2020

Three essays in behavioral economics

Sher Afghan Asad
Iowa State University

Three essays in behavioral economics

by

Sher Afghan Asad

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Economics

Program of Study Committee:

Joydeep Bhattacharya, Major Professor
Otávio Bartalotti
Elizabeth Hoffman
Desire Kedagni
Peter Orazem

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2020

DEDICATION

I dedicate this work to my mother who has gone through many hardships in life to get me where I am today. I am also thankful for the support and love of my sisters, Saman and Kiran, and brother Ali.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS

I am most thankful to my adviser, Joydeep Bhattacharya, for providing constant support and motivation throughout my research. Joydeep was not only an advisory to me but my best friend in Ames. Joydeep's insights and out-of-the-box thinking have inspired me to think bigger and aim for higher goals. He always created a nurturing environment which allowed me to discuss ideas, no matter how stupid!, without any inhibitions. A true academic experience. I am thankful for all the chats, meetings, drinks, chicken wings, and badminton among other things.

I would like to thank my committee members, Betsy Hoffman, Peter Orazem, Otavio Bartalotti, and Desire Kedagni, for their guidance and support throughout the course of this research. I am also thankful to my co-author Ritwik Banerjee for his guidance throughout, without which most of this research would not have been possible.

I want to acknowledge the support provided by the staff of the economics department. The experiment that I ran would not be possible without the logistical support provided by Christopher Jorgensen and the IT department. By the way, Chris has a hilarious letter series called Jackass Letters that's worth checking out.

Last but not the least, I would like to thank my friends, colleagues, the department faculty and staff for making my time at Iowa State University a wonderful experience.

ABSTRACT

Over the last few decades, behavioral economics, by introducing psychology in the decision making process, has changed the way we think about the policy problems of our age. Governments and non-governmental organizations all over the world are using insights from behavioral economics to solve wide ranging issues in the domains such as health, wealth, education, social security, environment etc.. In this dissertation research I use insights from behavioral economics to understand two unrelated but pressing problems of our age; labor-market discrimination and road-safety behavior.

In this dissertation, I first present a theory of worker side discrimination and highlight the importance of exploring worker side discrimination to improve our overall understanding of labor market discrimination. The main thesis of this work is that in many situations, workers' motivation to work for employer depends on employer's social group. More precisely, workers feel more motivated when they work for the same group employer and less so when working for an out-group employer. Workers' productivity differential provides incentive to the non-discriminatory employers to recruit workers from their own group and pay higher wages to own-group workers. The assortative matching of employers and workers leads to segregation of the labor force when there are enough same-group employers. However, under-representation of employers of one group leads to adverse labor market outcomes for the workers of that group in terms of wages. I show that this has implications for how we interpret the existence and source of discrimination in the labor markets. Specifically, I demonstrate that what is traditionally understood as discrimination by employers may, in fact, be a rational response to the worker's differential social preferences towards the employer's group identity. I also show that ignoring worker social preferences (and employer's beliefs about them) may lead to misleading conclusions about the sources of discrimination.

In another chapter of this dissertation, I (with my co-authors) explore the evidence for the above theorized channel of discrimination in an online labor market. Specifically, we examine whether workers in the online economy discriminate against their employers via their social-preferences / motivation. In this chapter, we focus on racial identity and ask, do workers discriminate (say, by under providing effort) for an out-race employer relative to an otherwise-identical, own-race one? We run a well-powered model-based experiment using subjects from Amazon's Mechanical Turk (M-Turk). Interestingly, we find that white workers do not discriminate against their out-group employers, in-fact they work harder for black employers as compared to white employers. The results are exciting because it reflects the lack of bias in workers preferences towards the minority group employers in the online economy. The results imply that as the economies transition to online jobs, the possibility of discriminatory behavior against minorities may diminish.

In the last chapter, I (along with my co-author) study another application of behavioral economics to understand road-safety behavior of the automotive drivers. Particularly, we look at the traffic-related messages such as “drive sober,” “x deaths on roads this year,” and "click it or ticket,”, displayed on major highways, on reported near-to-sign traffic accidents. To estimate the causal effect of these nudges, we build a new high-frequency panel dataset using the information on the time and location of messages, traffic incidents, overall traffic levels, and weather conditions using the data of the state of Vermont. We estimate several models that control for endogeneity of these messages, allow for spillover effects from neighboring messages, and look at the impact as the function of distance from the sign. We find that behavioral nudges, such as “drive sober” and “wear seat belt”, are at best ineffective in reducing the number of crashes while informational nudges, such as “slippery road” and “work zone”, actually lead to causal reduction in number of crashes. Our findings are robust to many different specifications and assumptions.

CHAPTER 1. GENERAL INTRODUCTION

Over the last few decades, behavioral economics, by introducing psychology in the decision making process, has changed the way we think about the policy problems of our age. Governments and non-governmental organizations all over the world are using insights from behavioral economics to solve wide ranging issues in the domains such as health, wealth, education, social security, environment etc.. In this dissertation research I use insights from behavioral economics to understand two unrelated but pressing problems of our age; labor-market discrimination and road-safety behavior.

Discrimination in the labor-markets is defined as the valuation of personal characteristics (such as race, gender, age, attractiveness etc.) that are unrelated to productivity. For example, discrimination occurs if a black worker systematically faces higher unemployment or gets lower wages than an *equally productive* white worker. Discrimination is an issue probably as old as humanity itself and the discrimination against blacks is well documented in the United States. The issue is how do we address it.

There is a large body of literature in economics and other social sciences that try to understand the nature of the existence of discrimination. Economists generally think of labor market discrimination as an outcome of *employer's* preferences or *employer's* beliefs about the discriminated workers. The focus here is on employer, i.e., the literature mostly assumes that discrimination is driven by the employers. However, in my dissertation research, I argue that discrimination may also be driven by the worker side and understanding this is important from a policy perspective. The next two chapters of this dissertation concern the understanding of this discrimination from the workers' side.

In Chapter 2, I present a theory of worker side discrimination and highlight the importance of exploring worker side discrimination to improve our overall understanding of labor market discrimination. The main thesis of this chapter is that in many situations workers' motivation to work for employer depends on employer's social group. More precisely, workers feel more motivated when they work for the same group employer and less so when working for an out-group employer. This implies that, in an incomplete contract environment (i.e. when workers' performance cannot be dictated), workers are more productive for the same group employer and therefore, from an employer perspective, it is optimal to hire workers from the same group even though the workers from the out-group are equally (or more) qualified. This shows that the wage differential between, for example, blacks and whites may arise because of productivity differential of equally qualified workers, and not because of discrimination.

I build this theory by outlining a simple principal-agent model in which wages are determined by the worker's outside option and the motivation towards the employer group. Wage differential between the two

groups arise if either the outside option of one group is different than the other, or the worker's motivation towards the employer's group varies. I show that if the two groups are well represented on the demand side of labor market, then there are no wage differential on the basis of group-based motivation, there is complete segregation of the labor market, and each worker is paid equal to their marginal product. However, the outside options of two groups vary if one group is underrepresented on the demand side of the labor market. The under-representation of employers of one group reduces the outside option of workers of that group because they have to work in cross organizations where they are less productive. This productivity differential translates into difference in wages as workers are paid equal to their marginal product. The disadvantaged workers get lower wages not only from the cross type employers but also from the same type employer because their outside option is lowered.

Chapter 2 also shows that ignoring worker social preferences (and employer's beliefs about them) may lead to misleading conclusions about the sources of discrimination. For example, ignoring worker discrimination may imply that employer discrimination, mostly understood by economists to be preference-based, may be somewhat belief-based because it may be driven, not by animus of the employer, but by beliefs a rational employer holds about the preferences of the workers from a particular group. That latter distinction is vital because designing an effective policy intervention to reduce discrimination crucially depends on the source of discrimination. I conclude this chapter by arguing that policy makers should consider the worker side of the market and employer's response to it to come up with more effective policies.

In Chapter 3, I (along with my co-authors) explore the evidence for this theorized channel of discrimination in an online labor market. Specifically, we examine whether workers in the online economy discriminate against their employers via their social-preferences / motivation. In this chapter, we focus on racial identity and ask, do workers discriminate (say, by under providing effort) for an out-race employer relative to an otherwise-identical, own-race one? We run a well-powered model-based experiment using subjects from Amazon's Mechanical Turk (M-Turk). Interestingly, we find that white workers do not discriminate against their out-group employers, in-fact they work harder for black employers as compared to white employers. The results are exciting because it reflects the lack of bias in workers preferences towards the minority group employers in the online economy. The results imply that as the economies transition to online jobs, the possibility of discriminatory behavior against minorities may diminish.

The experimental design in Chapter 3 is tightly connected to a simple structural model in which workers have race-dependent social preferences towards their employer and maximize utility from the provision of costly effort. We take the approach of revealing race indirectly via the revelation of skin color and voice: employer-subjects are videotaped while they read off a script explaining and demonstrating the task for the workers. The camera placement only captures the hand of the employer along with the movement of the

fingers alternating 'a' and 'b' button presses. Other identifiers, such as the face, are not revealed. In the neutral treatments, gloves and other clothing hide the skin entirely and the worker is aware of being matched to an employer but is unaware of any identity clues. In the experiment, we introduce a total of ten treatment variations. In the first three, we vary the piece-rate with an aim to identify and estimate the cost-of-effort function. Here, the worker is not given any information about the existence of (non-existent) employer; any earnings from his/her effort choices go entirely to the worker. The next set of three treatments aim to a) detect the baseline level of altruism towards the hidden race of the employer (altruism neutral) and b) estimate race-specific altruism towards the revealed race of the employer (altruism black and altruism white). The final treatments are designed to a) detect the baseline level of reciprocity towards the hidden race of the employer (reciprocity neutral) and b) estimate the race-specific variations in reciprocity towards the revealed race of the employer (reciprocity black and reciprocity white). Thus, the ten treatments help us identify the cost-of-effort function and social-preference parameters (altruism and reciprocity) of the structural model separately for neutral (hidden race), black, and white employers.

In Chapter 4, I (along with my co-author) study another application of behavioral economics to understand road-safety behavior of the automotive drivers. Influenced by the nudge revolution, this chapter tests the effectiveness of public information nudge in changing behaviors. Nudges are defined as "choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives." These benign behavioral interventions have become increasingly popular with the researchers and the governments all over the world to address various policy problems. There have been hundreds of studies which show that nudges are effective in influencing behavior ranging from donating organs, reducing energy consumption, and saving more money. While there is considerable evidence on the effectiveness of nudges, there is not much known about what kind of nudges are most effective and if poorly designed nudges can backfire? In this paper, we study the effectiveness of different nudges on road safety behavior and examine whether all nudges are created equal and if not which ones are most effective in ensuring road safety?

Particularly, we look at the traffic-related messages such as "drive sober," "x deaths on roads this year," and "click it or ticket,", displayed on major highways, on reported near-to-sign traffic accidents. To estimate the causal effect of these nudges, we build a new high-frequency panel dataset using the information on the time and location of messages, traffic incidents, overall traffic levels, and weather conditions using the data of the state of Vermont. We estimate several models that control for endogeneity of these messages, allow for spillover effects from neighboring messages, and look at the impact as the function of distance from the sign. We find that behavioral nudges, such as "drive sober" and "wear seat belt", are at best ineffective in reducing the number of crashes while informational nudges, such as "slippery road" and "work zone", actually

lead to causal reduction in number of crashes. Our findings are robust to many different specifications and assumptions.

The main challenge in Chapter 4 is the endogeneity of nudges and crashes, i.e., is the change in the number of crashes happening because of nudges, or would the crashes be lower (higher) anyway at the time these nudges are displayed, perhaps because these are displayed during times when conditions are relatively safe (unsafe) for driving. In particular, we observe that informational nudges (such as about weather and crashes on the road) are selectively displayed when conditions are precarious while behavioral nudges are selected in otherwise less risky conditions. We introduce various specifications to account for this endogeneity. In one specification, we run a structural equation model in which we explicitly model the process that determines these nudges along with the process that determine the crashes while allowing the latent factors to effect both nudges and crashes. In another specification, we run a lagged quasi first-difference Poisson model using generalized method of moment (GMM) to account for this endogeneity.

Finally, in Chapter 5 I conclude this dissertation by highlighting some of the insights from all these chapters, and outline the way forward for future research.

CHAPTER 2. A THEORY OF DISCRIMINATION WITH MOTIVATED WORKERS

Modified from a manuscript to be submitted to the Quarterly Journal of Economics.

Sher Afghan Asad

Iowa State University

2.1 Abstract

Economic literature on labor market discrimination has been mostly focused on the demand side. This paper shifts the focus from the demand side (employers) to the supply side (workers) for a more nuanced understanding of labor market discrimination. I outline a simple principal-agent model of the labor market and study the contract design when workers have differential social preferences depending on the employer's group identity (e.g., race). I show that worker's identity-based social-preferences imply that the workers are more productive (work harder) for the same identity employers. Workers' productivity differential provides incentive to the non-discriminatory employers to recruit workers from their own group and pay higher wages to own-group workers. The assortative matching of employers and workers leads to segregation of the labor force when there are enough same-group employers. Under-representation of employers of one group leads to adverse labor market outcomes for the workers of that group in terms of wages. I show that this has implications for how we interpret the existence and source of discrimination in the labor markets. Specifically, I demonstrate that what is traditionally understood as discrimination by employers may, in fact, be a rational response to the worker's differential social preferences towards the employer's group identity. I also show that ignoring worker social preferences (and employer's beliefs about them) may lead to misleading conclusions about the sources of discrimination.

Keywords: discrimination; worker-to-employer; identity; social-preferences; taste-based discrimination; statistical discrimination; contract theory

JEL Codes: D86, D91, J71, J41

This paper has benefitted from countless discussions with my excellent advisors and mentors, Joydeep Bhattacharya and Ritwik Banerjee. I also thank Otavio Bartalotti, Betsy Hoffman, Jian Li, Peter Orazem, and Bertan Turhan for their valuable comments and suggestions.

2.2 Introduction

Disparities in labor market outcomes based on social groups (such as gender, ethnicity, race and religion) are ubiquitous,[1] and a large body of literature in economics has argued that a part of these disparities is a consequence of discrimination.[2] Discrimination, in the empirical studies, is defined as an unexplained differential in labor market outcomes after controlling for factors such as age, education, or more generally the ability of the workers from different groups. These studies mostly focus on the employer side of the labor market and assume market discrimination as the outcome of *employer's* preferences or differential beliefs about ability of the workers from certain groups. Apart from few exceptions, the studies on discrimination ignore the worker side of the market. In this paper, I outline a theoretical model to study the *worker* side of the market and show that ignoring workers' preferences may lead to misleading conclusions about the existence as well as the sources of discrimination.

Suppose workers differ in their intrinsic motivation to work for employers based on employer's group identity. More precisely, workers feel more motivated when they work for the same group employer and less so when working for an out-group employer. This implies that, in an incomplete contract environment, workers are more productive for the same group employer and therefore, from an employer perspective, it is optimal to hire workers from the same group even though the workers from the out-group are equally (or more) qualified. In this setting, a researcher's neglect of the workers' differential motivation will lead to a wrong conclusion that differential in labor market outcomes is discrimination when in-fact the differences are driven by the productivity differentials.[3] These productivity differentials arise, for example, in a white-dominated firm a black worker feel out of place (e.g. because of cultural or social differences), and that reflects in sub-optimal performance by black worker even though the worker is otherwise of high ability.[4]

Fundamentally, economists view discrimination as arising in one of two ways: taste or statistical. Becker (1957) introduced the notion of taste-based discrimination postulating that discrimination exists because of prejudice/animus of the majority group towards the minority group. Phelps (1972) and Arrow (1973) instead view discrimination as statistical, in which an employer, lacking information about the worker's ability, forms belief about the worker's ability based on worker's group identity using the aggregate distribution

[1] For example, in 2010, black men were 28 percent less likely to be employed and earned 31 percent less annually conditional on employment than white men in the United States (Kahn-Lang, 2018).

[2] See Charles and Guryan (2011), Rich (2014), Bertrand and Duflo (2017), and more recently Lang and Kahn-Lang (2020) for a survey on evidence of discrimination.

[3] I use Arrow (1973) definition of discrimination as "the valuation in the market-place of personal characteristics of the worker that are unrelated to worker productivity".

[4] This is different from co-worker discrimination à la Becker (1957) where coworkers derive negative utility from working with minorities. Here, it is argued that a worker from minority will be less motivated to work, not because of distaste of coworkers, but because of worker's own lack of preference in working with the out-group.

of worker's group traits.[5] Distinguishing between the sources of discrimination is important from a policy perspective as each source of discrimination warrants different policy response. For example, reducing taste-based discrimination requires policy actions that increase competition or makes it costly for employers to discriminate (Levine, Levkov, & Rubinstein, 2008); on the other hand, statistical discrimination can be addressed by providing more information to the employers.

In this paper, I show that ignoring workers' identity-based preferences can lead to misclassification of the source of discrimination. More specifically, differential motivation by workers may imply that employer discrimination, mostly understood by economists to be taste-based, may be somewhat statistical because it may be driven, not by animus of the employer, but by beliefs a rational employer holds about the preferences of the workers from a particular group.[6] Relatedly, Stinebrickner, Stinebrickner, and Sullivan (2019) show that wage premium for attractive college graduates, traditionally interpreted as taste-based discrimination, is in fact due to the higher productivity of attractive workers.

The main thesis of this paper is that in many situations workers' motivation to work for employer depends on employer's social group. This argument has its roots in the psychology literature which postulates that group membership is an important component of one's social identity (Neuberg & Cottrell, 2006; Tajfel, 1970) and subconscious, knee-jerk negative associations can be triggered by exposure to members of the out-group (Bertrand, Chugh, & Mullainathan, 2005b; Cuddy, Fiske, & Glick, 2007). Within this paradigm it is natural that working for an out-group employer may challenge the sense of belonging or social identity of the worker, or analogously, working for the same-group employer give the sense of affiliation or purpose to the worker beyond the monetary compensation from working (Akerlof & Kranton, 2000, 2005). Thus a worker may exhibit positive social preferences and work more for employers belonging to the own-group and not so much for others.

There is growing amount of evidence that shows that workers' productivity may depend on the social identity of the employer. For example, Glover, Pallais, and Pariente (2017a) show that minority workers working for implicitly biased managers, under-perform as compared to when they work for unbiased managers. Glover et al. (2017a) attribute workers' under-performance to managers' implicit biases against workers' ethnicity; however, I argue that one plausible alternative interpretation is that workers under-provide effort just because they are not motivated enough when working for the biased employers.[7] Similarly, Plug, Webbink, and Martin (2014) shows that gay and lesbian workers shy away from prejudiced professions. While the authors

[5] See H. Fang and Moro (2011) for the nice review of models of discrimination.

[6] For example, in many empirical studies in economics, controlling for worker ability in a regression may not be enough to identify discrimination or its source, as the information on the identity-based motivation of the worker may be still missing from the regression.

[7] In fact, Glover et al. (2017a) find that workers underperform even though they do not face any *direct* discrimination or negative treatment from the implicitly-biased managers.

argue that this sorting of workers is consistent with employer and coworker taste-based discrimination, I argue that another plausible explanation is that workers may be selecting into those not because of employer and coworker prejudice but because they actually prefer to work in "tolerant" occupations. A more direct evidence comes from Oh (2019), in which the author shows that identity and social image concerns lead workers to opt-out of jobs that are associated with out-group in an experimental labor market in India.

It is obviously very difficult to identify the discrimination due to workers identity-based preferences.[8] Fortunately, there is a growing literature in economics which is directly testing for the evidence of employee-to-employer discrimination (see, for example, Abel (2019); Asad, Banerjee, and Bhattacharya (2020); Ayalew, Manian, and Sheth (2018); Chakraborty and Serra (2019); Grossman, Eckel, Komai, and Zhan (2019)). This literature generally finds that workers care about the identity of the employer and adjust their behavior accordingly. Taken together, these studies imply that worker preferences should be an important consideration when studying labor market discrimination.

This paper outlines a simple principal-agent model in which wages are determined by the worker's outside option and the motivation towards the employer group. Wage differential between the two groups arise if either the outside option of one group is different than the other, or the worker's motivation towards the employer's group varies. If the two groups are well represented on the demand side of labor market, then there are no wage differential on the basis of group-based motivation, there is complete segregation of the labor market, and each worker is paid equal to their marginal product. However, the outside options of two groups vary if one group is underrepresented on the demand side of the labor market. The underrepresentation of employers of one group reduces the outside option of workers of that group because they have to work in cross organizations where they are less productive. This productivity differential translates into difference in wages as workers are paid equal to their marginal product. The disadvantaged workers get lower wages not only from the cross type employers but also from the same type employer because their outside option is lowered.

This paper is particularly informed by the large body of economic literature on workers' intrinsic motivation and a desire to perform pro-social behavior in the workplace (see, for example, Besley and Ghatak (2005); Delfgaauw and Dur (2007); Ellingsen and Johannesson (2008); Frey (1997); Prendergast (2007)). This literature argues that workers care about the non-pecuniary aspect of their jobs and work hard if they feel motivated and find meaning in their work. This worker motivation is most relevant in settings in which complete labor contracts cannot be written or enforced, for it is in such environments that group identity

[8] Disentangling the differential labor market outcomes due to workers motivation from theories of discrimination would require data on productivity, on the preferences of workers, preferences of employers, and on perceptions of ability and identity-based motivation.

and social preferences are most likely to get activated.[9] For example, workers may work harder for employers by reciprocating disparately to "gifts" (Akerlof, 1982c) or by exhibiting disparate feelings of altruism depending on employer's group identity (Simon, 1993). See Akerlof and Kranton (2005); Benjamin, Choi, and Strickland (2010); Y. Chen and Li (2009) for understanding the role of identity and social preferences at workplace.

A related study to my paper is Craig and Fryer (2018), in which the authors build a theoretical model allowing for the possibility of statistical discrimination from both worker and employer side. In my paper, I abstract away from the possibility of statistical discrimination from the worker side and rather focus on preferences of the workers. To my knowledge, mine is the first study to theoretically explore the possibility of preference-based discrimination from the worker side.

This paper falls into a broad category of papers which study how differences in group preferences, rather than labor market discrimination, can lead to differential labor market outcomes (see Altonji and Blank (1999) for a nice review). This literature has mostly looked at group differences in preferences, such as a gendered preference over a profession, in explaining differential outcomes. My paper instead looks at difference in preference over the employer type as another potential source of discrimination. My paper is also related to the literature on assortative matching which argues that matching of agents on types may be efficiency enhancing because of within-group complementarities, see for example Becker (1973) and Durlauf and Seshadri (2003).

The rest of the paper proceeds as follows. In Section 2.3, I present a simple principal-agent model that outlines the preferences, payoffs, and the optimal contract for a worker and an employer. This section also study the equilibrium under the assumption of perfectly competitive labor market. In Section 2.4, I study the implications of the model on the interpretation of labor market discrimination and its sources. In Section 2.5, I conclude the paper by discussing the main insights of the paper.

2.3 Model

2.3.1 Environment

A "firm" consists of a risk-neutral employer denoted by j and a risk-neutral worker denoted by i. The employer needs the worker to carry out a task. The task's outcome x is stochastically determined by the

[9] A large literature in behavioral economics has established that other-regarding or social preferences play an important role in economic interactions when contracts cannot be perfectly defined or enforced. See Cooper and Kagel (2016) for a review of experimental literature on other-regarding preferences and Falk et al. (2018) for global evidence on existence of these preferences.

worker i's ability, a_i, and effort provided by worker i when working for employer j, e_{ij}, i.e.,

$$x = a_i + e_{ij} + \epsilon$$

where $\epsilon \sim N(0, 1/\tau_\epsilon)$ is a random shock with precision τ_ϵ which prevents an impartial third party (for contract enforcement) from knowing exactly the effort e_{ij} and ability a_i of the worker by observing the quality of the task. Higher variance allows for greater uncertainty in the determination of worker's effort and ability, conditional on quality of the task. The prior on ability, a_i, has a normal distribution with mean α_i and precision τ_a, i.e., $a_i \sim N(\alpha_i, 1/\tau_a)$. The belief about average ability, $\hat{\alpha}_i$, may be biased if $\hat{\alpha}_i \neq \alpha_i$. a_i and ϵ are independent. The firm and the worker share the prior belief about a_i, thus there is no asymmetric information and adverse selection. However, the effort, $e_{ij} \in \mathbb{R}^+$, is known only to the worker and hence is noncontractable. The employer and the worker observes x but the payment to the worker are not conditioned on the realizations of x.[10][11] The worker is paid a fixed wage w_{ij} before starting to work on the task. Effort is costly to workers with cost function given by $c(e_{ij}) = \frac{ce_{ij}^2}{2}$.

