure E.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 E.1: Survey of mTurkers: Balancedness test

	M v C	M v F	F v C	W v C	W v B	B v C
18 to 24	0.65	0.09	0.35	0.68	0.72	0.90
25 to 34	0.59	0.03	0.23	0.42	0.25	0.89
35 to 44	0.36	0.74	0.53	0.68	0.37	0.25
45 to 54	0.30	0.95	0.32	0.38	0.71	0.24
55 to 64	0.68	0.35	0.73	0.84	0.58	0.80
65 and older	0.72	0.21	0.21	0.29	0.55	0.52
Non-white	0.58	0.92	0.53	0.54	0.96	0.57
Female	0.55	0.87	0.47	0.65	0.61	0.38
High School	0.92	0.33	0.49	0.69	0.83	0.82
Some College	0.95	0.66	0.77	0.44	0.11	0.60
2-Year College	0.32	0.54	0.62	0.41	0.86	0.50
B.Sc.	0.33	0.82	0.43	0.46	0.73	0.31
Graduate	0.67	0.94	0.71	0.41	0.27	0.96

Notes: Reported are the Pvalues from a ranksum test of the differences in means between groups. W v C is white compared to control, W vs B is white compared to black, B v C is black compared to control, M v C is male compared to control, M v F is male compared to female and F v C is female compared to control.

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

	Experiment	Survey	Experiment	Survey
BF	0.45	0.41	0.40	0.41
BM	0.30	0.33	0.62	0.55
Control	0.83	0.88	0.82	0.88
WF	0.83	0.84	0.49	0.47
WM	0.80	0.78	0.69	0.62

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.