The employer and the worker can belong to one of the two types, i.e., $i, j \in \{B, W\}$. These types, exogenously assigned by the nature, represent a social identity of the employer and the worker. A social identity represents an association of an employer and a worker to a particular group such as race, gender, age, political affiliation, or any other identity. For this paper, I focus on racial identity and assume that these types represent association of the employer and the worker to either 'black', B, or 'white', W, racial groups. The racial identity of the employer makes workers value the return from the effort above/below any monetary income they receive from working. This follows a long tradition in labor economics where workers have preferences over their work environment - see, for example, Rosen (1986). This could also be based on the sense of identity of the worker à la Akerlof and Kranton (2000), where workers derive utility from behaving according to the norms of their identity. Therefore, workers when they work for same race employer derive extra motivation (utility) as compared to when they work for opposite group employer.

Following Besley and Ghatak (2005), I define worker i's social preference or motivation, θ_{ij}, towards the employer j's racial identity such that the worker of type B (type W) receives a non-pecuniary benefit of $\bar{\theta}$ if they work for an employer of type B (type W), and $\underline{\theta}$ if they work for an employer of type W (type B),

[10] This maybe due to high cost of monitoring performance (Weiss, 1990), fear of crowding out of effort from other tasks (in case of multitasking) which are difficult to monitor (Holmstrom & Milgrom, 1991), or fear of crowding out intrinsic motivation (Benabou & Tirole, 2003; Ellingsen & Johannesson, 2008; Frey, 1997).

[11] Introducing incentive payments do not change the qualitative implications of the model but make the analysis unnecessarily involved.

where $\bar{\theta} > \underline{\theta} \geq 0$, i.e.,

$$\theta_{ij} = \begin{cases} \underline{\theta} & \text{if } i \neq j \\ \bar{\theta} & \text{if } i = j \end{cases}$$

I assume, for now, that these non-pecuniary payoffs are contractible, and that they are independent of employer actions.[12] The overall payoff of a worker i who is matched with an employer j, getting a fixed payment of w_{ij} can be written as:

$$u_{ij} = w_{ij} + \theta_{ij} e_{ij} - \frac{c e_{ij}^2}{2}$$

The employer of type j receives a payoff of π_{ij} from the effort of worker i. I assume, for now, that employers do not derive non-pecuniary payoff from working with any type of workers, i.e., employers do not have taste bias against any group of workers. The employer's payoff from worker's effort can be written as:

$$\pi_{ij} = x - w_{ij}$$

The timing of the model is as follows;

1. The employer j observes worker type i and offers a wage (w_{ij}) to the worker which is paid regardless of worker's performance.

2. The worker i observes employer type j and accepts or rejects the contract. If the worker rejects the contract, he/she receives the outside option of $\bar{u}_i$. The reservation utility, $\bar{u}_i$, is the utility arising from the best outside option available to the worker. This could be a position in another firm or the value of leisure if the worker chooses not to work.

3. If the worker accepts the contract, he/she gets w_{ij} and then choose effort e_{ij}.

4. Nature draws ϵ, determining x.

The analysis of the model is in two steps. I first solve for the optimal contract for an exogenously given match of an employer of type j and a worker of type i. Second, I study matching of employers and workers where the reservation payoffs are endogenously determined.

[12] In reality, worker's motivation/social-preference towards the employer's group may be unobserved to employer and incorporating that asymmetry in information may reveal interesting insights.

2.3.2 Optimal Contracts

As a benchmark, consider the first best case where effort is contractible. This will result in effort being chosen to maximize the joint payoff of the manager and the worker. This effort level will depend on worker's social preferences and hence the employer-worker match. The contract offered to the worker, however, plays no allocative role in this case. Any value of w_{ij} such that the worker's expected payoff is at-least the reservation utility, $\bar{u}_i$, will work. It is straightforward to calculate that the first-best effort level is $\frac{1+\theta_{ij}}{c}$. The expected joint surplus from a given employer-worker match in this case is $\frac{(1+\theta_{ij})^2}{2c} + \alpha_i$.[13]

In the second best, effort is not contractible. The employer's optimal contracting problem now solves:

$$\max_{\{w_{ij}\}} \mathbb{E}[\pi_{ij}] = \max_{\{w_{ij}\}} \mathbb{E}\left[a_i + e_{ij} + u - w_{ij}\right] \tag{2.3.1}$$

subject to:

1. The participation constraint of the agent that:

$$\mathbb{E}[u_{ij}] = \mathbb{E}\left[w_{ij} + \theta_{ij} e_{ij} - \frac{c e_{ij}^2}{2}\right] \geq \bar{u}_i \tag{2.3.2}$$

2. The incentive compatibility constraint, which stipulates that the effort chosen by the worker maximizes the worker's private payoff given (w_{ij}):

$$e_{ij} = \arg \max_{e_{ij}} \mathbb{E}\left[w_{ij} + \theta_{ij} e_{ij} - \frac{c e_{ij}^2}{2}\right] \tag{2.3.3}$$

[13] The optimal contract for the manager solves the following problem when the effort is contractible.

$$\max_{e_{ij}, w_{ij}} \mathbb{E}\left[x - w_{ij}\right]$$

subject to the participation constraint of the worker

$$\mathbb{E}\left[w_{ij} + \theta_{ij} e_{ij} - \frac{c e_{ij}^2}{2}\right] \geq \bar{u}_i$$

and the incentive compatibility constraint

$$e_{ij} \in \arg\max_{e_{ij}} \mathbb{E}\left[w_{ij} + \theta_{ij} e_{ij} - \frac{c e_{ij}^2}{2}\right]$$

where expectations are taken over the distribution of a_i and ϵ. The problem is solved by first taking the effort as given $(e_{ij} = e_{ij}^{\star})$ and finding the optimal level of w_{ij} and then finding the optimal effort to maximize the profit function given the optimal payment scheme. Any value of w_{ij} that leaves the worker with exactly$\bar{u}_i$ will solve the above problem.

I restrict attention to the range of reservation payoffs for the worker in which the employer earns a non-negative payoff. The incentive-compatibility constraint can be simplified to:

$$e_{ij} = \frac{\theta_{ij}}{c}$$

The following proposition characterizes the optimal contract for a given employer-worker match. All proofs are presented in the Appendix.

Proposition 1. *The optimal contract* $(w^{\star}_{ij})$ *between an employer of type* j *and a worker of type* i *given a reservation payoff* $\bar{u}_i$ exists, and has the following features:

a) The wage is characterized by

$$w^{\star}_{ij} = \bar{u}_i - \frac{\theta^2_{ij}}{2c}$$

b) The optimal effort level is

$$e^{\star}_{ij} = \frac{\theta_{ij}}{c}$$

The first part of the proposition shows that the wage will be set such that the participation constraint of the worker binds. Any wage offer that does not satisfy the participation constraint of the worker will be rejected. Since a worker derive motivation from the effort and work is costly to worker, employer adjust the wage so that worker is compensated for the cost of effort and no rents are accrued to the worker. The reservation payoff plays an important role in setting w_{ij}, the higher the reservation payoff the higher the w_{ij}. Interestingly for a highly motivated worker ($\theta^2_{ij} > 2c\bar{u}_i$), wages can be negative (worker pay to work on the task).[14] It is clear that workers of the same type will be paid a lower wage than the workers of the opposite type for the same level of outside option. This is as though the same type worker is cheaper to the employer.

The second part characterize the resulting optimal effort. Because of the hidden effort, an employer cannot get the first-best level of effort $\left(\frac{1+\theta_{ij}}{c}\right)$. Note that in the absence of worker's social preferences, i.e., when $\theta_{ij} = 0$, workers do not provide any effort to the employer. This is because the wages are already paid and there is no monetary incentive to put effort (effort is otherwise costly).

The worker's social preference or motivation towards the manager plays an important role in the optimal contract. Given w_{ij}, a worker with higher social preference towards the employer group will provide higher effort. From the employer's point of view, θ_{ij} is sole reason for providing effort, and workers with higher social preferences towards employer's type provide higher effort at the optimum. Note that for a given employer-worker match, the beliefs about worker's ability do not effect the wage payment or the effort.

[14] This may be applicable, for example, in volunteer work jobs that require high motivation and do not pay.

I now present two corollaries of this proposition, which are useful in understanding its implication for optimal matching. These corollaries illustrate the importance of matching employers and workers.

Corollary 1. *Suppose that $\bar{u}_B = \bar{u}_W$, then effort is higher and the wage payment lower if the worker's type is the same as that of the employer.*

This can be seen from examining part a) of the Proposition 1 that the wage payment is decreasing in θ_{ij}. The wage has to be higher if $i \neq j$. Effort is higher when $i = j$, can be seen from observing part b) of the Proposition 1. Hence, this implies that organizations with same type of manager's and workers will have higher level of productivity holding the expected ability and the reservation payoff of the workers the same for both types.

Corollary 2. *Suppose that $\bar{u}_B = \bar{u}_W$, then effort and the wage payment are negatively correlated in a cross section of organizations.*

This follows directly from Corollary 1. Thus holding constant the reservation payoff; productivity and wages will be negatively correlated cross organizations. These corollaries highlight the cost of matching workers cross types and therefore, on the pure efficiency grounds, matching workers to the same type managers is optimal.

2.3.3 Competition

I now consider the outcomes when employers compete for workers. I do not model the competitive process explicitly but instead study the implications of stable matching. If the matching is not stable, i.e. an employer or a worker can do better by deviating from the contract, then I would expect re-matching to occur.

I follow Roth and Sotomayor (1989) to define a one-to-one matching function μ that takes as it's argument an employer or a worker and returns a match to a worker or an employer. Specifically, let $\mathcal{T}^\eta = \{\eta_B, \eta_W\}$ ($\mathcal{T}^\psi = \{\psi_B, \psi_W\}$) denote the set of types of employers (workers). The matching process can be described as $\mu : \mathcal{T}^\eta \cup \mathcal{T}^\psi \to \mathcal{T}^\eta \cup \mathcal{T}^\psi$ such that a) $\mu(\eta_j) \in \mathcal{T}^\psi \cup \eta_j$ for all $\eta_j \in \mathcal{T}^m$; b) $\mu(\psi_i) \in \mathcal{T}^\eta \cup \psi_i$ for all $\psi_i \in \mathcal{T}^\psi$; and c) $\mu(\eta_j) = \psi_i$ if and only if $\mu(\psi_i) = \eta_j$ for all $(\eta_j, \psi_i) \in \mathcal{T}^\eta \times \mathcal{T}^\psi$. If an employer (worker) is unmatched then $\mu(\eta_j) = \eta_j$ ($\mu(\psi_i) = \psi_i$).

Let n_j^η (n_i^ψ) denote the number of employers (workers) of type j (i) in the population. For simplicity, I assume that there are unlimited type W employers such that $n_W^\eta > n_W^\psi + n_B^\psi$. For type B employers, I first assume that $n_B^\eta > n_B^\psi$, i.e. there are as many type B employers as type B workers. I later allow for scarcity of type B employers such that type B employers are less than the type B workers, i.e. $n_B^\eta < n_B^\psi$. I

assume that the labor market is competitive (with no frictions) and therefore employers on the long side of the market gets none of the surplus.

From Corollary 1, it is clear that employers can generate more surplus by hiring workers of the same type as compared to the opposite type holding the beliefs about ability and worker's reservation payoff fixed. In other words, it is Pareto improving for an employer to hire a worker of her own type given that the employer believes the workers of both types are of equal ability. Since the worker's type (and therefore motivation) is known with certainty to the employer the employer set the $\theta_{ij} = \bar{\theta}$ in the optimal contract and therefore no more satisfy the participation constraint of the opposite type workers for which $\theta_{ij} = \underline{\theta}$. The following proposition summarizes the optimal contract in the presence of selection effects:

Proposition 2. *Suppose that $n_i^{\eta} > n_i^{\psi}$ for $i \in \{B, W\}$ (full employment of workers in both types), then the matching $\mu\{\psi_i\} = \eta_i$ is stable and the optimal contract $(w_{ij}^{\star})$ between an employer of type j and a worker of type i* has the following features:

a) The wage is characterized by

$$w_{ii}^{\star} = \hat{\alpha}_i + \frac{\bar{\theta}}{c}$$

b) The optimal effort level is

$$e_{ii}^{\star} = \frac{\bar{\theta}}{c}$$

This Proposition indicates complete segregation of workforce by type in equilibrium, i.e., white (black) workers work for white (black) employers. The wage differential between black and white workers can only arise if the beliefs about their ability are different, i.e., $\hat{\alpha}_B \neq \hat{\alpha}_W$, with higher expected ability workers getting higher wages in equilibrium. Given homogeneous beliefs about ability, $\hat{\alpha}_B = \hat{\alpha}_W$, there is no wage differential between black and white workers and each worker type provide effort based on their social preference towards the same race employer. The implications of Proposition 2 are analogous to Becker's model of employer's taste discrimination. In the Becker model, employer distaste creates an incentive for segregation and given enough non-prejudiced employers there is no wage differential. In my model, worker's intrinsic motivation (and resulting productivity differential) creates incentives for segregation and given enough black employers there is no wage-differential.

Now suppose there are not enough black employers to recruit all the black workers, $n_B^{\eta} < n_B^{\psi}$, and therefore some black workers end up working for white employers.[15] The following Proposition characterizes the equilibrium in this case.

[15] According to Korn Ferry (2019), since 1955, there have only been 15 Black CEOs at the helm of Fortune 500 companies, today, there are only four. Chakraborty and Serra (2019) finds that minority group leaders may select out of leadership roles for the fear of backlash by the workers.

Proposition 3. *Suppose that* $n_B^{\eta} < n_B^{\psi}$*, then the matching* $\mu\{\psi_B\} = \eta_j \ \forall \ j \in \{B, W\}$ *and* $\mu\{\psi_W\} = \eta_W$ *is stable and the optimal contract* $(w_{ij}^{\star})$ *between an employer of type* j *and a worker of type* i has the following features:

a) The wage is characterized by

$$w_{BW}^{\star} = \hat{\alpha}_B + \frac{\underline{\theta}}{c}$$

$$w_{BB}^{\star} = \hat{\alpha}_B + \frac{\underline{\theta}}{c} - \left(\frac{\bar{\theta}^2 - \underline{\theta}^2}{2c}\right)$$

$$w_{WW}^{\star} = \hat{\alpha}_W + \frac{\bar{\theta}}{c}$$

b) The optimal effort level is

$$e_{ii}^{\star} = \frac{\bar{\theta}}{c}$$

$$e_{BW}^{\star} = \frac{\underline{\theta}}{c}$$

This proposition shows that lower representation of blacks in the leadership positions will lead to wage differentials between black and white workers. Since black workers are unable to find employment with black employers, some of them work with white employers where their marginal product is lower because of lower intrinsic motivation. White employers pay the black workers according to their marginal product which is lower than white workers. The wage differential between black and white workers when working for white employers is given as

$$w_{WW}^{\star} - w_{BW}^{\star} = (\hat{\alpha}_W - \hat{\alpha}_B) + \frac{\bar{\theta} - \underline{\theta}}{c}$$

This wage differential arise whenever employers have differential beliefs about the ability of blacks and whites, and/or because of the differences in worker's motivation to work for a white employer. Given homogeneous beliefs about ability, black workers are paid lower than white workers by white employers. The following remark summarize this result

Remark 1. Given homogeneous beliefs about ability of workers, wage differential between blacks and whites when working for white employers represents the employer's rational response to the worker's differential social preferences towards the employer's race.

Interestingly, Proposition 3 shows that blacks are paid even lower wage when they work for black employers, i.e.,

$$w_{BW}^{\star} - w_{BB}^{\star} = \left(\frac{\bar{\theta}^2 - \underline{\theta}^2}{2c}\right) > 0$$

Technically this happens because black employers are in short supply which lowers the value of the outside option for the black worker. This allows black employers to pay less (and earn rents) to black workers and still bind the worker's participation constraint.[16] Intuitively, black workers are paid less because they prefer to work with black employers and are willing to take a wage cut. Note that the payoff of black worker from working for white or black employer is same even though the wage is different.

The important insight in this model is that the differences in wages may arise even in the absence of discrimination. Note that black and white workers are paid differential wages because of their differences in productivity for an employer. This insight has crucial implications for understanding of discrimination and its sources. The next section studies this in the context of models of discrimination.

2.4 Implications of the Model

I now study the implications of the insights develop in the last section on interpretation of discrimination and its sources in labor markets. Fundamentally, economists view labor market discrimination as arising in one of two ways, taste or statistical. Becker (1957) introduced the notion of *taste-based* discrimination postulating that discrimination exists because of prejudice/animus of the employers toward the workers from the disadvantaged group. Phelps (1972) and Arrow (1973) instead view discrimination as *statistical*, in which an employer, lacking information about the worker's productivity, forms belief about the worker's productivity based on worker's group identity using the aggregate distribution of worker's group traits. In this section, I incorporate the traditional employer taste-based and statistical discrimination in the model of Section 2.3 and study implications for labor market discrimination. I assume throughout that there are not enough black employers to recruit all the black workers and focus on white employers only.

To incorporate taste-based discrimination, I now assume that employers of type W derive negative utility, d, if they hire a worker of type B. Therefore the employer's payoff is now given as

$$\pi_{ij} = x - w_{ij} - \mathbb{1}_{i \neq j} d$$

where $d > 0$. As shown in Section 2.3, the worker i's effort is given by the i's motivation towards the employer j (θ_{ij}) which allows me to replace effort e with motivation θ_{ij} (subsuming c) and so the outcome x is given by

$$x = a_i + \theta_{ij} + \epsilon$$

[16] However, in the long run, this should mean that rents for black employers attract other blacks employers to enter the market until the rents are dissipated. The assumption here is that there are institutional constraints because of which it is not easy for black employers to enter the market and provide employment opportunities.

To make things more interesting I now assume that θ_{ij} is not known with certainty (unlike in Section 2.3) and instead employers and workers have common prior which are normally distributed with mean χ_{ij} and precision τ_θ, i.e., $\theta_{ij} \sim N(\chi_{ij}, 1/\tau_\theta)$. θ_{ij} is independently distributed from a_i and ϵ. The beliefs about average motivation of worker i towards employer j, $\hat{\chi}_{ij}$, may be biased if $\hat{\chi}_{ij} \neq \chi_{ij}$. The firm and the worker share the prior belief about θ_{ij}, thus there is no asymmetric information and adverse selection.

To introduce the notion of statistical discrimination, I follow Phelps (1972), and assume that the outcome x is unobserved and employer only observe the signal of the outcome, $s = x + \gamma$ where $\gamma \sim N(0, 1/\tau_\gamma)$. Employer use the observed signal s and the race of worker to extract information about the outcome of the task. In other words, employer derive an expression for $\hat{x}_i$ given observed signal for performance s and worker type i, i.e., $\hat{x}_i = \mathbb{E}[x|s, i]$.

Because a_i, θ_{ij} and ϵ are independent by assumption and a_i, θ_{ij}, and ϵ are joint normally distributed: $\begin{pmatrix} a_i \\ \theta_{ij} \\ \epsilon \end{pmatrix} \sim N\begin{pmatrix} 1/\tau_a & 0 & 0 \\ 0 & 1/\tau_\theta & 0 \\ 0 & 0 & 1/\tau_\epsilon \end{pmatrix}$ and because x is a linear combination of three normally distributed random variables, $x_i \sim N\left(\alpha_i + \chi_{ij}, 1/\tau_x\right)$ where $\tau_x = \frac{\tau_a \tau_\theta \tau_\epsilon}{\tau_\theta \tau_\epsilon + \tau_a \tau_\epsilon + \tau_a \tau_\theta}$. The employer combine the perceived distribution of the outcome with the observed signal s, which has distribution $s|x \sim N(x, 1/\tau_\gamma)$ and uses the Bayes rule to form the posterior belief about the outcome

$$\hat{x_i} \sim N\left(\frac{\tau_\gamma s + \tau_x \hat{\alpha}_i + \tau_x \hat{\chi}_{ij}}{\tau_\gamma + \tau_x}, \frac{1}{\tau_\gamma + \tau_x}\right)$$

Given these posterior beliefs and the competitive labor market, employer maximize the expected payoff by setting

$$w^\star_{ij} = \frac{\tau_\gamma s + \tau_x \hat{\alpha}_i + \tau_x \hat{\chi}_{ij}}{\tau_\gamma + \tau_x} - \mathbb{1}_{i \neq j} d$$

and the wage discrimination between blacks and whites is given as

$$w^\star_{WW} - w^\star_{BW} = \frac{\tau_x}{\tau_\gamma + \tau_x}(\hat{\alpha}_W - \hat{\alpha}_B) + \frac{\tau_x}{\tau_\gamma + \tau_x}(\hat{\chi}_{WW} - \hat{\chi}_{BW}) + d \tag{2.4.1}$$

Equation 2.4.1 provides various insights on the sources of wage discrimination between blacks and whites. Most importantly, the equation shows that the wage differential may arise because of: 1) differences in beliefs about the ability of blacks and whites, 2) differences in beliefs about motivation of workers to work for white employer, and/or 3) the distaste of employers towards the black workers. It is also clear that the

discrimination is decreasing in the precision of the signal τ_γ, i.e., if the outcome is observed with certainty ($\tau_\gamma \to \infty$) then there is no discrimination on the basis of beliefs and the only source of discrimination is due to the distaste of the employer.

The following proposition characterize one of the most important results of this paper

Proposition 4. *Suppose employers have same prior beliefs about the distributions of ability ($\hat{\alpha}_W = \hat{\alpha}_B$) and there is no distaste from hiring blacks ($d = 0$), then wage differential against blacks arises if employers have differential beliefs about motivation of workers towards whites $\hat{\chi}_{WW} > \hat{\chi}_{BW}$.*

This follows directly from the Equation 2.4.1 after substituting $\hat{\alpha}_W = \hat{\alpha}_B$ and $d = 0$. This proposition shows that wage discrimination against blacks may arise even in the absence of traditional explanation of ability-based statistical discrimination and taste-based discrimination. This highlights that to address the statistical discrimination, providing information about the true ability of workers alone may not be enough (as suggested by ability-based models of statistical discrimination) and rather one should also address the motivation of workers towards the opposite race.

Another important insight that follows from Equation 2.4.1 can be characterized as follows:

Proposition 5. *Suppose there is no distaste from hiring blacks ($d = 0$), then for any level of beliefs about ability, there exists a level of beliefs about motivation that yields an equivalent discrimination and vice versa.*

Proposition 5 shows that interpreting residual (from taste) discrimination as ability-based statistical discrimination may very well be discrimination due to differential beliefs about the motivation of the workers. I argue that distinction between ability-based statistical discrimination and motivation-based statistical discrimination is important from policy perspective and therefore one need to be careful in interpreting the source of beliefs in explaining discrimination. For example, ability-based statistical discrimination can be addressed by providing information about the true ability of the worker however addressing motivation-based statistical discrimination may require information beyond ability.

The distinction between the source of discrimination is arguably even more important when the traditional explanation may misinterpret statistical discrimination as taste-based discrimination. The following proposition characterize this result

Proposition 6. *Suppose employers have same prior beliefs about the distributions of ability ($\hat{\alpha}_W = \hat{\alpha}_B$), then discrimination against blacks arises if employers have differential belief about motivation $\hat{\chi}_{WW} > \hat{\chi}_{BW}$ or if employers have preference-based partiality towards whites.*

Proposition 6 can be easily seen from Equation 2.4.1 after substituting $\hat{\alpha}_W = \hat{\alpha}_B$. This proposition highlights an important consequence of ignoring motivation-based discrimination, i.e., after controlling for ability

based beliefs, ignoring motivation based discrimination may lead to misinterpretation of discrimination as taste-based, when in fact part of it may be motivation-based statistical discrimination. This proposition has important implications for identification of source of discrimination and highlights the fact that researchers need to be careful in interpreting source of discrimination as controlling for ability alone does not guarantee the identification of taste-based discrimination.

2.5 Conclusion

Discrimination in labor markets has received a lot of attention from economists and social scientists for many decades. There is a lot that we understand about discrimination and the reason for its existence and yet our understanding is limited. This paper attempts to deepen our understanding by exploring heretofore an ignored side of the labor market, i.e., the worker side. In this paper, I show why it is important to understand the worker side and what are the implications from ignoring it. I see the literature on intrinsic motivation and the literature on discrimination as naturally complementary and take a step towards bridging a gap between the two. This paper is a drop in the ocean of the issues that arise from the consideration of workers' identity-based social preferences.

It is highlighted that the worker-to-employer discrimination might be most relevant in settings where there is considerable degree of contact between the employer and the worker for social preferences to get activated. For example, Asad et al. (2020) finds that workers provide more effort for outgroup employers in an online environment where the social contact with the employer is minimal.

There are lessons for policy makers in this paper. For example, one of the implications is that policies aimed at encouraging entrepreneurship among blacks, such as providing non-discriminatory credit to black-owned firms, might be an effective strategy to reduce wage differential between blacks and whites, which is currently not in favor of blacks (Blanchflower, Levine, & Zimmerman, 2006). More generally, policy makers should consider the worker side of the market and employer's response to it to come up with more effective policies. For example, providing information about the ability of workers may not be enough to end statistical discrimination in the markets, rather one need to address the motivation of workers and employer's belief about it to truly address the issue. In this regard, future work should explore the implications of one-sided affirmative action policies and come up with policies which consider both sides of the market.

2.6 Appendix

Proof of Proposition 1

Proof. Substituting for $e_{ij} = \frac{\theta_{ij}}{c}$ using the incentive compatibility constraint, I can write the optimal contracting problem in Section 2.3.2 as:

$$\max_{\{w_{ij}\}} \mathbb{E}[\pi_{ij}] = \mathbb{E}[a_j + \frac{\theta_{ij}}{c} + \epsilon - w_{ij}]$$

subject to the participation constraints:

$$\mathbb{E}[u_{ij}] = \mathbb{E}[w_{ij} + \frac{{\theta_{ij}}^2}{2c} \geq \bar{u}_j]$$

This modified optimization problem involves one choice variables, w_{ij}, and one participation constraint. Since the wage does not effect the effort, employer would pay the minimum possible wage that will be accepted by the worker, i.e., an employer will chose the wage to bind the participation constraint of the worker. This gives;

$$w_{ij} = \bar{u}_j - \frac{\theta_{ij}^2}{2c}$$

To see if the optimal contract exists we need to make sure that the payoff of the employer and the worker are non-negative for each reservation payoff. The worker's expected payoff is exactly equal to its reservation payoff $\bar{u}_j$ which is non-negative since the lowest possible outside option for worker is to stay at home and get zero payoff. The employer's expected payoff is given by $\hat{\alpha}_j + \frac{\theta_{ij}}{c} + \frac{\theta_{ij}^2}{2c} - \bar{u}_j$. This is non-negative as long as $\bar{u}_j \leq \hat{\alpha}_j + \frac{\theta_{ij}}{c} + \frac{\theta_{ij}^2}{2c}$, which has to be true because otherwise the employer will not hire the worker. Note that our assumption is that reservation payoffs are such that both employer and manager make non-negative payoffs.

□

Proof of Proposition 2

Proof. Since the number of employers is greater than the number of workers, there are unmatched employers. Therefore, all employers must be earning zero profits, i.e., $\mathbb{E}[\pi_{ij}] = 0 \; \forall \; i, j$. This pins down the reservation payoff of the worker. From the proof of Proposition 1, we have

$$\mathbb{E}[\pi_{ij}] = \alpha_j + \frac{\theta_{ij}}{c} + \frac{\theta_{ij}^2}{2c} - \bar{u}_j$$

Equating this to zero we get

$$\bar{u}_j = \hat{\alpha}_j + \frac{\theta_{ij}}{c} + \frac{\theta_{ij}^2}{2c}$$

Substituting this in the optimal contract in Proposition 1, we get

$$w_{ij} = \hat{\alpha}_j + \frac{\theta_{ij}}{c}$$

and

$$e_{ij}^{\star} = \frac{\theta_{ij}}{c}$$

To prove that $\mu\{\psi_j\} = \eta_j$ is the stable match, suppose that it is not, i.e., $\mu(\psi_j) = \eta_i$ for $i \neq j$ is the stable match. Then, it must be that a worker of type j would not wish to switch to an employer of type $i = j$ from an employer of type $i \neq j$. This implies that the expected payoff of worker j from working with employer $i \neq j$ is greater than the expected payoff from working for an employer $j = i$, i.e., $\mathbb{E}[u_{ij}] > \mathbb{E}[u_{jj}]$ for $i \neq j$. From the proof of Proposition 1, we have the expected payoff of worker given by

$$\mathbb{E}[u_{ij}] = \bar{u}_j$$

which from above is

$$\bar{u}_j = \hat{\alpha}_j + \frac{\theta_{ij}}{c} + \frac{\theta_{ij}^2}{2c}$$

For $i \neq j$, we have

$$\mathbb{E}[u_{ij}] = \hat{\alpha}_j + \frac{\underline{\theta}}{c} + \frac{\underline{\theta}^2}{2c}$$

and

$$\mathbb{E}[u_{jj}] = \hat{\alpha}_j + \frac{\bar{\theta}}{c} + \frac{\bar{\theta}^2}{2c}$$

Taking the difference of the two

$$\begin{aligned}\mathbb{E}[u_{jj}] - \mathbb{E}[u_{ij}] &= \left[\hat{\alpha}_j + \frac{\bar{\theta}}{c} + \frac{\bar{\theta}^2}{2c}\right] - \left[\hat{\alpha}_j + \underline{\theta} + \frac{\underline{\theta}^2}{2c}\right] \\ &= \left(\frac{\bar{\theta}}{c} - \frac{\underline{\theta}}{c}\right) + \left(\frac{\bar{\theta}^2}{2c} - \frac{\underline{\theta}^2}{2c}\right)\end{aligned}$$

Which is positive since $\bar{\theta} > \underline{\theta}$. This is a contradiction and worker will be better off by switching to the same type employer. □

Proof of Proposition 3

Proof. Since the number of type B workers is greater than the type B employers, some type B workers are not able to find employment with type B employers, and end up working for type W employers. Since there are large number of white employers, white employers earn zero expected payoff from hiring type B worker and therefore type B worker expected payoff when working for the white employer is given as

$$\mathbb{E}[u_{BW}] = \hat{\alpha}_B + \frac{\underline{\theta}}{c} + \frac{\underline{\theta}^2}{2c}$$

Since this is the minimum payoff that a worker can obtain, it is therefore the reservation payoff of the type B worker. Substituting the value of reservation payoff in the optimal contract in Proposition 1, we get

$$w^{\star}_{BW} = \hat{\alpha}_B + \frac{\underline{\theta}}{c}$$

and

$$w^{\star}_{BB} = \hat{\alpha}_B + \frac{\underline{\theta}}{c} - \left(\frac{\bar{\theta}^2 - \underline{\theta}^2}{2c}\right)$$

For type W workers the outside option is same as shown in proof of Proposition 2. Therefore, there wages remain the same as in Proposition 2.

To prove that $\mu\{\psi_B\} = \eta_W$ is stable, it must be that a worker of type B would not wish to switch to an employer of type B from an employer of type W. This implies that the expected payoff of worker B from working with employer W is at-least as great as the expected payoff from working for an employer B, i.e., $\mathbb{E}[u_{BW}] \geq \mathbb{E}[u_{BB}]$. From above, we have the expected payoff of worker given by

$$\mathbb{E}[u_{BW}] = \hat{\alpha}_B + \frac{\underline{\theta}}{c} + \frac{\underline{\theta}^2}{2c}$$

and from the proof of Proposition 1

$$\begin{aligned}\mathbb{E}[u_{BB}] &= w_{BB} + \frac{\bar{\theta}^2}{2c} \\ &= \hat{\alpha}_B + \frac{\underline{\theta}}{c} + \frac{\underline{\theta}^2}{2c} \\ &= \mathbb{E}[u_{BW}]\end{aligned}$$

Therefore a worker of type B has no incentive to switch to employer of type B. Analogously, if the type B worker is working with type B employer, there is no incentive to switch to type W employer implying that $\mu(\psi_B) = \eta_i$ is stable.

Type W workers matching stability follows the same argument as shown in proof of Proposition 2 and is therefore omitted. □

Proof of Proposition 5

Proof. Suppose the employer has differential beliefs about motivation such that $\hat{\chi}_{WW} > \hat{\chi}_{BW}$, but same beliefs about ability, $\hat{\alpha}_W = \hat{\alpha}_B = \hat{\alpha}$. Then substituting these and $d = 0$ in Equation 2.4.1 I get

$$w^{\star}_{WW} - w^{\star}_{BW} = \frac{\tau_x}{\tau_\gamma + \tau_x}(\hat{\chi}_{WW} - \hat{\chi}_{BW})$$

Setting $\hat{\chi}_{WW} = \hat{\chi}_{BW} = \hat{\chi}$, $\hat{\alpha}_W = \hat{\chi}_{WW}$ and $\hat{\alpha}_B = \hat{\chi}_{BW}$ in 2.4.1 will yield equivalent discrimination. The proof in the other direction is analogous. □

References

Abel, M. (2019). *Do Workers Discriminate against Female Bosses?*

Akerlof, G. A. (1982c). Labor Contracts as Partial Gift Exchange. *Quarterly Journal of Economics*, *97*(4), 543–569.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and Identity. *Quarterly Journal of Economics*, 715–753.

Akerlof, G. A., & Kranton, R. E. (2005). Identity and the Economics of Organizations. *Journal of Economic Perspectives*, *19*(1), 9–32.

Altonji, J. G., & Blank, R. M. (1999). Chapter 48 Race and gender in the labor market. *Handbook of Labor Economics*, *3 PART*(3), 3143–3259.

Arrow, K. J. (1973). The Theory of Discrimination. In *Discrimination in labor markets* (pp. 3–33).

Asad, S. A., Banerjee, R., & Bhattacharya, J. (2020). *Do workers discriminate against their out-group employers? Evidence from an online platform economy.*

Ayalew, S., Manian, S., & Sheth, K. (2018). *Discrimination from Below: Experimental Evidence on Female Leadership in Ethiopia.*

Becker, G. S. (1957). *The economics of discrimination.* University of Chicago Press.

Becker, G. S. (1973). A Theory of Marriage: Part I. *Journal of Political Economy*(4), 813–846.

Benabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, *70*(3), 489–520.

Benjamin, D. J., Choi, J. J., & Strickland, A. J. (2010). Social Identity and Preferences. *American Economic Review*, *100*(September), 1913–1928.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005b). Implicit Discrimination. *American Economic Review*, *95*(2), 94–98.

Bertrand, M., & Duflo, E. (2017). Field Experiments on Discrimination. In *Handbook of economic field experiments* (Vol. 1, pp. 309–393). North-Holland.

Besley, T., & Ghatak, M. (2005). Competition and Incentives with Motivated Agents. *American Economic Review*, *95*(3), 616–636.

Blanchflower, D. G., Levine, P. B., & Zimmerman, D. G. (2006). Discrimination in the Small-Business Credit Market. *Review of Economics and Statistics*, *85*(4), 930–943.

Chakraborty, P., & Serra, D. (2019). *Gender differences in top leadership roles: Does worker backlash matter?*

Charles, K. K., & Guryan, J. (2011). Studying discrimination: Fundamental challenges and recent progress. *Annu. Rev. Econ.*, *3*(1), 479–511.

Chen, Y., & Li, S. X. (2009). Group Identity and Social Preferences. *The American Economic Review*, *99*(1), 431–457.

Cooper, D. J., & Kagel, J. H. (2016). Other-Regarding Preferences: A Selective Survey of Experimental Results. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics* (chap. 4).

Craig, A. C., & Fryer, R. G. (2018). *Complementary Bias : A Model of Two-Sided Statistical Discrimination.*

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, 631–648.

Delfgaauw, J., & Dur, R. (2007). Signaling and screening of workers' motivation. *Journal of Economic Behavior and Organization*, *62*(4), 605–624.

Durlauf, S. N., & Seshadri, A. (2003). Is assortative matching efficient? *Economic Theory*, 475–493.

Ellingsen, T., & Johannesson, M. (2008). Pride and Prejudice: The Human Side of Incentive Theory. *American Economic Review*, *98*(3), 990–1008.

Falk, A., Becker, A., Thomas, D., Enke, B., Huffman, D., & Sunde, U. (2018). Global Evidence on Economic Preferences. *Quarterly Journal of Economics*, *133*(4), 1645–1692.

Fang, H., & Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In *Handbook of social economics* (Vol. 1A, pp. 133–200).

Frey, B. S. (1997). *Not Just for the Money: An Economic Theory of Personal Motivation.* Edward Elgar Pub.

Glover, D., Pallais, A., & Pariente, W. (2017a). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. *Quarterly Journal of Economics*, 1219–1260.

Grossman, P. J., Eckel, C. C., Komai, M., & Zhan, W. (2019). It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior and Organization*, *161*, 197–215.

Holmstrom, B., & Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, *7*, 24–52.

Kahn-Lang, A. (2018). *Missing Black Men? The Impact of Under-Reporting on Estimates of Black Male Labor Market Outcomes.*

Korn Ferry. (2019). *The Black P&L Leader: Insights and Lessons from Senior Black P&L Leaders in Corporate America* (Tech. Rep.). Korn Ferry.

Lang, K., & Kahn-Lang, A. (2020). Race Discrimination: An Economic Perspective. *Journal of Economic Perspectives*, *34*(2), 68–89.

Levine, R., Levkov, A., & Rubinstein, Y. (2008). *Racial discrimination and competition.*

Neuberg, S. L., & Cottrell, C. A. (2006). Evolutionary Bases of Prejudices. *Evolution and social psychology*, 163–187.

Oh, S. (2019). *Does Identity Affect Labor Supply?*

Phelps, E. S. (1972). The Statistical theory of Racism and Sexism. *American Economic Review*, *62*(4), 659–661.

Plug, E., Webbink, D., & Martin, N. (2014). Sexual orientation, prejudice, and segregation. *Journal of Labor Economics*, *32*(1), 123–159.

Prendergast, C. (2007). The Motivation and Bias of Bureaucrats. *American Economic Review*, *97*(1), 180–196.

Rich, J. (2014). *What Do Field Experiments of Discrimination in Markets Tell Us? A Meta Analysis of Studies Conducted since 2000.*

Rosen, S. (1986). The theory of equalizing differences. In *Handbook of labor economics* (pp. 641–692).

Roth, A. E., & Sotomayor, M. (1989). The college admissions problem revisited. *Econometrica*, *57*, 559–570.

Simon, H. A. (1993). Altruism and Economics. *American Economic Review: Papers & Proceedings*, *83*(2), 156–161.

Stinebrickner, R., Stinebrickner, T., & Sullivan, P. (2019). Beauty, job tasks, and wages: A new conclusion about employer taste-based discrimination. *Review of Economics and Statistics2*, *101*(4), 602–615.

Tajfel, H. (1970). Experiments in Intergroup Discrimination. *Scientific American*, 96–103.

Weiss, A. (1990). *Efficiency Wages: Models of Unemployment, Layoffs, and Wage Dispersion.* Princeton University Press.

CHAPTER 3. DO WORKERS DISCRIMINATE AGAINST THEIR OUT-GROUP EMPLOYERS?

Modified from a manuscript to be submitted to the Journal of Labor Economics.

Sher Afghan Asad, Iowa State University
Ritwik Banerjee, IIM, Bangalore
Joydeep Bhattacharya, Iowa State University

3.1 Abstract

We study possible *worker-to-employer* discrimination manifested via social preferences in an online labor market. Specifically, we ask, do workers exhibit positive social preferences for an out-race employer relative to an otherwise-identical, own-race one? We run a well-powered, model-based experiment wherein we recruit 6,000 workers from Amazon's M-Turk platform for a real-effort task and randomly (and unobtrusively) reveal to them the racial identity of their non-fictitious employer. Strikingly, we find strong evidence of race-based altruism – white workers, even when they do not benefit personally, work relatively harder to generate more income for black employers. Our results suggest the possibility that pro-social behavior of whites toward blacks, atypical in traditional labor markets, may emerge in the online economy where associative (dis)taste is naturally muted due to limited social contact.

Keywords: Discrimination; Worker-to-Employer; Social Preferences; Taste-based discrimination; Online Economy; Mechanical Turk; Structural Behavioral Economics.

JEL Codes: J71,D91, C93

3.2 Introduction

By construction, *Homo economicus* is self-interested and only takes actions that maximize his/her payoffs. By way of contrast, *Homo behavioralis*, in addition to being self-interested is also endowed with social preferences, a concern for how his/her actions affect the payoffs of others. These "others" could belong to his in-group, a group he identifies with and whose membership gives him a sense of belonging. Everyone else, by definition, is in his out-group. *Homo behavioralis* may harbor negative social preferences urging him to

We thank, without implicating, Martin Abel, Mackenzie Alston, Otavio Bartalotti, Michael Best, Kerwin Charles, Stefano DellaVigna, Catherine Eckel, Betsy Hoffman, Alex Imas, Peter Orazem, Devin Pope, Joshua Rosenbloom, Arka Roy Chaudhuri, and participants at the 15th ISI-Delhi Annual Conference, the Advances with Field Experiments (AFE, 2019) conference in Chicago, and the National Economic Association 2020 conference in San Diego for their invaluable input. We thank Bernard Fay and Alejandro Ruiz Ortega for excellent research assistance. We also acknowledge partial funding from the Russell Sage Foundation and from the Economics Department at Iowa State University. This research was approved by the IRB at Iowa State University, and was registered on the AEA RCT registry, ref. AEARCTR-0003885.

discriminate against others; or the preferences could be positive and take the form of prosocial behavior – actions taken with an intent to benefit others with no expectation of personal benefit.

This paper is aimed at detecting evidence of positive or negative social preferences within the context of labor markets. The experimental setting is an U.S. based online platform labor market and group identity is assumed to be racial in origin. Within this environment, we ask, is there evidence that whites systematically treat blacks differently from how they treat fellow whites? We depart from a half-century of research in labor economics that views this issue largely as unidirectional, emanating from employers and directed toward their employees.[1] Instead, we ask, is there evidence that white *workers* in the online economy treat their black *employers* better or worse than how they treat their otherwise-identical, white employers?

A series of questions come up right away. Why is it interesting to study discrimination or pro-social behavior of workers toward employers? Is there any evidence of this? And, why the online economy? We take these up one by one. That workers may treat their out-race employers differently may, at first glance, seem implausible; after all, it is mostly bosses who get to frame labor contracts and surely within the bounds of such contracts there cannot be much room left for workers to mistreat out-group bosses. Our view is that this first-pass line of thinking is limited. While admittedly it is easier for bosses to maltreat out-group workers, the latter are also keenly aware that the effort they put in, the diligence or care they show on the job, crucially affects the bottomline of their bosses. Moreover, as is well known, labor contracts are often "incomplete": they leave workers a considerable degree of discretion over work effort. It is therefore conceivable that a worker with substantial leeway over effort makes very different effort choices reflecting his underlying differential social preferences. For instance, a black worker may choose to work harder for a black boss because of his desire to a) see his boss succeed even if it does not benefit him personally (**altruism** à la Simon (1993)), and b) return any respect or kindness he receives from his boss (**reciprocity** à la Akerlof (1982b)).

Second, there is important evidence that workers care about the social identity of their bosses and differentially perform for in versus out-group employers. Sundstrom (1994), focusing on U.S. urban labor markets 1910-1950, notes "one of the most widely noted rules of the southern labor market was that blacks were not to supervise whites...[because it] would plainly invert the appropriate hierarchy" which meant "blacks were generally absent from supervisory positions". White employees simply did not wish to receive orders from (or work under) black supervisors. More recently, Glover, Pallais, and Pariente (2017a) study whether discriminatory beliefs held by bosses directly affect minority workers' job performance in a real-world workplace. They investigate the performance of cashiers in a French grocery store chain, and find when "minority cashiers, but not majority cashiers, are scheduled to work with managers who are biased

[1] See Charles and Guryan (2011), Rich (2014), Bertrand and Duflo (2017), and Neumark (2018) for a review of this literature.

(as determined by an Implicit Association Test), they are absent more often, spend less time at work, scan items more slowly, and take more time between customers." The upshot is, workers *do* adjust their effort based on the social identity of their bosses, and may perform better when paired with own-group managers than out-group ones.[2]

And why study this question in the confines of the online economy? To be clear, an online labor market platform economy is one where independent workers are paid by the gig (i.e., for a task or a project) as opposed to the traditional economy where workers are paid a salary or hourly wage as part of a contract. One important distinction is that in the online economy, particularly of the digital-platform type, there is little scope for familiarity or closeness or repeated interactions between the employer and the employee; hence, associative distaste or liking is unlikely to be activated.[3] This means, if we are to detect any race-based differences in social preferences (altruism or reciprocity) in our online economy setting, it will not be because of associative distaste or liking. Another critical difference is the vastly dissimilar "power dynamic" between worker and boss. In an online economy, workers retain a lot of power in the worker-employer relationship: they may shirk under a particular employer or easily switch employers without losing much "employment rent".[4] This new power dynamic makes the online economy an ideal setting to study *worker-to-boss* discrimination, much more so than the conventional labor market setting.[5,6]

To the end of answering our research question, we run a well-powered, AEA preregistered, model-based experiment using 6,000 white subjects from one of the largest online economy platforms: Amazon's Mechanical Turk (M-Turk).[7] Specifically, our experimental design uses U.S. based subjects from M-Turk (recruited as "workers") and black and white student subjects (recruited as "employers") from a major U.S. public university. The interaction between a worker and an employer is kept one-shot, as is typical in the online economy, so that confounding reputation effects (of the kind that naturally emerge in Glover et al. (2017a)) do not

[2]Oh (2019) finds that 43% of Indian workers "refuse to spend ten minutes working on tasks associated with other castes, even when offered ten times their daily wage" suggesting the important role of social identity in determining work-related decisions.

[3]While our work is focused on an online labor market, others such as TaskRabbit offer tasks situated in the physical world and cover household errands and skilled tasks such as minor home repairs, assembling Ikea furniture, where the scope for more interaction between worker and boss, and hence, more associative (dis) taste, is clearly higher.

[4]After all, a typical Uber driver (or a M-Turker), each a worker, may work for ten "employers" in a day and ten different ones the very next day!

[5]Allport (1954) classic *The Nature of Prejudice*, (Chapter 16 'The Effect of Contact') argued for bringing members of different groups together in face-to-face encounters to reduce inter-group hostility. Significantly, he was of the view that direct inter-group contact would effectively reduce out-group prejudice if it involved equal status among the participants. We posit that the online economy allows the worker and the employer to be of "equal status" and that, in and of itself, may reduce inter-group hostility even when no direct contact à la Allport is initiated.

[6]There are ancillary reasons why our focus on the online economy is pertinent. The argument is often made that blacks, often the victim of discrimination in conventional labor markets, would gravitate to the online economy because of reduced expectations of discrimination in the latter. We would want to know, are those expectations likely to be satisfied? Also, other than Cook, Diamond, Hall, List, and Oyer (2019), it is not known whether long-established routes of discrimination researched on traditional labor markets with conventional worker-boss power dynamics will continue to operate in the dawning online economy.

[7]Roughly 50% of M-Turkers are from the United States. Based on 2015 data, about 77% are non-Hispanic white and only 6% are non-Hispanic black (Hitlin, 2016). The results reported below are for U.S.-based white workers, by far the vast majority of workers on M-Turk and in our sample.

enter. In the experiment, workers engage in a real-effort task for a pre-assigned, non-fictitious employer. The real-effort task (unlike monetized costs in studies such as Charness, Rigotti, and Rustichini (2007); Fershtman and Gneezy (2001)) entails a real utility cost of effort because it requires a worker to alternately press the 'a' and 'b' buttons on a keyboard for up to 10 minutes.[8] Workers do not get to select their employer but are free to decide how much effort to provide on the task (an 'incomplete contract' environment).[9] The worker's performance is measured by the number of times the buttons are alternately pressed, and the worker is informed (truthfully) of the payoff the employer will receive due to the worker's performance. Race-dependent social preferences are potentially activated in some treatments by unobtrusively revealing the employer's race to the matched worker.

The design is tightly connected to a simple structural model à la DellaVigna, List, Malmendier, and Rao (2016), in which workers have race-dependent social preferences towards their employer and maximize utility from the provision of costly effort. Inspired by Doleac and Stein (2013b), we take the approach of revealing race indirectly via the revelation of skin color and voice: employer-subjects are videotaped while they read off a script explaining and demonstrating the task for the workers. The camera placement only captures the hand of the employer along with the movement of the fingers alternating 'a' and 'b' button presses. Other identifiers, such as the face, are not revealed. This allows us to reveal or conceal race without sacrificing either privacy or anonymity. In the neutral treatments, gloves and other clothing hide the skin entirely. The worker is aware of being matched to an employer but is unaware of any identity clues. We make every effort to check that race, when revealed, is correctly perceived. In the experiment, we introduce a total of ten treatment variations. In the first three, we vary the piece rate with an aim to identify and estimate the cost-of-effort function. Here, the worker is not given any information about the existence of (non-existent) employer; any earnings from his/her effort choices go entirely to the worker. The next set of three treatments aim to a) detect the baseline level of altruism towards the hidden race of the employer (altruism neutral) and b) estimate race-specific altruism towards the revealed race of the employer (altruism black and altruism white). The final treatments are designed to a) detect the baseline level of reciprocity towards the hidden race of the employer (reciprocity neutral) and b) estimate the race-specific variations in

[8]The task is admittedly artificial in the sense that workers, in reality, do not routinely engage in such meaningless tasks. We offer three arguments for choosing such a task. First, we wanted the task to not require any special ability on the part of workers, for in that case, our results would be tainted by the unobservability of the underlying ability distribution attached to the task. Second, the task is exactly the one used in DellaVigna and Pope (2018) thereby facilitating comparisons across our paper and theirs, more so because they use workers from M-Turk (we restrict participation to U.S. workers, they don't). And third, however meaningless the task may be, it requires substantial focus and effort, both of which contribute to measurable actual earnings.

[9]Traditionally, discrimination in labor markets is understood to arise in two main ways. Becker (1957) introduced the notion of taste-based discrimination postulating that discrimination exists because of a prejudice/animus towards the members of the disadvantaged group. On the other hand, Phelps (1972) theorized that discrimination might be statistical – an employer, lacking information about a job-seeker's productivity, forms beliefs about it based on the person's group identity and the aggregate productivity distribution of the group to which the person belongs. In our experiment, the employers do not get to make any strategic choices (such as wage offers, payments, minutes of work, work times, etc.). This eliminates most channels for statistical discrimination by workers.

reciprocity towards the revealed race of the employer (reciprocity black and reciprocity white). Thus, the ten treatments help us identify the cost-of-effort function and social-preference parameters (altruism and reciprocity) of the structural model separately for neutral (hidden race), black, and white employers.

Our findings reported in terms of average effort by white workers are as follows. First, not surprisingly, incentives via piece rates have a strong, statistically-significant effect on effort. This observation lends credence to the idea that the MTurk population is generally representative of a typical labor force; MTurkers, like most workers, work harder when they receive higher wages. Second, as in DellaVigna and Pope (2018), we detect statistically significant evidence for altruism: workers put more effort when they know their work benefits an employer of unknown race ("altruism-neutral treatment") as compared to the piece rate 0-cent treatment where neither the worker nor the employer earns any payoff attributable to worker effort. Parenthetically, there is no evidence of reciprocity.

Strikingly, white workers are significantly more altruistic towards black employers than white employers – categorically, they do not discriminate against their black employers. In addition to being statistically significant at the 2% level, the difference in effort provision is non-trivial. To see this, consider a baseline level of altruism, defined as the differential effort provided by white workers knowing their effort enhances the payoff of an unknown race employer versus their effort when the piece rate is 0-cent and no employer exists. Our results indicate that the differential effort by white workers knowing their effort enhances the payoff of a white vs. a black employer is 75% of this baseline. Also, the differential effort by workers knowing their effort enhances the payoff of a black vs. an unknown-race employer is 45% of the baseline. The structural estimation exercise also reveals that black employers get 5% more effort than white employers at a 0-piece rate. Collectively, these represent persuasive evidence of pro-social behavior by whites toward black employers.

What explains this pro-social behavior? Is it racial heterophily? Is it "white guilt"?[10] The short answer is, we do not know. We did not collect data from M-Turk workers on any measure of racial bias such as the Implicit Association Test (IAT).[11] . In an attempt to understand heterogeneities in the response, we sliced the data based on the results of our demographic survey. We combined IAT data from Project Implicit with county-level knowledge of worker residence. We find that the pro-social response towards black employers

[10] An alternative explanation for why whites worked harder for blacks relates to statistical discrimination related to beliefs about differential standards for white/ black employers? For example, one could argue that if white workers think black employers have higher standards than white employers to approve a task, they may put in more effort. While this explanation has a lot of merit in reality, our design is such that employers do not have to approve a task for the worker to be paid. As soon as a worker "returns the HIT", they are automatically paid, irrespective of whether they worked hard or not as deemed by some unknown, employer-specific standard.

[11] Perhaps the most well known (though problematic) measure of racial bias is the Implicit Association Test (IAT) which measures the "strength of association between categories such as European American versus African-American and words such as joy, laughter, and happy versus hurt, evil, and awful that represent categories of good versus bad." Upwards of 80% of whites in nationally representative American samples have shown an implicit preference for whites over blacks (Triplett, 2012).

is partially driven by workers from areas with low implicit bias against blacks. Peeking further, we find if we split the IAT data into two halves (top and bottom), the pro-black altruism is highly significant for workers in the bottom half – those residing in the "least racist" counties – and is insignificant for those in the top half.[12] We also test (albeit, crudely) and reject the hypothesis that the differentially altruistic response toward black employers is driven by worker beliefs about the income status of their employers. We find workers who are self-declared Republicans and Independents exert significantly more effort for their black employers as compared to Democrats.

In terms of the value-added to the literature, our primary contribution is to showcase the importance of looking at the worker-to-employer social preference angle. Our finding is interesting because it raises the possibility that positive social preference toward blacks, rarely detected in traditional labor markets, may emerge in environments such as the online platform economy where *associative distaste* is naturally muted. (Of course, long distance racial animus, or a desire to see an out-group person suffer losses, may still prevail.) Bear in mind, ours is a well-powered, AEA pre-registered experiment which would have detected preference-based discrimination had it existed on the M-Turk platform; the fact we don't is encouraging, seeing how the online economy is expanding (Katz & Krueger, 2019). Further, it is oft-repeated that the relative lack of success of black-owned businesses or the diminished presence of blacks in leadership positions in the United States is a major concern among policy makers; more so, because "business ownership has historically been a route of economic advancement for disadvantaged groups" (Fairlie & Robb, 2007). Do entrepreneurial blacks shy away from business because they rationally fear discrimination by majority white workers? Our study does not document such fears on the part of black employers but offers some reason to question those fears in online-platform economies if they exist. Curiously, our finding also shuts down another line of thinking connected to the issue of anticipation of discrimination. There is some evidence that establish that the employer-to-employee discrimination is taste-based (see, for example, (Charles & Guryan, 2008)). What if it is being miss-classified? What if an employer discriminates against his out-race workers because he rationally believes/anticipates being discriminated against by them? In that case, the employer-to-employee discrimination ought to be characterized as statistical. Within the confines of our environment, our finding that workers do not discriminate against their out-race employers essentially shuts down any rational expectation of bias an employer may have. Incorrect beliefs may persist, though (Bohren, Haggag, Imas, & Pope, 2019).

[12]It is tempting to draw conclusions about "white guilt", a supposedly collective guilt felt by whites for their group's actions toward blacks, not necessarily for their own actions. As Chudy, Piston, and Shipper (2019) point out "...whites who hold collective guilt acknowledge that their group is responsible for black suffering and that the inter-group relationship needs to be repaired." Just because someone lives in a county where an average person registers low animus toward African Americans in an IAT test does not mean such people will wish to do something to repair the aforementioned inter-group relationship. In our case, though, unlike research that relies on survey-based measures of white guilt, we are able to detect evidence of whites doing *something extra for blacks even when they do not need to.*

Our research is related to an emerging literature in economics studying discrimination by subordinates (Abel, 2019; Ayalew, Manian, & Sheth, 2018; Chakraborty & Serra, 2019; Grossman, Eckel, Komai, & Zhan, 2019). This literature focuses on gender as group identity and mostly finds belief-based discrimination against female leaders. Another study on Amazon's Mechanical Turk by Abel (2019) finds that workers do not discriminate in effort choices when they work for women leaders, even though the feedback from them is perceived as being less pleasant than from a male leader. Ours is the first to investigate the possibility of race-based discrimination by subordinates in the U.S. Evidence from Benson, Board, and Meyer-ter Vehn (2019) suggests that workers' performance is influenced by the social identity of their boss. They chalk it to the fact that bosses can better screen applicants from their own race. Our study shuts down this "selection effect" and yet finds no evidence of race-based discrimination by workers. Our result, along with that in Abel (2019), reaffirms our conclusion that worker-to-boss discrimination is less likely to elicit itself in an online economy.

The rest of the paper proceeds as follows. In Section 2, we present the model of behavior and produce the treatments to identify the parameters of interest. In Section 3, we present the experiment design. Section 4 summarizes the data. In Section 5, we present the results followed by structural estimation in Section 6; concluding remarks are in Section 7.

3.3 Model and Treatments

In this section, we present the model of behavior that is used to design the experiment. The model explains a worker's effort choice given the monetary and non-monetary incentives and costs of working for an employer. Our design is inspired by DellaVigna et al. (2016) modified to permit discrimination from the workers' side. In the setup, workers choose how much effort to provide on a real-effort task.

A risk-neutral worker, working for an employer j, $j \in \{Neutral, Black, White\}$, receives utility[13]

$$U_j \equiv (\ F + (s + \rho_j \mathbb{1}_{Gift} + \alpha_j v + p)e_j - c(e_j)). \tag{3.3.1}$$

Here, e_j is the number of points (on the button-pressing task) scored by the worker when working for an employer j, F is the fixed participation fee he receives , and s captures a sense of duty, norm, intrinsic motivation, and competitiveness of the worker towards the task and is independent of the employer. ρ_j is the reciprocity parameter per unit of effort which is activated whenever employer j awards a gift to the worker à la Gneezy and List (2006). $\mathbb{1}_{Gift}$ is an indicator function which assumes a value 1 when a gift is rewarded

[13] We assume risk neutrality because the stakes are too small for the curvature of the preferences to matter. It also leaves us with one less parameter to estimate.

by the employer, 0 otherwise. α_j captures the altruistic preference of a worker towards employer j per unit of effort, where v is the (race independent and exogenous) value to the employer of a unit of effort by the worker. Note that our notion of altruism captures "pure altruism" as well as "warm glow" of the workers (DellaVigna et al., 2016): we don't aim to disentangle the two. p is the piece rate per unit of effort. $c(e_j)$ is the cost of effort function, assumed, for now, to be the same for all workers. We assume the regularity conditions $c'(.) > 0$, $c''(.) > 0$, and $lim_{e\to\infty} c'(e) = \infty$. The upshot is that effort is costly but helps generate both a) a private benefit (via, F, s and p) that would appeal to *Homo economicus*, and b) a part (via α and ρ) that would appeal to *Homo behavioralis*. Following DellaVigna and Pope (2018) and DellaVigna et al. (2016), we analyze the optimality conditions assuming two different functional forms for the cost of effort function : a power function and an exponential function i.e.,

$$c(e) = \frac{ke^{1+\gamma}}{1+\gamma}, \tag{3.3.2}$$

and

$$c(e) = \frac{kexp^{\gamma e}}{\gamma} \tag{3.3.3}$$

The power cost function (3.3.2) characterizes a constant elasticity of effort with respect to return to effort given by $1/\gamma$, while the exponential function (3.3.3) represents decreasing elasticity of effort with respect to return to effort given by $1/log(r/k)$, where r is the return to the effort. Workers' effort at different piece rates can be used to identify and structurally estimate both parameters of the cost-of-effort functions, namely, k and γ.

A worker solves the problem, $\underset{e_j \geq 0}{max}\, U_j$. The interior solution is characterized by:

$$e_j^\star = c'^{-1}\left(s + \rho_j \mathbb{1}_{Gift} + \alpha_j v + p\right) \tag{3.3.4}$$

which, for the power cost function, yields :

$$e_j^\star = \left(\frac{s + \rho_j \mathbb{1}_{Gift} + \alpha_j v + p}{k}\right)^{1/\gamma},$$

and

$$e_j^\star = \frac{1}{\gamma} ln\left(\frac{s + \rho_j \mathbb{1}_{Gift} + \alpha_j v + p}{k}\right)$$

for the exponential form.

We start by making the simplifying assumption that workers are homogeneous given a treatment i.e., they will make the same effort choice as any other worker assigned to the same treatment. We later relax this assumption to account for heterogeneity in effort within a treatment. Our goal is to identify the parameters of the model just described. To that end, we design our treatments by varying the incentives and behavioral motivators for the workers.

3.3.1 Piece Rate Treatments

Here, all else same, each worker works on a task at a given piece rate of either 0, 3, 6 or 9 cents per unit of effort (calibrated to 100 points scored on the task). The piece rates generate income in addition to the \$1 fixed participation fee, F. By M-Turk standards, this amount of variation in piece rates is substantial enough to elicit significant changes in effort thereby allowing us to estimate the baseline parameters (s, k, and γ) which, in turn, are used to estimate other behavioral parameters.

Formally, in the piece rate treatments, a worker observes a piece rate p and then chooses effort e_j. There is no corresponding employer j present in these treatments. This shuts down altruism and reciprocity right away: for any worker, $\alpha_j = 0$ and $\mathbb{1}_{Gift} = 0$. The equilibrium efforts $e_j^\star$ in these treatments is thus given as:

$$e_p^\star = c'^{-1}(s + p) \text{ for } p \in \{0, 3, 6, 9\}$$

The solution of effort has one behavioral unknown (s), and two unknowns from the cost function (k and γ). To back these out, we use effort corresponding to three different piece rates which gives us three equations to identify these parameters.

3.3.2 Altruism Treatments

In the altruism treatments, each worker is matched (see below for details) to an employer (truthfully) and he/she observes the (true) value of his/her effort to the matched employer. Specifically, each participant knows that an employer earns 1 cent for every 100 points scored by the matched worker. So as to not contaminate social preference with individual benefit, we set the piece rate to 0 in the three altruism treatments. In the first treatment (altruism baseline) a worker knows he/she has been matched to an employer but does not observe the employer's identity. In the 'altruism black' and 'altruism white' treatments, the worker observes the matched employer to be black and white, respectively.

Formally, in the altruism treatments, a worker observes the zero piece rate ($p = 0$), the value of the unit of effort to the employer j ($v = 0.01$), and then chooses effort e_j by maximizing (3.3.1). There is no gift from the employer implying $\mathbb{1}_{Gift} = 0$. The equilibrium efforts $e_j^\star$ in these treatments is, thus, given as:

$$e_j^\star = c'^{-1}\left(s + \alpha_j v\right) \text{ for } j \in \{Neutral, Black, White\}.$$

We are implicitly assuming that the altruism parameter can vary by the employer's group identity. For instance, $\alpha_{White} > \alpha_{Black}$ ($\alpha_{White} < \alpha_{Black}$) represents stronger (weaker) altruistic feelings for white as opposed to black employers. (As will be clear soon, all the workers in our sample are white which means α_j represents the strength of altruism a white worker feels for the jth employer.) Notice, since the piece rate is held fixed at 0 and reciprocity is shut out, the difference in effort provision between the 'altruism white' and 'altruism black' treatments is identifiable as resulting solely from the employer-race-dependent altruistic preferences of the workers. The three altruism treatments help us identify $\alpha_{Neutral}$, α_{Black}, and α_{White}, given the baseline parameters.

3.3.3 Reciprocity Treatments

Reciprocity treatments build on the altruism treatments and add a positive monetary gift (20 cents) from the employer to the worker. The remaining details are exactly the same as in altruism treatments. Thus, the equilibrium effort is given as;

$$e_j^\star = c'^{-1}\left(s + \alpha_j v + \rho_j\right) \text{ for } j \in \{Neutral, Black, White\}$$

As above, we are implicitly assuming that the reciprocity parameter may be different for each employer's group identity. In other words, controlling for the differences in altruism, the difference in effort between the treatments 'reciprocity white' and 'reciprocity black' is interpreted as resulting solely from the differential reciprocity preferences of the workers. The three reciprocity treatments help us identify $\rho_{Neutral}$, ρ_{Black}, and ρ_{White} given the baseline and altruism parameters.

3.4 Experiment Design

The main goal of this study is to investigate the possibility of discrimination by workers towards their out-group employers in an online labor market. Our variable of choice is effort provision and the margin of choice is intensive. Our experiment is designed to ensure that observed differences in effort provision can only realize because of the race-dependent social preferences of workers. That is, if we detect any discrimination, it will be entirely driven by taste parameters; after all, we rule out the possibility of statistical discrimination by making it clear that worker choices in no way can affect their future earning prospects on M-Turk and

employers will not get to make any payoff-relevant (or otherwise) choices after workers are done with the task.

3.4.1 Task

We need a task that is costly, effort-wise, to workers but is not meaningful in any way to a particular race. The task must require no special ability either. We settled on a button-pressing task as in DellaVigna and Pope (2018). The task involves alternating presses of "a" and "b" on a keyboard for 10 minutes. We chose it because it is simple to understand and has features that parallel clerical jobs: it involves repetition, it gets tiring (and boring), and therefore tests the motivation of the workers to stick to it and bring benefits to himself or his employer.

3.4.2 Race Revelation

We take the approach of revealing race via the revelation of skin color (Doleac & Stein, 2013b). To that end, we record videos of employers in otherwise-identical scenarios as they read off a script explaining and demonstrating the task. The camera placement only captures the hand of the employer along with the movement of the fingers alternating 'a' and 'b' button presses. Other identifiers, such as the face, are not captured in the video to avoid psychological confounds often associated with faces, such as attractiveness and trustworthiness (C. C. Eckel & Petrie, 2011). The employer's hand is bare or covered (with full sleeves and latex gloves) depending on the assigned treatment. For black employers, we restrict the sample to participants with darker skin tone to avoid any ambiguity about the race of the person. We mute the voice for the videos in the neutral treatments. We program each video to play with subtitles to aid easier understanding of the instructions. The sample video links for each treatment are given in Table 3.1.

3.4.3 Experiment Flow

The experiment proceeds as follows: (1) First, we recruit employers, students from a major public university in the U.S. Midwest and record videos of them explaining the task, 2) next, we post a HIT on Amazon's Mechanical Turk inviting M-Turkers to take a screener survey , (3) we invite those who meet the recruitment criteria (undisclosed) and consent to participate to initiate the experiment, (4) upon initiation, we assign each subject to one of the aforediscussed treatment groups. Following Czibor, Jimenez-Gomez, and List (2019), we use the blocked randomization design to assign subjects to treatments. We define blocks based on demographic information collected in the screener survey (Gender, Age, Race, Education,

Income, Political Party Affiliation, and the Most-Lived U.S. state),[14] Next, (5) we present instructions to each subject in a pre-recorded video (based on the assigned treatment). We program our study to *require* each worker to watch the assigned video, (6) we elicit incentivised beliefs from each worker about their matched employer,[15,16] and 7) workers start to work on the task for a maximum of 10-minutes.

3.4.3.1 Piece Rate Treatments

In the piece rate treatments, each worker sees a video demonstrating a task with a script: "*On the next page, you will play a simple button-pressing task. The object of the task is to alternately press the 'a' and 'b' buttons on your keyboard as quickly as possible for ten minutes. Every time you successfully press the 'a' and then the 'b' button, you will receive a point. Note that points will only be rewarded when you alternate button pushes: just pressing the 'a' or 'b' button without alternating between the two will not result in points. Buttons must be pressed by hand only (key-bindings or automated button-pushing programs/scripts cannot be used), or task will not be approved. Feel free to score as many points as you can.*" The final line is tailored to the assigned treatment (0, 3, 6 or 9 cents). The wording is provided in Table 3.1. Even though piece rates are framed in units of 100 points, workers are paid continuously for each point scored and are able to see the earned bonus in real time as they score points.

3.4.3.2 Social Preference Treatments

In the altruism and reciprocity treatments, each video starts with the introduction by the employer: "*Hi, I am another participant in this study who is matched to you. In this study, you will work on a simple button-pressing task, and I will earn some money depending on how well you do on the task.*" Thereafter, the script follows the same instructions as in piece rate treatments with the last paragraph being tailored to the social preference treatment in question. The wording is provided in Table 3.1. There are three treatments each in the category of altruism and reciprocity. Altruism-baseline and reciprocity-baseline conceals the skin color of the employer in the video using latex gloves. The voice in the baseline treatments is also muted so as not to reveal any racial markers present in the voice. We recruit an equal number of black and white employers

[14] See Cavaille (2018) for instructions on implementing sequential blocked randomization for online experiments.

[15] The elicitation of beliefs *before* workers start work on the task serves two purposes: 1) it provides us with data on workers' beliefs about the identity of their employer, and 2) it allows for the identity of the employer to become salient to the worker; importantly, it renders prominence to the seemingly-obvious fact that the worker is indeed matched to a real person whose payoff will be influenced by the worker's choices. We believe prior belief elicitation serves to increase salience of employer identity but does not amount to targeted priming courtesy the between-subject design of the study. It is important to note that workers' beliefs are elicited on a variety of identities (gender, age, income, and so on), not just race. We do not reveal our desire to know about their beliefs on race; it is just one of *six* different categories they are asked to report their beliefs on. As such, we are confident, our results are not tainted by experimenter demand effects. Parenthetically, in post experiment comments, not one worker identified ours to be a study about race or discrimination.

[16] To discourage random guessing in the belief elicitation part, participants are informed that an incorrect guess will lead to a deduction of 2 cents from their final earnings.

Table 3.1: Summary of treatments

Category	Treatment Wording	Voice	Skin Color	Sample Video
(1)	(2)	(3)	(4)	(5)
Piece Rate	Your score will not affect your payment in any way.	Muted	Concealed	Link
	As a bonus, you will be paid an extra 3 cents for every 100 points that you score.	Muted	Concealed	Link
	As a bonus, you will be paid an extra 6 cents for every 100 points that you score.	Muted	Concealed	Link
	As a bonus, you will be paid an extra 9 cents for every 100 points that you score.	Muted	Concealed	Link
Altruism	I will earn 1 cent for every 100 points that you score. Your score will not affect your payment in any way.	Muted	Concealed	Link
	I will earn 1 cent for every 100 points that you score. Your score will not affect your payment in any way.	Black	Black	Link
	I will earn 1 cent for every 100 points that you score. Your score will not affect your payment in any way.	White	White	Link
Reciprocity	I will earn 1 cent for every 100 points that you score. In appreciation to you for performing this task, I have decided to pay you extra 20 cents as a bonus. Your score will not affect your payment in any way.	Muted	Concealed	Link
	I will earn 1 cent for every 100 points that you score. In appreciation to you for performing this task, I have decided to pay you extra 20 cents as a bonus. Your score will not affect your payment in any way.	Black	Black	Link
	I will earn 1 cent for every 100 points that you score. In appreciation to you for performing this task, I have decided to pay you extra 20 cents as a bonus. Your score will not affect your payment in any way.	White	White	Link

Notes: The table list all the treatments in this study. Each piece rate treatment differs just in the last line of the script, uses no audio, and conceals the skin color of the hand. Social preference treatments (altruism and reciprocity) begin with the introduction of the employer (in the first person), explain the task using the same script as in piece rate treatments and then differ only in the last paragraph of the script. Both altruism and reciprocity categories have three treatments, each with black, white, and concealed skin tone of the employer (using gloves). In the social preference treatments of concealed skin tone, the ratio of black and white employers is 1:1.

in the neutral treatments. The videos shown to workers in the altruism black (white), and reciprocity black (white) clearly reveal the black (white) skins of the employers respectively.

3.4.4 Recruitment of Subjects

3.4.4.1 Recruitment of Employers

To recruit employers, we invite male student subjects over the age of 18 from a major public university in the U.S. Midwest who racially identify as either African American or Caucasian. We restrict our sample to male and U.S.-based employer-subjects to avoid confounds arising from identity effects of gender and nationality effects. Holding the sample size fixed and restricting it to one social identity give us extra statistical power and thereby ability to draw more credible inferences. We restrict the sample to employer subjects who are either black or white (we exclude Asians and Latinos, for example) because we believe our race-revelation mechanism works best in the context of these two races. We call these student subjects "employers" because they assign tasks to the workers who, in turn, work for these subjects and receive compensation (as is typical in most employer-worker relationships). Workers, at no point, know that the "employers" are students.[17] When an employer-subject arrived at the lab, they filled out a short demographic survey and was then randomly assigned to one of six social preference treatments. Based on the assigned treatment, subjects read from the script and demonstrate the task on a video. Each subject was paid $5 for participation and an additional variable amount (average of $17.5) depending on their matched worker's performance. Our final sample include six employers in each social preference treatment (in all, 36 employers, 18 blacks and 18 whites).

3.4.4.2 Recruitment of Workers

We recruit U.S. based workers from Amazon's Mechanical Turk, a popular crowd-sourcing web-service that allows employers (called requester) to get tasks (called Human Intelligence Tasks (HITs)) executed by employees (called workers) in exchange for a wage (called reward). Mechanical Turk is a widely used platform for research in economics and allows access to a large pool of applicants at an affordable rate.[18]

We post a screener survey as the HIT on M-Turk with the following description "*Fill out this 2-minute screener survey to qualify for a study that starts immediately, take up to 15 minutes, and pays participation bonus $1 with scope to earn extra. You will be required to watch and listen to a video. Do NOT take this study on mobile.*". The responses to the screener survey allows us to pick participants that satisfy the

[17] These employer subjects are framed as "other participant" to the workers so as not to introduce any imaginative effects from the use of make-believe language such as "employer".

[18] See Paolacci, Chandler, and Ipeirotis (2010) and Paolacci and Chandler (2014) for a discussion on demographic characteristics and representation of subjects from M-Turk.

criteria listed above. We allowed both black and white workers to participate. As per our pre-registration commitment, we recruited black workers only for race-salient, social-preference treatments. However, in the end, we could only recruit 711 (U.S.-based) black workers in the four social preference treatments combined. Power considerations, therefore, precluded their inclusion in the final analysis. Perforce, we restrict attention to white workers and study their effort choices for black versus white employers. We paid 15-cents to each potential subject for filling out the screener survey. On average, the workers in our sample earned $1.72 (including $0.15 for the screener survey, $1.0 for participation, and up-to $0.1 for belief elicitation questions). These payments are sizable as per M-Turk standards for a 10-minute task and come close to the pro-rated, federal minimum wage in the United States.

3.4.5 Pre-registration

We pre-registered the design on AEA RCT registry as AEARCTR-0003885. Since our task is the same as used in DellaVigna and Pope (2018), we can use results from their study to determine the sample size needed to achieve sufficient power for our study. DellaVigna and Pope (2018) find that the points scored across all treatments have a standard deviation of around 660 . Assuming this standard deviation for each treatment and assuming a minimum detectable effect of 0.2 standard deviations between two treatments, we needed around 400 subjects in each treatment to have a power of 80 percent. This implies that we needed 400 x 10 = 4,000 observations in total for all ten treatments. We pre-registered the rule for sample size collection: we aimed to recruit 6,000 worker-subjects from M-Turk within the first three weeks of posting the experiment. Our data collection went slower than anticipated, and we ended up recruiting subjects from August 5th, 2019 to October 24th, 2019. In our registration, we had also planned to recruit self-identified black workers, which as explained above, did not work out.

3.5 Data

3.5.1 Employers

The demographic characteristics of the employer subjects in each treatment are presented in Table B1.

3.5.1.1 Pre-Testing of Videos

To verify whether the videos accurately reveal race , we test them using an independent sample of U.S.-based, white subjects from Academic Prolific, a data collection platform. We used them instead of M-Turk to ensure that our M-Turk recruits could not have watched these videos before they participate in our

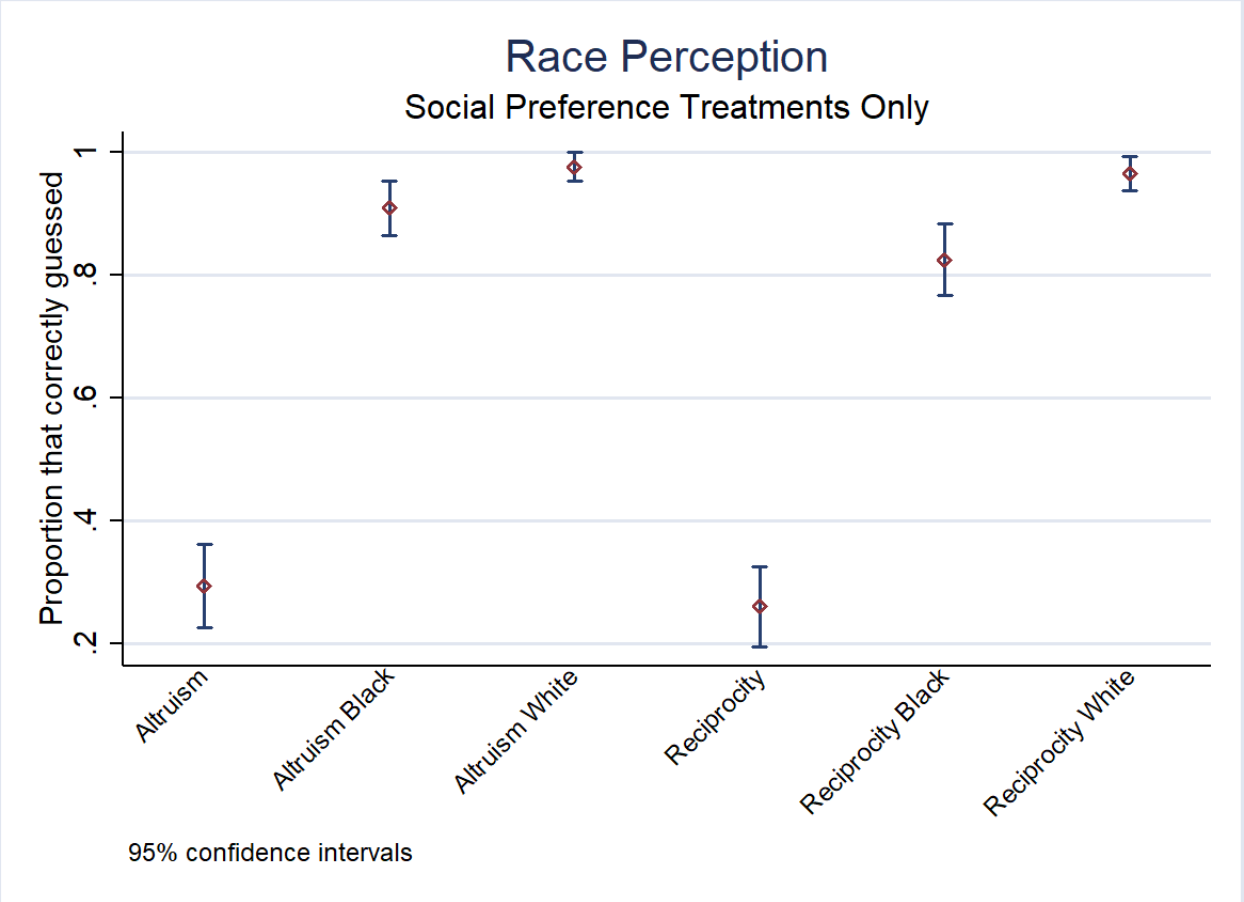


Figure 3.5.1: Race Perception

Notes: This figure shows the proportion (with confidence interval) of individuals who were able to correctly guess the race of the employer after watching a video.

experiment. Each subject was asked to identify the race of the person in a randomly-assigned video. See Figure 3.5.1 for a graphical representation of average perception of race across treatments. Overall, race is correctly perceived more than 80 percent of the time for all the race-salient treatments: – our race revelation mechanism works. For the race-neutral treatments, only less than 30 percent of the participants could guess the race, probably the result of random guessing. The pairwise comparisons of race perception among these treatments is presented in Table B2. The results suggest that the race-neutral treatments (altruism and reciprocity) are statistically indistinguishable from each other and significantly different from race-salient treatments. The perception of race in the treatments 'Altruism Black' and 'Altruism White' is statistically indistinguishable; however 'Reciprocity Black' is not perceived as accurately as 'Reciprocity White'.

Participants also evaluated the videos in race-salient treatments for perception of skin color; the results are presented in Figure 3.5.2. Overall, blacks' skin is correctly perceived to be of darker tone and whites' of lighter tone. The pairwise comparisons of skin color perception among these treatments is presented in

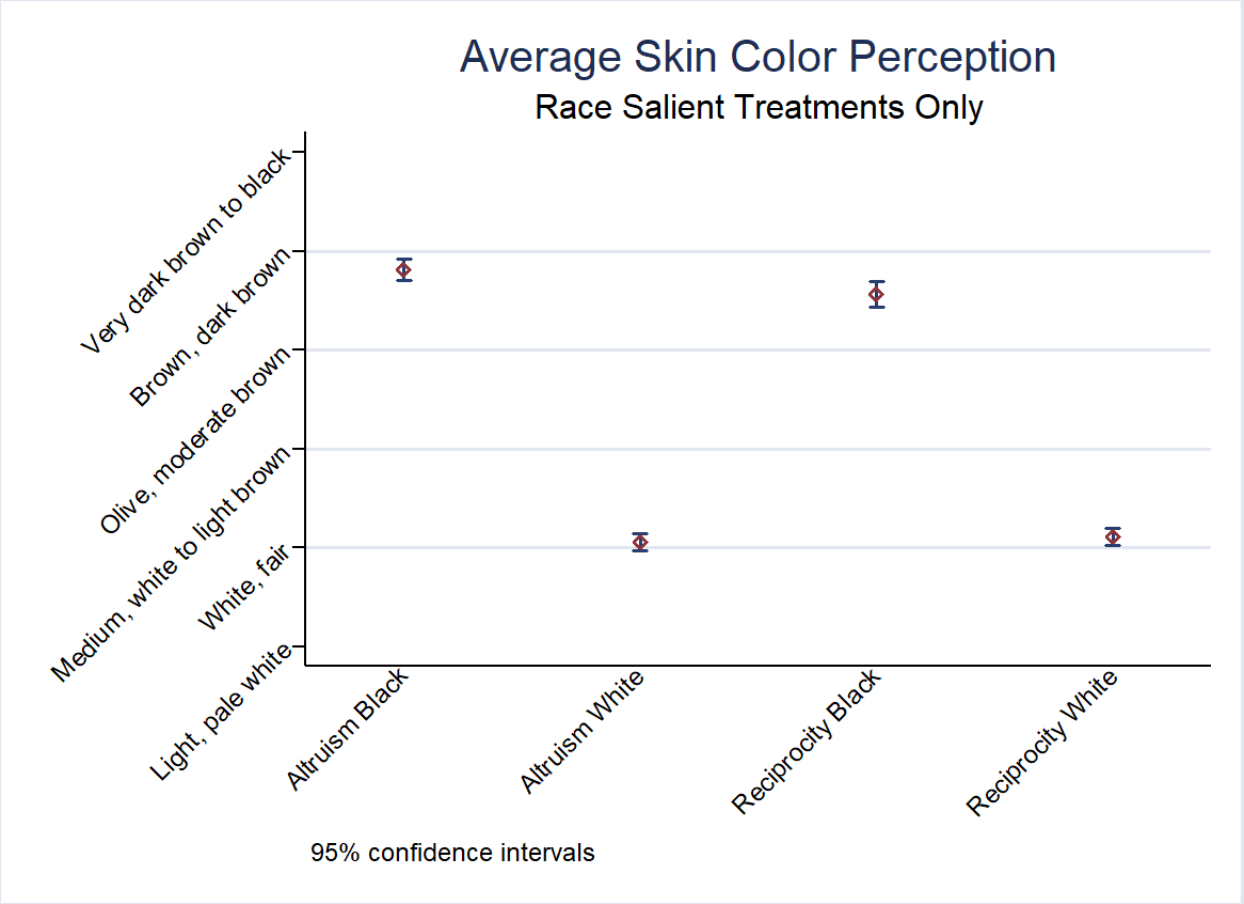


Figure 3.5.2: Skin Color Perception

Notes: This figure shows the average (with confidence interval) perceived skin tone across the race salient treatments.

Table B2. The results suggest black treatments are statistically indistinguishable from each other and are significantly different from white treatments.

Finally, to check whether subjects in the videos were not perceived differently on soft personality traits such as friendliness, professionalism, clarity etc., we get these videos rated on those traits. The results for positive and negative traits are presented in Figure A1 and A2 of the Appendix A respectively. Pairwise comparisons of means across all the social-preference treatments suggest only the reciprocity-black treatment is perceived to be significantly higher on positive traits while all other treatments are statistically indistinguishable from each other on both positive and negative traits (see Table B3 of the Appendix B). This confirms that the only difference between the black and white employer' videos is the perceived race of the employer.

3.5.2 Workers

As per the pre-registration, we apply the following restrictions to the collected data: (1) we drop 17 participants who scored above 4,000 points as this is physically impossible in the 10- minute time-frame – likely, these users used some automated programs to score points;[19], 2) we drop 64 workers who scored zero points as this may reflect some malfunction or technical problem in the recording of points;[20] (3) we drop 4 observations of workers who participate more than once;[21] (4) we dropped two observations from workers who somehow managed to take this study from outside the United States.[22]

The final sample consists of 5,945 workers and the summary statistics are presented in Table 3.2. Our sample over represents women, young, educated, middle-income, and Democrats as compared to the U.S. labor force. This is typical of the population on online platforms. We present results of productivity by demographics in Table B8. Overall in our sample, men and younger workers are more productive than women and older workers respectively. We present test-of-balance of demographic variables across ten treatments in Table B4 of the the Appendix B. The treatments are balanced on all the observed variables, no surprise since we use blocked randomization to assign subjects to treatments. Since worker characteristics are balanced across treatments, there is no reason to believe that more/less productive workers are assigned to any specific treatment.

3.6 Results

We present average effort by workers (recall, all our workers are white) against each treatment in column 1 of Table 3.3 and in Figure 3.6.1. Overall, it is evident that incentives have a strong effect on effort, raising performance from 1627 points (piece rate of 0) to 2060 points (3-cent piece rate) and 2127 points (9-cent piece rate).[23] However, the average effort for 3-cent and 6-cent treatments is statistically the same, reflecting a low elasticity of effort beyond an initial increase in effort from 0 to 3 cents. The standard error for the mean effort per treatment is around 30 points or less, implying that differences across treatments larger than 85 points are statistically significant.

How do we detect altruism? We compute the average effort of workers in the altruism-neutral treatment; recall, these are workers who know they are matched to an employer but don't know his race. We compare

[19] We instructed each worker up-front to not use any automated scripts/programs .

[20] These workers are spread across all treatments, and there is no systematic difference in workers scoring zero points for any particular treatment or employer.

[21] A worker can participate in our study only once; these exceptions must be an error on the part of M-Turk.

[22] The study was restricted to U.S.-based workers. Presumably, these participants used a proxy server or VPN to mask their origin but we could spot them from the GPS coordinates recorded by Qualtrics.

[23] Workers' positive effort in the 0-cent treatment is explained by the parameter s of the model in Section 2. Part of the positive effort could also be workers' unsubstantiated fear of being rejected for not scoring enough points.

Table 3.2: Summary Statistics, Worker Sample

	(1) Sample	(2) US Labor Force
Gender		
Female	0.58	0.47
Male	0.41	0.53
Race		
White or Caucasian	1.00	0.78
Age		
18-24	0.12	0.11
25-30	0.38	0.14
31-40	0.26	0.22
41-50	0.14	0.21
51-64	0.08	0.25
65 and over	0.03	0.06
Education		
Less than high school	0.01	0.14
High school or equivalent	0.13	0.39
Some college	0.28	0.35
College graduate	0.41	0.30
Graduate or professional degree	0.18	0.18
Income		
Less than $20,000	0.17	0.20
$20,000 - $44,999	0.31	0.26
$45,000 - $99,999	0.38	0.33
$100,000 - $149,999	0.09	0.12
$150,000+	0.03	0.08
Political Affiliation		
Democrat	0.39	0.31
Independent	0.28	0.38
Republican	0.27	0.29
Most lived US State		
Blue	0.31	0.47
Red	0.20	0.14
Swing	0.49	0.39
Observations	5945	162075000

Notes: The table presents demographic information of worker subjects. Column (1) presents proportion of the worker subjects by their gender, race, age, education, income, party, and the most lived state in the United States. Column (2) presents these demographics for US labor force based on 2018 numbers from Bureau of Labor Statistics/Current Population Survey. Estimates of population by political affiliation and by blue, red, and swing state are based on Gallup polling survey 2019.

Table 3.3: Effort by Treatment

	(1) All Workers		(2) Correctly Perceived Race	
	N	Mean (s.e)	N	Mean (s.e)
Piece Rate - 0 cents	599	1627.07	599	1627.07
		(28.56)		(28.56)
Piece Rate - 3 cents	595	2059.83	595	2059.83
		(24.19)		(24.19)
Piece Rate - 6 cents	592	2046.68	592	2046.68
		(23.62)		(23.62)
Piece Rate - 9 cents	588	2127.37	588	2127.37
		(23.01)		(23.01)
Altruism - Neutral	591	1746.06	261	1724.87
		(29.15)		(43.70)
Altruism - Black	601	1798.37	494	1807.68
		(27.55)		(29.58)
Altruism - White	592	1708.09	557	1715.24
		(28.90)		(29.52)
Reciprocity - Neutral	608	1771.15	265	1766.99
		(27.95)		(41.63)
Reciprocity - Black	590	1803.61	470	1818.78
		(26.95)		(29.73)
Reciprocity - White	589	1798.23	561	1803.75
		(29.58)		(30.33)
Total	5945	1848.08	4982	1865.98
		(8.80)		(9.49)

Notes: The table presents the effort choices in each treatment. Column 1 reports the effort choices by all the workers, column 2 reports the effort choices by workers who were able to correctly perceive the race of the employer as neutral, black or white in social preference treatments.

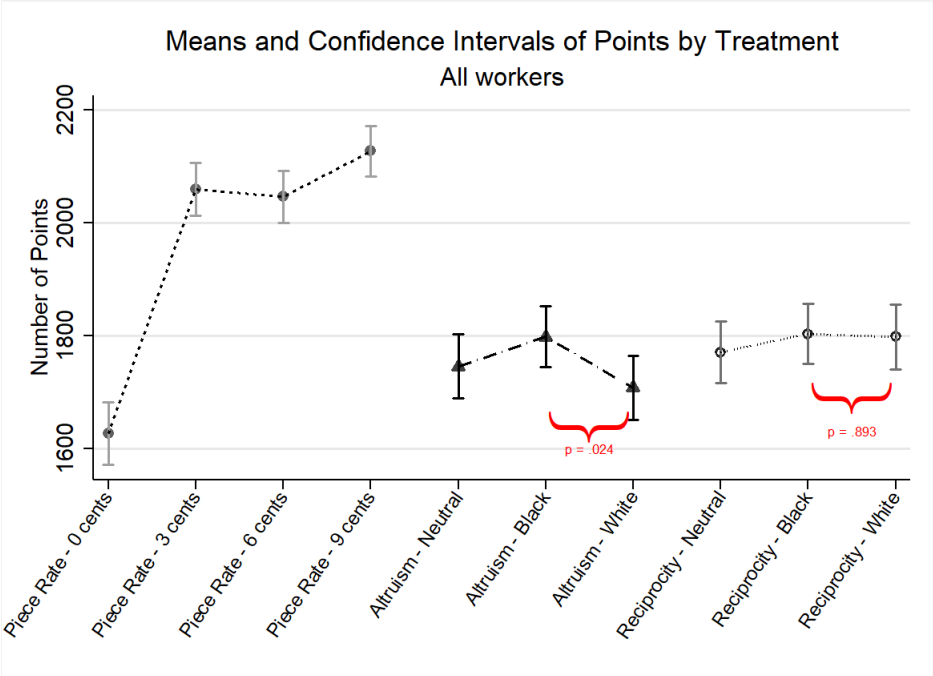


Figure 3.6.1: Effort by Treatment - All Workers

Notes: This figure presents the average score and confidence interval for each of ten treatments for all workers. Each treatment has about 590 participants.

that to the average effort of workers who are not matched to any employer and are offered a 0-cent piece rate. We find statistically significant evidence for altruism: workers put more effort in the altruism-neutral treatment as compared to the piece rate 0-cent treatment. The one cent return to the employer induces an effort of 1746 points as compared to 1627 points in the 0-cent treatment.

Next, we wish to compare the strength of the altruism preference across black and white employers. To that end, we compute the average effort of workers in the altruism-black treatment to those in the altruism-white treatment. Strikingly, workers are significantly more altruistic towards black employers than white employers. The average effort for black employers is 1798 points, which is significantly higher ($p = 0.024$) than the same for white employers (1708 points). However, note that in the altruism treatments, average effort for any race employer is not significantly different from the same for a race-neutral employer .

In the reciprocity treatments, the worker receives an unanticipated gift of 20-cents (in addition to all other earnings) from the employer, unconditional on performance. This gift does not induce a significant increase in effort as compared to the altruism treatment (1771 points in the reciprocity-neutral treatment as compared to 1746 points in the altruism-neutral treatment). The reciprocal response to the employer's racial identity is also insignificant, implying that, on average, workers do not reciprocate towards any employer

Table 3.4: Social Preference Treatments - Robustness

	Altruism			Reciprocity		
	(1)	(2)	(3)	(4)	(5)	(6)
Black or African American	52.31	53.68	56.49	32.46	37.15	13.60
	(40.31)	(40.74)	(41.42)	(39.74)	(40.43)	(40.92)
White or Caucasian	-37.97	-30.84	-33.82	27.08	34.33	1.521
	(40.46)	(41.07)	(41.78)	(39.76)	(40.67)	(41.60)
Constant	1746.1***	1648.9***	1705.9***	1771.1***	1801.6***	1781.8***
	(28.62)	(254.7)	(262.6)	(27.89)	(267.4)	(278.1)
Demographic Controls	No	Yes	Yes	No	Yes	Yes
Employer Perception	No	No	Yes	No	No	Yes
N	1784	1706	1701	1787	1719	1717
Black - White	90.28**	84.53**	90.30**	5.379	2.822	12.08
	(40.29)	(40.89)	(41.17)	(40.06)	(40.94)	(41.20)

Notes: The table presents the estimates from an OLS regression of Points in the social preference treatments on the employer's race. The omitted category is the employer with concealed race. Demographic controls include age, gender, education, income, political affiliation and the voting pattern of the most lived state (red, blue, or swing) of the worker. Employer Perception include worker's belief about the income, age, and education of the employer. Black - White represents the difference in the coefficients of black and white employers in each model. Standard errors in parentheses. for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$.*

race. This result is consistent with the literature which finds weak evidence for positive reciprocity (such as Kube, Maréchal, and Puppe (2006)).

In column 2 of Table 3.3, we restrict the analysis to only workers who could correctly guess the race of the employer in the social preference treatments. This does not substantially affect the direction or magnitude of the results.

Although our treatments are balanced on observed worker and employer characteristics, for robustness sake we present the regression results from regressing "Points" scored on the employer racial identity and controlling for these variables in the Table 3.4. We observe that workers' pro-altruistic response for black employers stays significantly different from white employers even after controlling for the demographic variables and the perception about the employer's income, age, and education. The latter highlight the fact that the higher altruism towards blacks is not driven by the differences in beliefs about the employer's income, age, and education. The reciprocity response stays statistically indifferent from zero for all the specifications.

3.6.1 Distribution of Effort

Beyond average effort, we present the distribution of effort from all the treatments in Figure A3 of the Appendix A and by each treatment in Figure A4 of the Appendix A. Overall, very few workers score below 500 points and even fewer score above 3000 points.

Figure 3.6.2a presents the cumulative distribution function for the piece rate treatments. Incentives induce a clear rightward shift in effort relative to the 0-cent treatment. However, there is not much change in effort between the 3-cent and the 6-cent treatments. Figure 3.6.2b shows strong evidence for altruistic preferences as observed by the clear rightward shift of the effort distribution in the altruism treatment as compared to the 0-cent treatment. The effort distribution in the reciprocity treatment is indistinguishable from the altruism treatment, implying a lack of reciprocal preferences. Figure 3.6.2c shows that altruism is stronger towards blacks as compared to whites while the cumulative density function is indistinguishable for reciprocity-black and reciprocity-white treatments. Quantile regression estimates for effort (Table B6 of Appendix B) show that black employers get higher effort than white employers at both the 0.25 and 0.5 quantile for the altruism treatments. This shows that the altruistic response for the black employers is mainly coming from the *lower part of the* effort distribution. On the other hand, there is no difference between black and white employers for the reciprocity treatments at any quantile.The Kolmogorov-Smirnov test of equality of distribution functions is presented in Table B5 of Appendix B.

3.6.2 Evolution of Effort

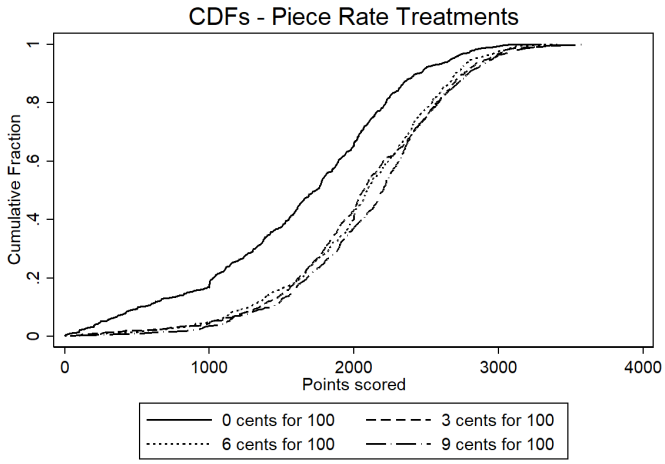
We present the evolution of effort over the 10-minute period in Figure 3.6.3. Figure 3.6.3b and 3.6.3c shows that, in the social preference treatments, effort declines over time presumably due to boredom and tiredness. And yet, interestingly, the piece rate treatments are able to sustain consistently higher effort throughout the entire time interval (Figure 3.6.3a), with workers in the 9-cent treatment pushing extra hard near the end. In the altruism treatments (Figure 3.6.3b), the effort for black employers is higher than effort for white employers throughout the 10-minutes period. However, in reciprocity treatments (Figure 3.6.3c), the effort for black and white employers are indistinguishable from each other.

3.6.3 Heterogeneity

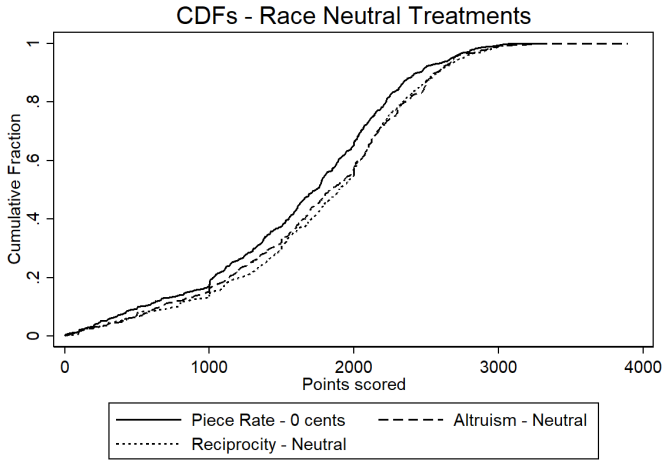
3.6.3.1 Heterogeneity by Demographics

To examine the heterogeneity in our average treatment effects based on demographic characteristics of the sample, we present the differences in treatment effects in Table 3.5 for both altruism and reciprocity treatments. Overall, we do not find evidence of heterogeneity in treatment effects on the basis of gender, age, and education in both altruism and reciprocity. However, we do find some evidence of heterogeneity in altruism on the basis of party affiliation and voting pattern of the state. Republicans are more altruistic towards black employers (vis-à-vis white employers) as compared to the Democrats ($p = 0.12$ for test of White $\times$ Democrat = White $\times$ Republican in Table 3.5). Similarly, workers from Red states are significantly more

(a)



(b)



(c)

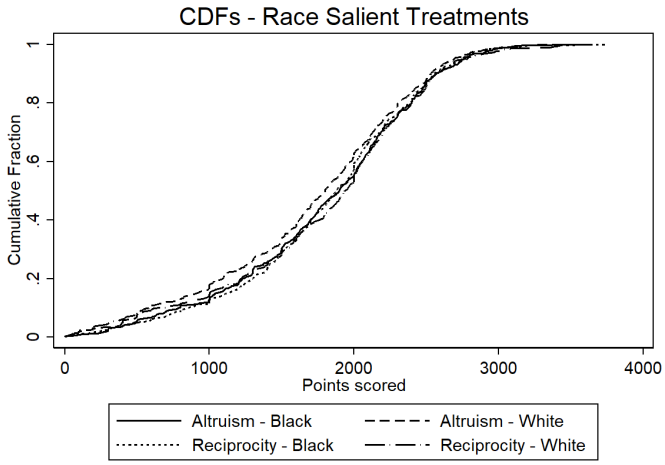
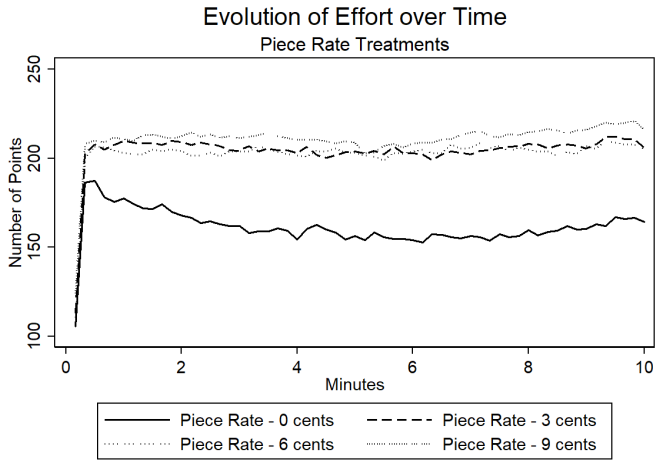


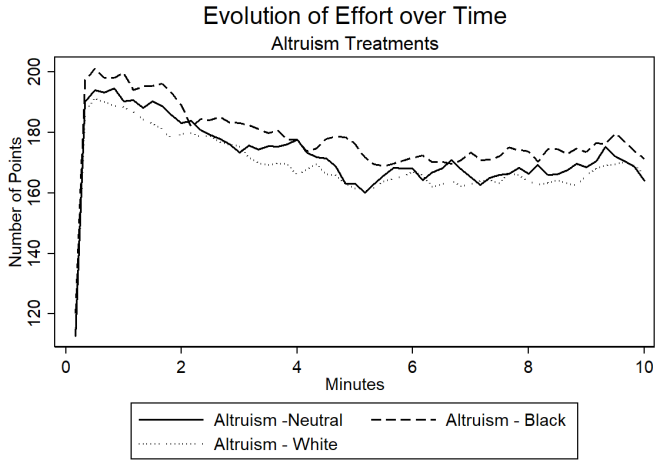
Figure 3.6.2: Cumulative distribution function

Notes: The figure presents the cumulative distribution function of points for the workers in each of the treatments featured. The sample size in each treatment is approximately 590 subjects. Figure a features the four piece rate treatments (no piece rate, 3-cent per 100 points, 6 cents per 100 points, and 9 cents per 100 points). Figure b presents the results for the race-neutral treatments. Figure c presents the results for the race-salient treatments.

(a)



(b)



(c)

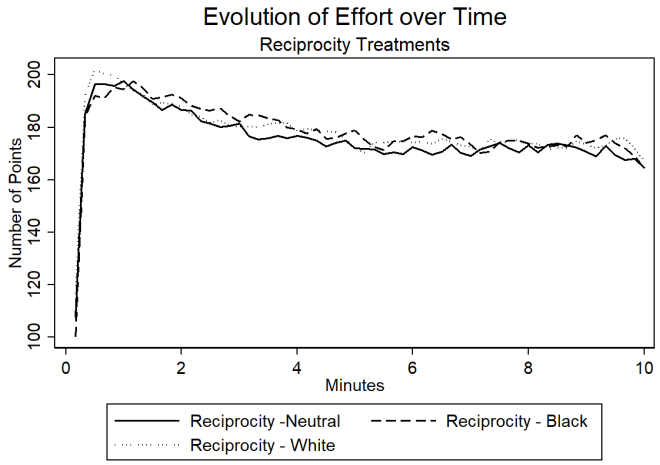


Figure 3.6.3: Evolution of effort over time

Notes: This figure presents the effort over time for piece rate (figure (a)), altruism (figure (b)), and reciprocity (figure (c)) treatments. The y axis indicates the average number of points scored in that treatment per minute.

altruistic ($p = 0.089$ for the test of White $\times$ Blue = White $\times$ Red in Table 3.5) towards black employers (vis-à-vis white employers) as compared to workers from Blue states. We also find that higher income workers reciprocate significantly more to the black employers (vis-à-vis white employers) as compared to lower income workers.

3.6.3.2 Heterogeneity by the share of black population in the neighborhood

Following Andreoni, Payne, Smith, and Karp (2016), we explore the effects of local racial composition on social preferences of the workers in our sample. We condition on the zip code level racial composition of the worker and examine the difference in effort provided to black versus white employers. Table 3.6 presents the conditional average treatment effects for top and bottom quantile of the share of black population for workers who correctly perceived the employer race for both altruism and reciprocity treatments. Consistent with the hypothesis 'familiarity breeds contempt', we find that workers from neighborhoods with lower share of black population are significantly more altruistic towards black employers. We do not find evidence for differential reciprocal response on the basis of share of black population in the neighborhood.

3.6.3.3 Heterogeneity by Geographical Area

It is a well established fact that racial disparities are not equally distributed across the U.S. We present the summary of workers performance by their geographical area in Table 3.7. Interestingly, there is an evidence in favor of workers from South being relatively more altruistic to black employers. This is surprising given that the average implicit bias against blacks (see next subsection) in the South is higher than in the other regions of the U.S.

3.6.3.4 Heterogeneity by Implicit Biases

We examine the heterogeneity in treatment effects based on the implicit biases of workers as measured by the implicit association test (IAT). IATs are widely used in social psychology to measure implicit and unconscious biases towards a particular group. The test involves categorizing two sets of words to the left or right hand side of the computer screen. The implicit bias is measured by a time difference in associating good or bad words to the relevant group identities. The idea is that making a response is easier when closely related items share categorization to the same side of the screen. In case of race IAT, we would say that one has an implicit preference for white people relative to black people if they are faster to categorize words when white face and good words (friend, glorious, enjoy, joyous, terrific, beautiful, magnificent, and fabulous) share a response key and black faces and bad words (detest, poison, nasty, disgust, pain, despise, sadness, evil) share a response key, relative to the reverse.

Table 3.5: Heterogeneity by Demographics

	(1) Altruism	(2) Reciprocity
Employer Race		
White	-113.6	-12.80
	(112.0)	(111.9)
Gender		
Female	-151.8***	-142.5**
	(57.40)	(57.75)
White × Female	99.12	93.07
	(81.74)	(82.53)
Age		
35 and above	-41.70	-142.9**
	(57.40)	(57.01)
White × 35 and above	-19.01	5.892
	(81.87)	(81.27)
Education		
College and above	-109.0*	-160.7***
	(60.53)	(61.04)
White × College and above	-110.8	94.07
	(85.69)	(86.46)
Income		
≥ $45,000	65.78	153.9**
	(59.30)	(60.01)
White × ≥ $45,000	30.76	-258.9***
	(84.40)	(85.27)
Party Affiliation		
Democrat	-85.17	-107.1
	(66.99)	(67.38)
Republican	33.12	-40.87
	(74.23)	(74.28)
White × Democrat	139.0	115.0
	(95.84)	(95.38)
White × Republican	-16.87	116.8
	(104.8)	(107.2)
State Voting Pattern		
Blue	-82.25	4.238
	(64.43)	(65.03)
Red	74.99	2.641
	(76.72)	(76.32)
White × Blue	23.33	-99.36
	(92.51)	(92.78)
White × Red	-174.9	-31.83
	(107.2)	(109.3)
Constant	1974.4***	2010.7***
	(76.18)	(77.00)
Observations	1171	1149

*Notes: The table presents the differences in average treatment effects by the demographics of the workers for both altruism (column 1) and reciprocity (column 2) treatments. The omitted employer is the Black employer. The reference categories for gender, age, education, income, party affiliation, and state voting pattern are female, below 35, below college, under $45,000, independent, and swing state respectively. Standard errors in parentheses. * for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$.*

Table 3.6: CATE by Black Share Quantiles - Bottom and Top

	Altruism		Reciprocity	
	(1)	(2)	(3)	(4)
	Lower Share	Higher Share	Lower Share	Higher Share
White	-184.0***	-57.15	-45.66	-15.27
	(65.09)	(59.62)	(62.96)	(64.87)
Constant	1854.7***	1431.5***	1494.5***	1671.1***
	(532.3)	(390.8)	(543.6)	(387.5)
Demographic Controls	Yes	Yes	Yes	Yes
Employer Perception	Yes	Yes	Yes	Yes
Observations	487	515	498	477

Notes: The table presents the conditional average treatment effect by the bottom and top quantile of the share of black population in the worker's zip code for both altruism (column 1 and 2) and reciprocity (column 3 and 4) treatments. Measure of conditional treatment effect is obtained by restricting to workers who could correctly perceive the employer race and running a regression of Points on Employer Race for bottom (column 1 and 3) and top (column 2 and 4) quantile of share of black population in the worker's zip code while controlling for demographics and employer perception. Standard errors in parenthesis. for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$.*

For this study, we did not conduct IAT test for individual workers. Nor did we employ survey measures so as to avoid revealing the purpose of the study.[24] Instead we proxy the IAT score of individual worker by using the Geo-coded race IAT data by Project Implicit, which provides historical record of tests taken on the project's website. These tests can be taken by anyone from anywhere in the world. For our purpose, we restrict to white individuals from the United States and use the data from more than two million test takers between 2006 to 2018. We map the county level (lowest available resolution) IAT score to workers in our sample based on the worker's geographic location. Our worker sample comes from 190 counties spanning all 50 states in the U.S.

Typical thresholds found in the literature (Greenwald, Nosek, & Banaji, 2003; Hahn, Judd, Hirsh, & Blair, 2014; Rooth, 2010) are as follows: IAT scores below -0.15 indicate some preference for minorities; scores between -0.15 and 0.15 indicate little to no bias; scores between 0.15 and 0.35 indicate a slight bias against minorities; and scores above 0.35 show moderate to severe bias against minorities. The average score (standard deviation) of white test takers in our sample is 0.38 (0.42) implying, on average, white people have moderate to severe implicit bias against blacks. Like black share, we explore the effects of local IAT score on the social preferences of workers in our sample. We condition on the county level IAT score of the worker and examine the difference in effort provided for black versus white employers. Restricting to two quantiles of IAT score clearly shows (Table 3.8) that workers with lower implicit bias are significantly more

[24] M-Turkers often communicate with each other on various platforms; as such, we wanted to make sure that the purpose of the study is not broadcasted even after the worker is done with the task.

Table 3.7: Heterogeneity by Geographical Area

(a)

	Altruism			
	(1)	(2)	(3)	(4)
	North East	Mid West	South	West
White or Caucasian	16.42	-76.81	-155.4**	18.00
	(96.74)	(90.16)	(67.89)	(111.3)
Constant	2344.7***	1486.9***	1452.0***	1777.9***
	(314.5)	(289.2)	(432.0)	(621.0)
Demographic Controls	Yes	Yes	Yes	Yes
Employer Perception	Yes	Yes	Yes	Yes
Observations	190	243	409	206

(b)

	Reciprocity			
	(1)	(2)	(3)	(4)
	North East	Mid West	South	West
White or Caucasian	-76.02	7.934	-34.81	-62.23
	(111.4)	(77.90)	(75.22)	(112.5)
Constant	3305.1***	1236.2***	941.1	1651.3**
	(874.7)	(449.8)	(758.1)	(822.7)
Demographic Controls	Yes	Yes	Yes	Yes
Employer Perception	Yes	Yes	Yes	Yes
Observations	186	286	360	198

Notes: The table presents the conditional average treatment effect by the geographical location of the worker for altruism (table (a)) and reciprocity (table (b)) treatments. Measure of conditional treatment effect is obtained by restricting to workers who could correctly perceive the employer race and running a regression of Points on Employer Race for each geographical region while controlling for demographic and employer perception. Standard errors in parenthesis. for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$.*

Table 3.8: CATE by IAT Quantiles - Bottom and Top

	Altruism		Reciprocity	
	(1)	(2)	(3)	(4)
	Lower Bias	Higher Bias	Lower Bias	Higher Bias
White or Caucasian	-172.4***	-25.33	-90.86	45.18
	(64.51)	(58.33)	(63.76)	(60.10)
Constant	1218.6***	2164.4***	1457.3***	2006.7***
	(467.2)	(413.2)	(408.2)	(503.1)
Demographic Controls	Yes	Yes	Yes	Yes
Employer Perception	Yes	Yes	Yes	Yes
Observations	529	513	526	495

Notes: The table presents the conditional average treatment effect by the bottom and top quantile of the IAT score of the worker's county for both altruism (column 1 and 2) and reciprocity (column 3 and 4) treatments. Measure of conditional treatment effect is obtained by restricting to workers who could correctly perceive the employer race and running a regression of Points on Employer Race for bottom (column 1 and 3) and top (column 2 and 4) quantile of IAT score while controlling for demographics and employer perception. Standard errors in parenthesis. for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$.*

altruistic towards black employers vis-à-vis white employers while workers with higher implicit bias provide statistically similar effort to both employer groups.

3.7 Estimates of Behavioral Parameters

We designed our experiment to go with the structural model outlined in Section 2. The advantage of designing field experiments on the basis of a model of behavior is that it allows researchers to estimate the nuisance parameters in the environment that are relevant to decision making (DellaVigna, 2018). Because of the simplicity of our task, there are only three nuisance parameters we need to estimate. We use data from the piece rate treatments to identify these parameters. Subsequently, we estimate the deeper behavioral parameters of interest using data from the social preference treatments. We closely follow the estimation procedure in DellaVigna and Pope (2018).

3.7.1 Minimum-Distance Estimation

We first use minimum-distance estimation method to estimate these parameters. In minimum distance estimation, one identifies the set of moments in the data (average effort) and then finds the set of model parameters that minimizes the distance between the empirical moments and the theory-predicted moments. To estimate nuisance parameters, we use the average effort corresponding to the three piece rates (0 cents, 3 cents and 9 cents), to estimate γ, s, and k. Specifically, in the case of the power cost function, to estimate nuisance parameters, we use first moments from the piece rate treatments and solve the following equations

$$\bar{e}_p = \frac{1}{\gamma}\left[log(s+p) - log(k)\right] \text{ for } p \in \{0, 0.03, 0.09\}$$

where $\bar{e}_p$ is the average effort in the piece rate p treatment. These parameters estimates are used to draw the marginal cost and marginal benefit curve in Figure 3.7.1.

Once these parameters are estimated, we use average effort corresponding to altruism neutral, altruism black and altruism white treatment to estimate behavioral parameters $\alpha_{Neutral}$, α_{Black}, and α_{White} respectively. Specifically, for the power cost function, we solve the following equations for α_j for $j \in \{Neutral, Black, White\}$ taking estimates of γ, s, and k as given

$$log(\bar{e}_{\alpha_j}) = \frac{1}{\gamma}\left[log(s+\alpha_j v) - log(k)\right] \text{ for } j \in \{Neutral, Black, White\}$$

where $\bar{e}_{\alpha_j}$ is the average effort in the altruism j treatment.

Similarly, to calculate reciprocity parameters for neutral ($\rho_{Neutral}$), black (ρ_{Black}) and white (ρ_{White}) employers, we use average effort from reciprocity neutral, reciprocity black, and reciprocity white treatments and solve the following equations taking estimates of γ, s, k, and α_j for $j \in \{Neutral, Black, White\}$ as given:

$$log(\bar{e}_{\rho_j}) = \frac{1}{\gamma}\left[log(s+\rho_j+\alpha_j v) - log(k)\right] \text{ for } j \in \{Neutral, Black, White\}$$

where $\bar{e}_{\rho_j}$ is the average effort in the reciprocity j treatment.

Estimates using the exponential cost function are similarly calculated. Table 3.9 presents the parameter estimates for power cost function (column 1) and exponential cost function (column 3). The standard errors for these parameter estimates are estimated using a bootstrap procedure with a thousand draws.

3.7.2 Non-Linear Least Squares Estimation

The minimum distance estimator solely relies on the moment, and hence, does not use all the variation in the data. There are methods such as maximum likelihood and non-linear least squares that can be used to estimate these parameters using all the variation present in the data. We use non-linear least square method to estimate these parameters allowing for the heterogeneous cost of effort. Allowing for a heterogeneous marginal cost of effort in 3.3.1, we assume for a worker i, for a power cost case, $c(e_{ij}) = \frac{ke_{ij}^{1+\gamma}}{1+\gamma} exp(-\gamma\epsilon_{ij})$ with $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The first order condition 3.3.4 can then be written as;

$$s + 1_{Gift}\rho_j + \alpha_j v + p - ke_{ij}^{\gamma} exp(-\gamma\epsilon_{ij}) = 0$$

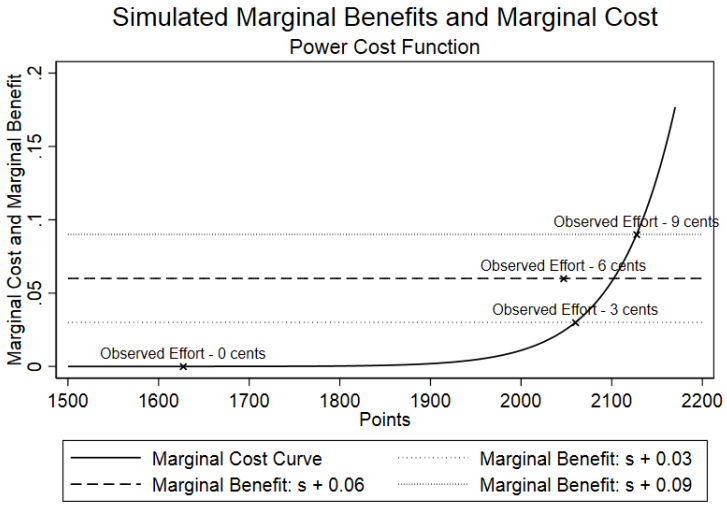


Figure 3.7.1: Illustration of the Model: Marginal Benefits and Cost Curves
Notes: The figure presents the marginal benefit and marginal cost curves using minimum-distance estimates for the power cost function.

Taking the last term to the right and taking logs, we obtain

$$log(s + 1_{Gift}\rho_j + \alpha_j v + p) + \epsilon_{ij} = log(k) + \gamma log(e_{ij}) - \gamma\epsilon_{ij}$$

Solving for $log(e_{ij})$, we obtain the estimating equation

$$log(e_{ij}) = \frac{1}{\gamma}\left[log(s + 1_{Gift}\rho_j + \alpha_j v + p) - log(k)\right] + \epsilon_{ij}. \tag{3.7.1}$$

Similarly using exponential cost function, we get

$$e_{ij} = \frac{1}{\gamma}\left[log(s + 1_{Gift}\rho_j + \alpha_j v + p) - log(k)\right] + \epsilon_{ij}. \tag{3.7.2}$$

Equations 3.7.1 and 3.7.2 can be estimated using non-linear least squares (NLS). Table 3.9 presents the NLS parameter estimates for power cost function (column 2) and exponential cost function (column 4). The NLS parameter estimates are nearly identical to those computed with minimum-distance estimation for the exponential cost case. The model predictions are also very similar .

The NLS estimates for the power cost function yield a lower curvature than the minimum-distance estimates ($\hat{\gamma}_{NLS} = 20.29$ versus $\hat{\gamma}_{MD} = 34.05$). The NLS model matches expected log effort, while the minimum-distance matches the log of expected effort. Both NLS and minimum-distance fit the in-sample moments and make similar predictions for the 6-cent piece rate treatment.

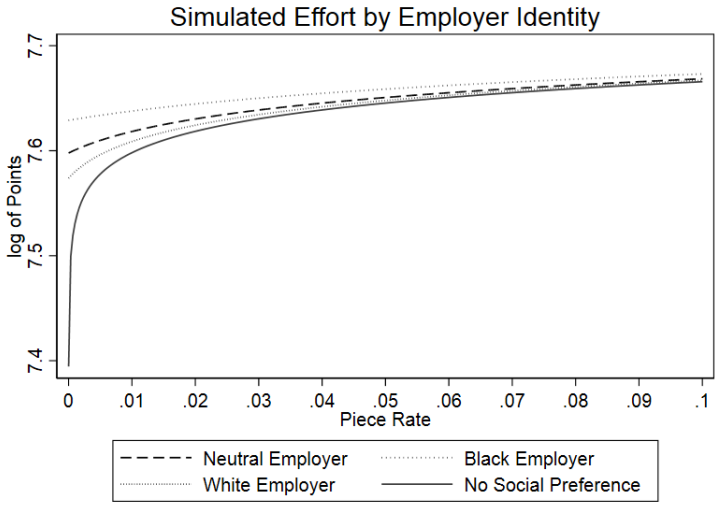


Figure 3.7.2: Simulated Effort by Employer Race at Different Piece Rates

Notes: The figure presents the simulated effort using the parameter estimates from table 3.9 *for power cost, minimum distance specification. Neutral/Black/White employer uses the respected social preference parameter estimates to calculate the predicted effort at each piece rate. No Social Preference assumes that altruism and reciprocity estimates are zero.*

The parameter estimate for 'altruism black' is significantly higher than 'altruism white' in all the specifications, indicating that white workers have significantly higher altruistic preferences for black employers as compared to white employers. The reciprocity estimates indicate a null effect from the gift for any employer in all the specifications. Even though the parameter values are close to zero, but they translate to meaningful differences in effort provided to black and white employers at the zero piece rate. Figure 3.7.2 presents the simulated effort for neutral, black and white employer using parameter estimates along with zero social-preference case. Black employers receive around five percent higher effort than white employers at the zero piece rate. The difference between black and white employers becomes negligible at higher piece rates because workers respond much more to monetary incentives as compared to social preferences.

3.8 Conclusion

Economic historians record a time in US labor history when white workers openly militated against receiving orders from (or working under) black supervisors. Things have changed. "[While overt racism was implicated in the past, it is behavioral differences that lie at the root of racial inequality in contemporary America" (Loury, 1998). What are these behavioral differences? Now that overt racism is either deemed illegal or too difficult to practice openly, have white workers stopped discriminating against black employers? This paper uses insights from behavioral and experimental economics to shed light on this enduring issue in

Table 3.9: Parameter Estimates

	Power Cost		Exponential Cost	
	Minimum Distance	NLS	Minimum Distance	NLS
	(1)	(2)	(3)	(4)
Baseline Parameters				
γ	34.05	20.30	0.0163	0.0163
	(7.3)	(8.85)	(.0102)	(.00807)
s	0.00000977	0.00000802	0.0000264	0.0000264
	(.000199)	(.000032)	(.000327)	(.000101)
k	4.50e-115	2.98e-70	8.58e-17	8.58e-17
	(3.6e-49)	(2.5e-68)	(3.1e-09)	(1.5e-15)
Altruism Parameters				
α_{Neutral}	0.00983	0.000426	0.0156	0.0156
	(.00739)	(.0017)	(.00966)	(.0427)
α_{Black}	0.0285	0.000776	0.0402	0.0402
	(.0181)	(.00274)	(.0214)	(.0953)
α_{White}	0.00413	0.000270	0.00722	0.00722
	(.00367)	(.00129)	(.00531)	(.0215)
Reciprocity Parameters				
ρ_{Neutral}	0.0000676	0.0000272	0.0000921	0.00124
	(.000134)	(.000103)	(.000182)	(.00318)
ρ_{Black}	0.0000307	0.0000395	0.0000381	0.00220
	(.000271)	(.00014)	(.000321)	(.00513)
ρ_{White}	0.000243	0.0000255	0.000328	0.00200
	(.000216)	(.0001)	(.00027)	(.00477)
Implied effort - 6-cents	2102	7.746	2102	2102.4

Notes: This table reports the structural estimates of the model in section 2. Column (1) and (3) use a minimum-distance estimator employing three moments (average effort in three piece rate treatments) and three parameters (γ , s and k), and is thus exactly identified. Column (2) and (4) use a non-linear least squares employing individual effort in all the treatments and thus estimating all the parameters simultaneously. We use power cost (column 1 and 2) and exponential cost (column 3 and 4) function to estimate the model. Implied effort is calculated using estimated parameters for each model. The observed effort for 6-cents treatment is 2047 points or log 7.624. For the altruism parameters, the baseline parameters are taken as given and the average effort for neutral, black, and white employers is used to estimate $\alpha_{Neutral}$, α_{Black}, and α_{White} from the altruism treatments. Similarly for the reciprocity parameters, the baseline and altruism parameters are taken as given and the average effort corresponding to reciprocity neutral, reciprocity black, and reciprocity white is used to estimate $\rho_{Neutral}$, ρ_{Black}, and ρ_{White}. Standard errors for minimum-distance estimator are calculated by taking a bootstrap sample of 1000 draws and recalculating these parameters for each draw.

American labor markets. The narrower question we ask is, do workers with a considerable degree of discretion over work effort display differential, race-dependent social preferences toward their out-race employers?

The experimental setting is an online labor market - Amazon's Mechanical Turk (M-Turk). In this online economy, workers and employers are at arms length and the worker is involved in a real-effort task for a pre-assigned, non-fictitious, black or white employer. The possibility of race-dependent social preferences is activated by unobtrusively revealing the employer's race to the matched worker. We detect statistically significant evidence for altruism: workers put more effort when they know their work benefits the employer (altruism neutral treatment) as compared to a treatment where neither the worker nor the employer benefits from worker effort. Most importantly, white workers are significantly more altruistic towards black employers than white employers. Not only is this finding statistically significant at the 2% level, the difference in effort provision is economically powerful as well. There is suggestive evidence that the higher effort towards black employers is driven by workers with relatively low implicit bias against blacks.

Our results suggest that preference-based discrimination against minorities may dampen as traditional labor markets get replaced with gig economy ones. Indeed, our results are roughly in line with a new body of research that finds a general erosion of racially-motivated discrimination in hiring in U.S. labor markets. Lahey and Oxley (2018) finds while "younger white applicants are preferred to younger black applicants, this preference diminishes with age as white applicants become less attractive and black applicants become more attractive. Indeed, we find no preference for white compared to black applicants in their 50s, and black applicants are even preferred in some specifications."

That is not to say that racial discrimination is disappearing or will soon, nor do we suggest that pro-social behavior of whites towards blacks is omnipresent. Indeed, the pro-sociality may vanish in settings where employer-worker engagement is longer and involves physical interaction. Likewise, we recognize that unlike the current focus on the intensive margin of worker effort, understanding social preferences on the extensive margin may be equally important; after all, it is possible workers from the dominant group may systematically select out of (not even apply for) jobs posted by disadvantaged-group employers, thereby limiting the labor resources at the disposal of said employers. In short, if workers are given agency in who they work for, they may well avoid out-race employers. We aim to study the extensive margin angle to worker-to-employer discrimination in future research. Our work also does not suggest that asking whether workers differentially treat their out-race bosses in traditional labor markets is not interesting. Far from it. The questions we ask, could in principle, be asked in the sort of field setting studied in Breza, Kaur, and Shamdasani (2017) where the researchers set up their own factory workshops in Odisha, India to employ 378 workers full-time for one month in seasonal manufacturing jobs. We leave this to future research.

We recognize that we do not offer a clear answer to the question, what explains the pro-social behavior of white workers towards black employers? That question deserves full attention and it will in our future work. An explanation that is sometimes offered – social desirability bias: people want to appear as nice (non-racist) to the experimenter which is why they put more effort for black employers. For one, the identity of the experimenters was never made salient to the subjects. Second, in order for a white subject to put more effort for black relative to white employers because of his social desirability bias, he must know what the experimenter deems as appropriate levels of effort toward black and white employers. Also bear in mind that the game ends after the worker has finished the task. Subsequent to that, there is no possibility of communication (and everyone knows this ex ante) between the employer and the worker implying the worker could not be motivated by a desire to signal anything to the employer. In particular, given the one-shot nature of the encounter, even if the worker has an innate preference to be well regarded by the other race, there is no possibility for the other race to react to that preference.

Another concern might be that worker subjects, knowing they are part of an experiment, may be responding to experimenter demand effects or, more accurately, the Hawthorne effect. Our contention is that this concern is unlikely to be critical for the following reasons: 1) our treatment revelation mechanism is decidedly subtle, and taken in tandem with the between-subject design, it is almost impossible for subjects to pinpoint the true purpose of the study to be race related; 2) even if there are demand effects they are likely to be mild – the recent paper de Quidt, Haushofer, and Roth (2018) finds that the range of (weak) demand effect for our 'a-b' task is expected to not exceed 11 percent of the treatment effects, which means the differential effort choices for black employer relative to the white, is likely to be between 85.33 and 95.27.[25] Taken together, these arguments imply that demand effects, if any, likely have a modest impact on our results.

[25] Arguably, the demand effects in de Quidt et al. (2018) are not entirely driven by race. Relatedly, Mummolo and Peterson (2019) find, using a vignette study approach, that the experimenter demand effect in studies on racial discrimination is modest.

3.9 Appendix A: Miscellaneous Figures

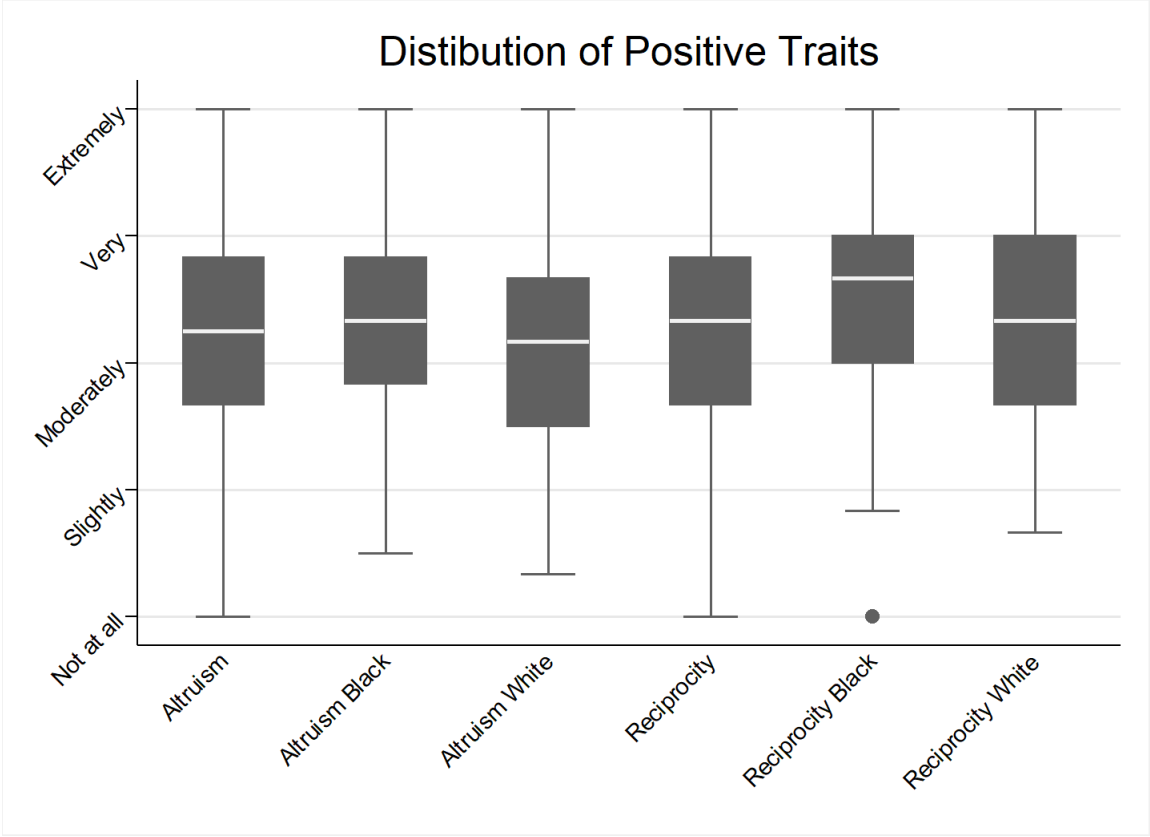


Figure A1: Perception of Positive Personality Traits

Notes: This figure presents the box-plot of average of positive traits as rated by the evaluators. After the evaluators watched the video they were asked "Please rate the following characteristics about the the person in the above video". The positive traits were friendliness, confidence, encouragement, trustfulness, clarity, and motivation .

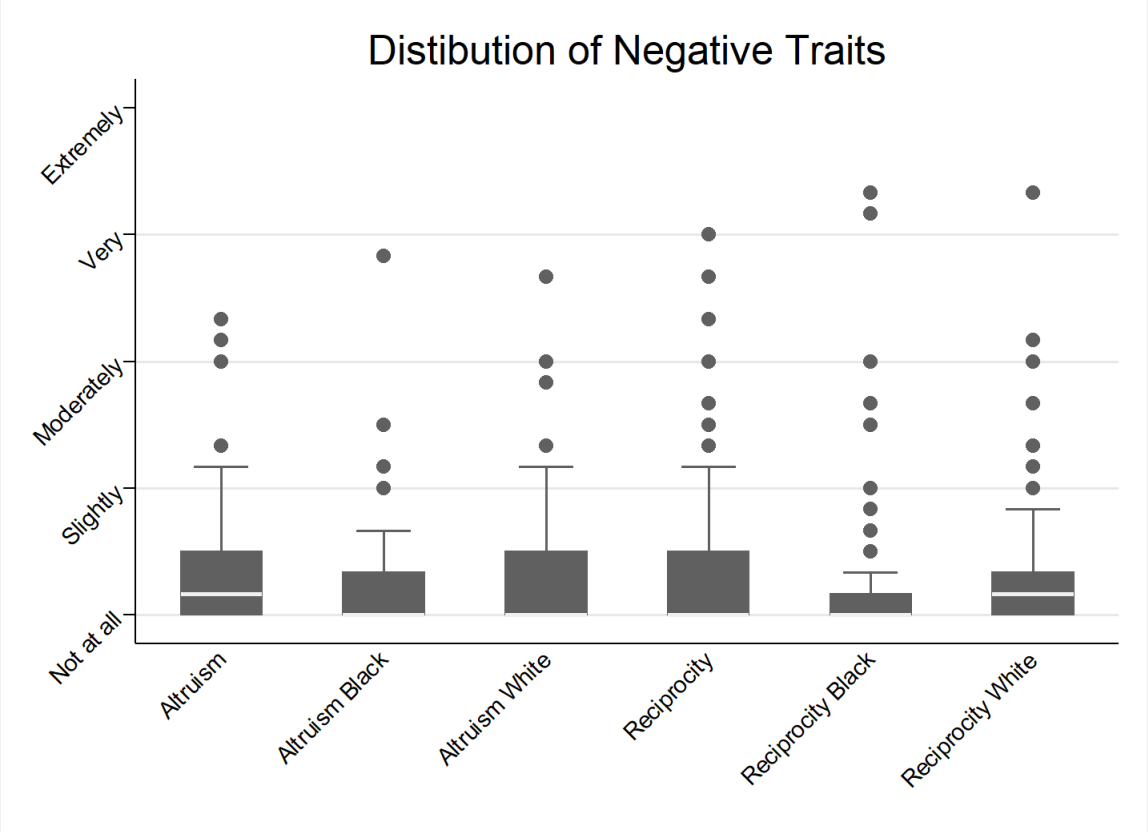


Figure A2: Perception of Negative Personality Traits

Notes: This figure presents the box-plot of average rating of negative traits by the evaluators. After the evaluators watched the video they were asked "Please rate the following characteristics about the person in the above video". The negative traits were arrogance, laziness, bossiness, rudeness, hostility, and undermining.

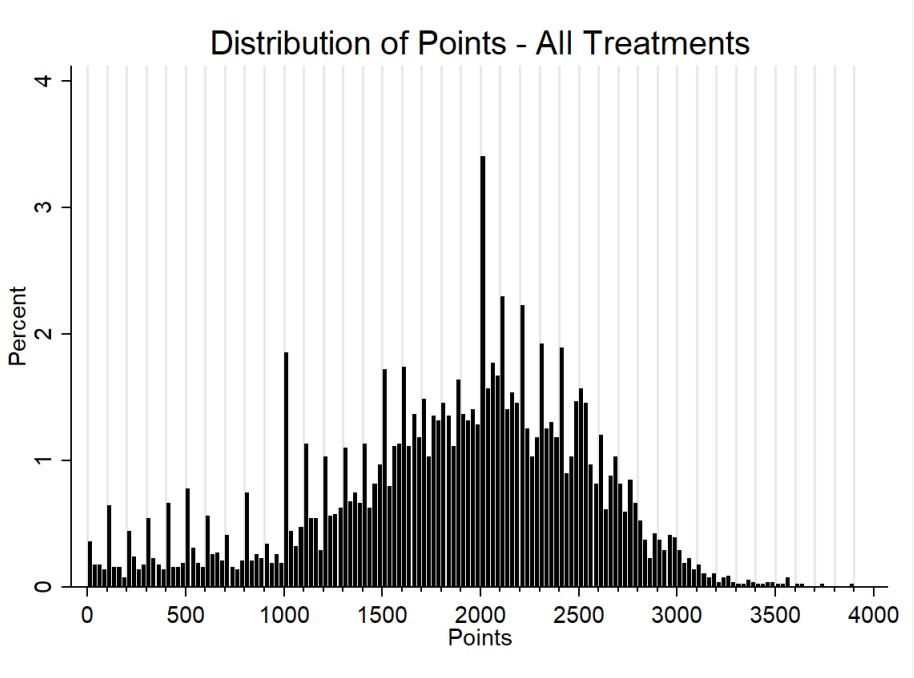


Figure A3: Distribution of effort

Notes: This figure plots a histogram of the observed points over all 10 treatments.

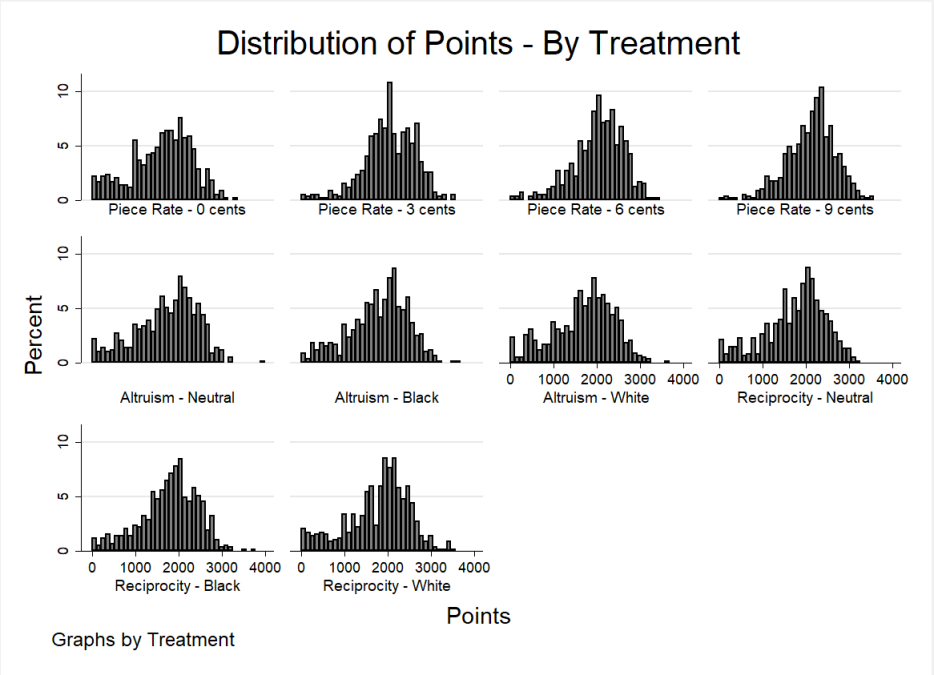


Figure A4: Distribution of effort by Treatment

Notes: This figure plots a histogram of the observed points by each of the 10 treatments.

3.10 Appendix B: Miscellaneous Tables

Table B1: Demographic Information of Employer Subjects

	(1) All Subjects	(2) Blacks	(3) Whites
Gender			
Male	1.00	1.00	1.00
Female	0.00	0.00	0.00
Race			
Black or African American	0.50	1.00	0.00
White or Caucasian	0.50	0.00	1.00
Age			
18-24	0.78	0.61	0.94
25-34	0.14	0.22	0.06
35-44	0.06	0.11	0.00
45-54	0.03	0.06	0.00
Education			
High school or equivalent	0.06	0.00	0.11
Some college	0.64	0.50	0.78
College graduate	0.19	0.28	0.11
Master's degree	0.08	0.17	0.00
Doctoral degree	0.03	0.06	0.00
Most lived state			
Blue	0.28	0.22	0.33
Red	0.03	0.06	0.00
Swing	0.69	0.72	0.67
Observations	36	18	18

Notes: The table presents demographic information of employer subjects. Column (1) presents proportion of all the employer subjects by their gender, race, age and education. Column (2) and column (3) presents these information for only black and white employers respectively.

Table B2: Test of Difference of Perception of Race and Skin Color

Panel A: Average Perception of Race

	(1) Race Perception		
	Proportion	SE	Group
Altruism	0.29	(0.03)	1
Altruism Black	0.91	(0.03)	23
Altruism White	0.98	(0.03)	3
Reciprocity	0.26	(0.03)	1
Reciprocity Black	0.83	(0.03)	2
Reciprocity White	0.96	(0.03)	3
Degrees of Freedom	1016		

Panel B: Average Perception of Skin Color

	(1) Skin Color Perception		
	Mean	SE	Group
Altruism Black	4.81	(0.05)	1
Altruism White	2.05	(0.05)	2
Reciprocity Black	4.57	(0.05)	1
Reciprocity White	2.11	(0.05)	2
Degrees of Freedom	667		

Notes: Panel A presents the proportion of subjects who could correctly guess the race of the employer in the video. Panel B presents the average skin color as perceived by the subjects in each treatment. The skin color can vary from 1 to 6 where 1 represents the 'light, pale white' while 6 represents the 'very dark brown to black' skin tone. Proportions sharing a digit in the 'Group' column are not significantly different at the 5% level. The comparisonwise error rate is adjusted using the Bonferroni method.

Table B3: Test of Difference of Personality Traits

	(1) Positive Traits			(2) Negative Traits		
	Mean	SE	Group	Mean	SE	Group
Altruism	3.27	(0.07)	12	1.33	(0.04)	12
Altruism Black	3.33	(0.07)	12	1.19	(0.04)	1
Altruism White	3.15	(0.07)	1	1.30	(0.04)	12
Reciprocity	3.26	(0.07)	12	1.38	(0.04)	2
Reciprocity Black	3.51	(0.07)	2	1.24	(0.04)	12
Reciprocity White	3.28	(0.07)	12	1.28	(0.04)	12
Degrees of Freedom	852			929		

Notes: The table presents the average of perceived positive and negative traits across the social preference treatments. The perception of the trait can vary from 1-Not at all to 5-Extremely. Positive Trait is constructed by taking an average of the ratings on; friendliness, confidence, encouragement, trustfulness, clarity, and motivation . Negative Trait is constructed by taking an average of the ratings on; arrogance, laziness, bossiness, rudeness, hostility, and undermining. Means sharing a digit in the group label are not significantly different at the 5% level. The comparisonwise error rate is adjusted using the Bonferroni method.

Table B4: Balance Check

	$\chi 2$ (p-value)
Gender	
Female	8.414 (0.493)
Age	
25-30	11.03 (0.273)
31-40	10.98 (0.277)
41-50	14.95 (0.0924)
51-64	11.04 (0.273)
65 and over	10.19 (0.335)
Education	
High school or equivalent	3.744 (0.927)
Some college	2.884 (0.969)
College graduate	3.511 (0.941)
Graduate or professional degree	2.753 (0.973)
Income	
$20,000 - $44,999	6.928 (0.645)
$45,000 - $99,999	10.38 (0.321)
$100,000 - $149,999	10.13 (0.340)
$150,000+	11.01 (0.275)
Most lived US State	
Blue	4.953 (0.838)
Red	9.193 (0.420)
Party	
Democrat	5.939 (0.746)
Republican	12.65 (0.179)
Observations	5945

Notes: The table presents the $\chi 2$ and corresponding p-values of the likelihood ratio (LR) test of the equality of each coefficient from multinomial-logit regression of Treatment status on the demographic variables.

Table B5: Tests of Equality of Distributions

	0-cents	3-cents	6-cents	9-cents	A-Neutral	A-Black	A-White	R-Neutral	R-Black
0-cents	1.000	.	.	.	.	.	.	.	.
3-cents	0.000	1.000	.	.	.	.	.	.	.
6-cents	0.000	0.484	1.000	.	.	.	.	.	.
9-cents	0.000	0.001	0.003	1.000	.	.	.	.	.
A-Neutral	0.004	0.000	0.000	0.000	1.000	.	.	.	.
A-Black	0.001	0.000	0.000	0.000	0.624	1.000	.	.	.
A-White	0.147	0.000	0.000	0.000	0.302	0.074	1.000	.	.
R-Neutral	0.001	0.000	0.000	0.000	0.716	0.800	0.088	1.000	.
R-Black	0.000	0.000	0.000	0.000	0.202	0.613	0.092	0.821	1.000
R-White	0.000	0.000	0.000	0.000	0.131	0.442	0.016	0.629	0.415

Notes: The table presents the p-values from the pairwise comparison of distribution functions of points scored for each pair of treatments using the Kolmogorov-Smirnov test. "A" and "R" in the column and row headings represent Altruism and Reciprocity respectively.

Table B6: Quantile Regression Results

	Altruism	Reciprocity
	(1)	(2)
q25		
White	-198.6***	-69.30
	(71.93)	(87.97)
Constant	1207.9***	1126.5**
	(357.2)	(491.2)
q50		
White	-101.0*	3.073
	(54.53)	(43.99)
Constant	1745.1***	1702.0***
	(279.5)	(436.7)
q75		
White	-42.92	31.67
	(50.30)	(38.00)
Constant	1929.7***	2376.8***
	(259.3)	(487.3)
Demographic Controls	Yes	Yes
Employer Perception	Yes	Yes
Observations	1030	1003

*Notes: The table presents the result of the 0.25, 0.5, and 0.75 simultaneous quantile-regression. The estimate of VCE is obtained via bootstrapping, and the VCE includes between-quantile blocks. Standard errors in parentheses. * for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$.*

Table B7: Social Preference Treatments - Robustness, Employer Race Correctly Perceived

	Altruism			Reciprocity		
	(1)	(2)	(3)	(4)	(5)	(6)
Black or African American	66.90	72.62	61.40	6.968	0.753	-31.62
	(60.68)	(61.44)	(62.46)	(59.72)	(60.94)	(62.29)
White or Caucasian	-25.54	-18.04	-31.36	-8.057	-7.958	-47.19
	(59.79)	(60.63)	(61.53)	(58.35)	(59.42)	(61.23)
Constant	1740.8***	1552.8***	1708.2***	1811.8***	1764.4***	1708.9***
	(52.26)	(295.4)	(306.2)	(50.67)	(292.5)	(305.5)
Demographic Controls	No	Yes	Yes	No	Yes	Yes
Employer Perception	No	No	Yes	No	No	Yes
N	1223	1167	1164	1214	1166	1164
Black - White	92.45**	90.66**	92.76**	15.02	8.711	15.57
	(42.36)	(43.30)	(43.68)	(42.86)	(43.82)	(44.21)

Notes: The table presents the estimates from an OLS regression of Points in the social preference treatments on the employer's race for workers who could correctly perceive the race of the employer. The omitted category is the employer with concealed race. Demographic controls include age, gender, education, income, political affiliation and the voting pattern of the most lived state (red, blue, or swing) of the worker. Employer Perception include worker's belief about the income, age, and education of the employer. Black - White represents the difference in the coefficients of black and white employers in each model. Standard errors in parentheses. for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$.*

Table B8: Overall Productivity by Demographics

	(1) Points
Gender	
Female	-135.42*** (17.77)
Age	
25-30	-26.53 (29.58)
31-40	-83.18*** (31.39)
41-50	-126.63*** (35.09)
51-64	-257.55*** (40.42)
65 and over	-356.25*** (58.48)
Education	
Some college	1.78 (29.12)
College graduate	-96.92*** (28.06)
Graduate or professional degree	-97.23*** (32.92)
Prefer not to answer	-1260.07*** (472.82)
Income	
$20,000 - $44,999	33.00 (25.98)
$45,000 - $99,999	40.73 (26.24)
$100,000 - $149,999	84.57** (37.01)
$150,000+	91.32* (54.65)
Party	
Democrat	-60.48*** (20.59)
Republican	-25.35 (22.64)
Most lived US State	
Blue	-47.50** (20.02)
Red	-13.10 (23.06)
Constant	2074.68*** (38.74)
Observations	5945
R^2	0.034
F	11.68

*Notes: The table presents the estimates of an OLS regression of points scored on worker demographics. The reference category for gender, age, education, income, party, and most lived state is male, 18 - 24, below some college, below $20,000, independent, and swing state respectively. Standard errors in parentheses. * for $p < 0.10$, ** for $p < 0.05$, and *** for $p< 0.01$.*

3.11 Appendix C: IRB Approval Letter

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 03/25/2019

To: Sher A Asad

From: Office for Responsible Research

Title: **Discrimination in Fairness: Evidence from an Online Labor Market**

IRB ID: **18-201**

Submission Type: Modification **Review Type**: Full Committee

Approval Date: 03/25/2019 **Approval Expiration Date: 07/17/2020**

The project referenced above has received approval from the Institutional Review Board (IRB) at Iowa State University according to the dates shown above. Please refer to the IRB ID number shown above in all correspondence regarding this study.

To ensure compliance with federal regulations (45 CFR 46 & 21 CFR 56), please be sure to:

- **Use only the approved study materials** in your research, including the **recruitment materials and informed consent documents that have the IRB approval stamp**.
- **Retain signed informed consent documents for 3 years after the close of the study,** when documented consent is required**.**
- **Obtain IRB approval prior to implementing any changes** to the study or study materials.
- **Promptly inform the IRB of any addition of or change in federal funding for this study.** Approval of the protocol referenced above applies only to funding sources that are specifically identified in the corresponding IRB application.
- **Inform the IRB if the Principal Investigator and/or Supervising Investigator end their role or involvement with the project** with sufficient time to allow an alternate PI/Supervising Investigator to assume oversight responsibility. Projects must have an eligible PI to remain open.
- **Immediately inform the IRB of** (1) **all serious and/or unexpected adverse experiences** involving risks to subjects or others; and (2) **any other unanticipated problems involving risks** to subjects or others.
- IRB approval means that you have met the requirements of federal regulations and ISU policies governing human subjects research. **Approval from other entities may also be needed.** For example, access to data from private records (e.g., student, medical, or employment records, etc.) that are protected by FERPA, HIPAA, or other confidentiality policies requires permission from the holders of

those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **IRB approval in no way implies or guarantees that permission from these other entities will be granted.**

- Your research study may be subject to **post-approval monitoring by Iowa State University's Office for Responsible Research**. In some cases, it may also be subject to formal audit or inspection by federal agencies and study sponsors.

- Upon completion of the project, transfer of IRB oversight to another IRB, or departure of the PI and/or Supervising Investigator, please initiate a Project Closure to officially close the project. For information on instances when a study may be closed, please refer to the IRB Study Closure Policy.

If your study requires continuing review, indicated by a specific Approval Expiration Date above, you should:

- **Stop all human subjects research activity if IRB approval lapses**, unless continuation is necessary to prevent harm to research participants. Human subjects research activity can resume once IRB approval is re-established.

- **Submit an application for Continuing Review** at least three to four weeks prior to the **Approval Expiration Date** as noted above to provide sufficient time for the IRB to review and approve continuation of the study. We will send a courtesy reminder as this date approaches.

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.

3.12 Appendix D: Experiment Material

Iowa State University
Department of Economics
Consent for Participation in Research

Title of Study: Decisions in Labor Market
Investigator: Sher Afghan Asad, Ritwik Banerjee, Joydeep Bhattacharya

This brief screener is a part of a research project at Iowa State University. You will receive $0.05 for completing the screener, which is used to see if you are eligible for the full study. Individuals who qualify for the study will be invited to participate in a 15-minute study for the pay of 1 dollar plus bonus. If you do not qualify for participation based on this screening questionnaire, all the information about you will be destroyed.

Description of Procedures
To be considered for participation in the study, you will have to answer a few demographic questions. Once you have answered those questions, you may be invited to participate in the full study. In the full study, you may be randomly matched with another participant and you will then work on a simple task that may affect your and your matched participant earnings. The experiment will last for approximately 15 minutes. You will be given more information about the structure of the study in the instructions.

Risks or Discomforts
There are no foreseeable risks currently in participating in the study.

Benefits
If you decide to participate in this study, there are no direct benefits to you. It is hoped that the information gained in this study will benefit the field of economics by providing more insight into the process of how decisions are made in the labor markets.

Costs and Compensation
You will not bear any costs from participating in this study. If you participate you will spend no longer than 15 minutes completing procedures. Participants will earn $1 for participating in the experiment and a bonus amount depending on the decisions in the experiment. Your final compensation will vary depending on your and your randomly matched participant choices.

Participant Rights

Participating in this study is completely voluntary. You may choose not to take part in the study or to stop participating at any time, for any reason, without penalty or negative consequences. If you have any questions about the rights of research subjects or research-related injury, please contact the IRB Administrator, 515-294-4566, IRB@iastate.edu, or Director, 515-294-3115, Office for Responsible Research, Iowa State University, Ames, Iowa 50011.

Confidentiality

This consent form and any other documents identifying participants will be kept confidential to the extent permitted by applicable laws and regulations and will not be made publicly available. However, federal government regulatory agencies, auditing departments of Iowa State University, and the Institutional Review Board (a committee that reviews and approves human subject research studies) may inspect and/or copy study records for quality assurance and data analysis. These records may contain private information. This experiment is approved by the Institutional Review Board at Iowa State University (ISU IRB: 18-201-01, Approved Date: 03/25/2019, Expiration Date: 07/17/2020). It is assured that the confidentiality of your data and the choices that you make in the study will be strictly maintained. To ensure confidentiality to the extent permitted by law, the following measures will be taken: Data will be stored on a secure cloud-based drive (Dropbox) under password protection. Your identifiable information will be separated from your decisions in the experiment. When we report results, we will group responses in aggregate; individual responses will not be shared. Please be aware that any work performed on Amazon MTurk can potentially be linked to information about you on your Amazon profile. We will not be accessing any information about you that you may have put on your Amazon public profile page. We will store your MTurk worker ID separately from the other information you provide to us.

Future Use of Data

De-identified information collected about you during this study may be shared with other researchers or used for future research studies. We will not obtain additional informed consent from you before sharing the de-identified data.

Questions

You are encouraged to ask questions at any time during this study. For further information about the study, contact Sher Afghan Asad at 515-735-6309 or saasad@iastate.edu or Joydeep Bhattacharya at joydeep@iastate.edu.

Consent and Authorization Provisions

By clicking the box below, you acknowledge, that you voluntarily agree to participate in this

study, that the study has been explained to you, that you have been given the time to read the document, and that your questions have been satisfactorily answered. You may print a copy of this informed consent document for your records.

If you don't agree with this consent document, then close this form and return the HIT.

◯ I acknowledge that I have read the material above and I agree to participate in the study.

Subjects who consent to participating in the study will fill out this screener survey before being considered for participation in this study.

Thank you for participating. Now that you have started, **you may not restart** this survey at any point or else your HIT will be rejected.

Please answer the following questions to the best of your ability.

Gender you most closely identify with:

- ◯ Male
- ◯ Female
- ◯ Prefer not to answer
- ◯ Other

[ ]

Race you most closely identify with:

- ◯ American Indian or Alaskan Native
- ◯ Asian
- ◯ Black or African American
- ◯ Hispanic or Latino
- ◯ Native Hawaiian or other Pacific Islander
- ◯ White or Caucasian
- ◯ Prefer not to answer
- ◯ Other

[ ]

If “White or Caucasian” is not selected, survey will end with 5 cents compensation.

Age (in years):

- ◯ Under 18
- ◯ 18 - 24
- ◯ 25 - 34
- ◯ 35 - 44
- ◯ 45 - 54
- ◯ 55 - 64
- ◯ 65 or older
- ◯ Prefer not to answer

If "Under 18" is selected, survey will end with 5 cents compensation.

Highest education level reached:

- ◯ Less than high school
- ◯ High school or equivalent
- ◯ Vocational / Technical School
- ◯ Some college
- ◯ College graduate
- ◯ Master's degree
- ◯ Professional degree
- ◯ Doctoral degree
- ◯ Prefer not to answer

Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else?

- ◯ Republican
- ◯ Democrat
- ◯ Independent
- ◯ Other

[ ]

- ◯ No preference

Annual pre-tax income

- ◯ Less than $10,000
- ◯ $10,000 - $19,999
- ◯ $20,000 - $29,999
- ◯ $30,000 - $44,999
- ◯ $45,000 - $99,999
- ◯ $100,000 - $149,999
- ◯ $150,000 - $199,999
- ◯ $200,000+
- ◯ Prefer not to answer

In which US state have you resided the longest?

[ ]

Subjects who report their race as "White or Caucasian", age as above 18, and their device type is not mobile will be shown the following screen. Rest of them will be shown the exit screen with 5 cents compensation.

Congratulations! You meet the criteria to participate in the full study.

This study will take up to 10 minutes, pay a bonus of 1 dollar and possibly an additional amount depending on your decisions in the study.

Make sure that you are not distracted for the next 10 minutes. Once you click next, you may not restart this study at any point or else your HIT will be rejected. When you are ready, click next to begin.

You may have to click the next button multiple times to move forward.

Participants will be blocked randomized to one of the ten treatments when they click next.

Instructions for each treatment will be explained in the video.

The script of each video will differ only on the incentive and bonus structure, the video format will be same for each treatment. The video will only show the hands of the other participant demonstrating the task. The skin will be revealed/concealed (using gloves) in the video depending on the assigned treatment. The next few pages presents the interface for each treatment.

Instructions for piece rate treatments. The videos have the hands covered in gloves and the audio is muted.

The following video explains what you are supposed to do in this study. You **MUST watch** this ~1-minute video to continue with the study.

The video has no sound, please carefully read the captions.

Below is an example of how the task will work. Try pressing `a' and `b' alternatively to score points. We have limited the point total below to a maximum of 5 as this is just practice, but the overall task will not have a limit.

Press `a' then `b'...

Points: 0

Proceed to the next page when you are ready to play the task. Your 10-minute task will begin immediately when the page loads.

The next button will appear only after you have finished watching a video. PLEASE WATCH AND LISTEN TO THE VIDEO TO CONTINUE.

Instructions for race neutral treatments. The videos have the hands covered in gloves and the audio is muted.

The following video explains what you are supposed to do in this study. You **MUST watch** this ~1-minute video to continue with the study.

The person in the video is **another participant** in the study. The video has no sound, please carefully read the captions.

The payment to the other participant will be paid in a couple of weeks. The proof of payment will be posted here. The ID of your other participant (assigned by us) is 18.

Below is an example of how the task will work. Try pressing `a' and `b' alternatively to score points. We have limited the point total below to a maximum of 5 as this is just practice, but the overall task will not have a limit.

Press `a' then `b'...

Points: 0

The next page will ask you some questions about the other participant. You will play the task after answering those questions.

The next button will appear only after you have finished watching a video. PLEASE WATCH AND LISTEN TO THE VIDEO TO CONTINUE.

Instructions for race salient treatments. The videos have the bare hands and the audio is not muted.

The following video explains what you are supposed to do in this study. You **MUST watch** this ~1-minute video to continue with the study.

The person in the video is **another participant** in the study.

The payment to the other participant will be paid in a couple of weeks. The proof of payment will be posted here. The ID of your other participant (assigned by us) is 62.

Below is an example of how the task will work. Try pressing `a' and `b' alternatively to score points. We have limited the point total below to a maximum of 5 as this is just practice, but the overall task will not have a limit.

Press `a' then `b'...

Points: 0

The next page will ask you some questions about the other participant. You will play the task after answering those questions.

The next button will appear only after you have finished watching a video. PLEASE WATCH AND LISTEN TO THE VIDEO TO CONTINUE.

These questions are presented only in the race salient and race neutral treatments.

Before you play the task, please give your **best guess** about the participant in the video. For each question, you will be paid **an extra 2 cents** as bonus if your guess is correct, we will **deduct 2 cents** from your final bonus payment if your guess is incorrect. Select "Cannot decide" if you cannot decide between the two options, in which case **no extra amount** will be rewarded or deducted for that question.

The other participant is either male or female, please guess the gender of the other participant?

- ◯ Male
- ◯ Female
- ◯ Cannot decide

The other participant's income is either less than or greater than $45,000, please guess the income of the other participant?

- ◯ Less than $45,000
- ◯ Greater than $45,000
- ◯ Cannot decide

The other participant's education is either 'below college' or 'some college or above', please guess the highest education level attained by the other participant.

- ◯ Below college
- ◯ Some college or above
- ◯ Cannot decide

The other participant is either black or white, please guess the race of the other participant?

- ◯ Black or African American
- ◯ White or Caucasian
- ◯ Cannot decide

The other participant is either 'under 35' or '35 or above', please guess the age group of the the other participant?

- ◯ Under 35
- ◯ 35 or above
- ◯ Cannot decide

Proceed to the next page when you are ready to play the task. Your 10-minute task will begin immediately when the page loads.

Task screen for Altruism Black treatment

0845

Press 'a' then 'b'...

Points: 155

Your bonus payout: $1

Other participant's earning: $ 0.016

The other participant will be paid 1 cent for every 100 points that you score.

Your score will not affect your payment in any way.

Demonstration of the task by the other participant

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Altruism Neutral treatment

0 9 1 8

Press 'a' then 'b'...

Points: 110

Your bonus payout: $1
Other participant's earning: $ 0.011

The other participant will be paid 1 cent for every 100 points that you score.

Your score will not affect your payment in any way.

Demonstration of the task by the other participant

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Altruism White treatment

0924

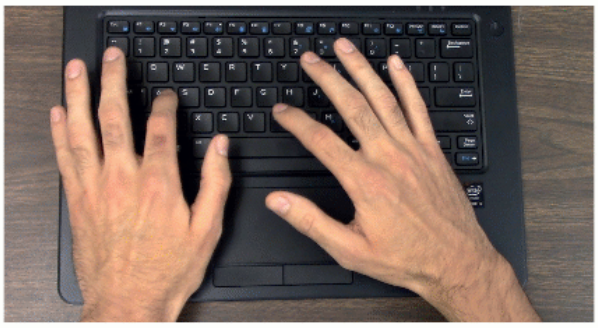
Press 'a' then 'b'...

Points: 132

Your bonus payout: $1
Other participant's earning: $ 0.013

The other participant will be paid 1 cent for every 100 points that you score.

Your score will not affect your payment in any way.



Demonstration of the task by the other participant

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Piece Rate - 0 cents treatment

0938

Press 'a' then 'b'...

Points: 57

Your bonus payout: $1

Your score will not affect your payment in any way.

Demonstration of the task

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Piece Rate - 3 cents treatment

0934

Press 'a' then 'b'...

Points: 44

Your bonus payout: $1 + 0.013

As a bonus, you will be paid an extra 3 cents for every 100 points that you score.

Demonstration of the task

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Piece Rate - 6 cents treatment

0 9 0 0

Press 'a' then 'b'...

Points: 38

Your bonus payout: $1 + 0.023

As a bonus, you will be paid an extra 6 cents for every 100 points that you score.

Demonstration of the task

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Piece Rate - 9 cents treatment

0916

Press 'a' then 'b'...

Points: 68

Your bonus payout: $1 + 0.061

As a bonus, you will be paid an extra 9 cents for every 100 points that you score.

Demonstration of the task

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Reciprocity Black treatment

0929

Press 'a' then 'b'...

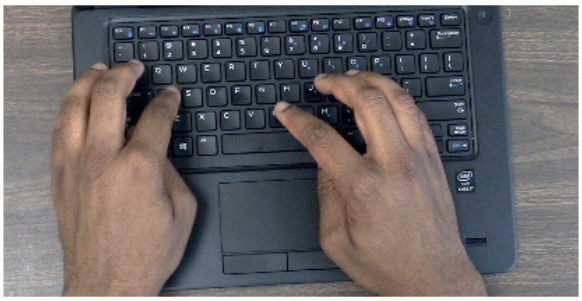
Points: 117

Your bonus payout: $1 + 0.2
Other participant's earning: $ 0.012

The other participant will be paid 1 cent for every 100 points that you score.

In appreciation to you for performing this task, the other participant has decided to pay you an extra 20 cents as a bonus.

Your score will not affect your payment in any way.



Demonstration of the task by the other participant

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Reciprocity Neutral treatment

0914

Press 'a' then 'b'...

Points: 114

Your bonus payout: $1 + 0.2
Other participant's earning: $ 0.011

The other participant will be paid 1 cent for every 100 points that you score.

In appreciation to you for performing this task, the other participant has decided to pay you an extra 20 cents as a bonus.

Your score will not affect your payment in any way.

Demonstration of the task by the other participant

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Task screen for Reciprocity White treatment

0 9 0 8

Press 'a' then 'b'...

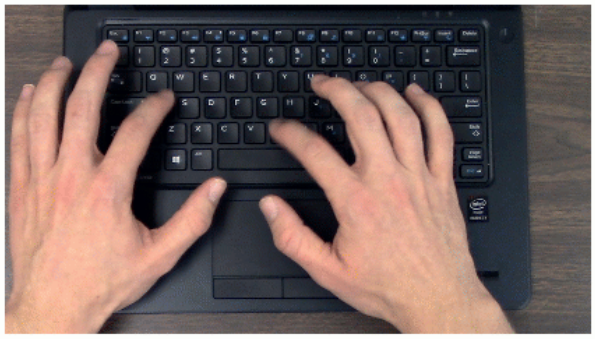
Points: 138

Your bonus payout: $1 + 0.2
Other participant's earning: $ 0.014

The other participant will be paid 1 cent for every 100 points that you score.

In appreciation to you for performing this task, the other participant has decided to pay you an extra 20 cents as a bonus.

Your score will not affect your payment in any way.



Demonstration of the task by the other participant

This page will automatically submit after 10 minutes are over. Do NOT refresh / reload this page.

Here is the summary of what happened in the experiment.

Points Scored: 38
Your Bonus Payout: $1.023

Please note that any bonus payment must be approved before they are given. Your bonus amount (if any) will be paid in 24 hours.

Did you have any questions, concerns or comments about this study? If so, enter them here.:

On the next screen, you will be given a survey code that you must enter into the textbox on Mechanical Turk to get paid.

Thank you for participating in this study.

Your MTurk completion code is: 28377

It is **very important** that you do not share any of your results and that you do not provide any details about this study to other potential participants. We trust in you to keep this study and your results confidential.

References

Abel, M. (2019). *Do Workers Discriminate against Female Bosses?*

Akerlof, G. A. (1982b). Labor Contracts as Partial Gift Exchange. *Quarterly Journal of Economics*, *97*(4), 543–569.

Allport, G. W. (1954). *The Nature of Prejudice*. Addison-Wesley Publishing Company.

Andreoni, J., Payne, A. A., Smith, J., & Karp, D. (2016). Diversity and donations: The effect of religious and ethnic diversity on charitable giving. *Journal of Economic Behavior and Organization*.

Ayalew, S., Manian, S., & Sheth, K. (2018). *Discrimination from Below: Experimental Evidence on Female Leadership in Ethiopia.*

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago Press.

Benson, A., Board, S., & Meyer-ter Vehn, M. (2019). *Discrimination in Hiring: Evidence from Retail Sales.*

Bertrand, M., & Duflo, E. (2017). Field Experiments on Discrimination. In *Handbook of economic field experiments* (Vol. 1, pp. 309–393). North-Holland.

Bohren, J. A., Haggag, K., Imas, A., & Pope, D. (2019). *Inaccurate Statistical Discrimination.*

Breza, E., Kaur, S., & Shamdasani, Y. (2017). The morale effects of pay inequality. *Quarterly Journal of Economics*, *133*(2), 611–663.

Cavaille, C. (2018). *Implementing Blocked Randomization in Online Survey Experiments.*

Chakraborty, P., & Serra, D. (2019). *Gender differences in top leadership roles: Does worker backlash matter?*

Charles, K. K., & Guryan, J. (2008). Prejudice and Wages : An Empirical Assessment of Becker's The Economics of Discrimination. *Journal of Political Economy*, *116*(5), 773–809.

Charles, K. K., & Guryan, J. (2011). Studying discrimination: Fundamental challenges and recent progress. *Annu. Rev. Econ.*, *3*(1), 479–511.

Charness, G., Rigotti, L., & Rustichini, A. (2007). Individual Behavior and Group Membership. *American Economic Review*, *97*(4), 1340–1352.

Chudy, J., Piston, S., & Shipper, J. (2019). Guilt by Association: White Collective Guilt in American Politics. *Journal of Politics*, *81*(3).

Cook, C., Diamond, R., Hall, J., List, J. A., & Oyer, P. (2019). *The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers.*

Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). *The dozen things experimental economists should do (more of).*

DellaVigna, S. (2018). Structural Behavioral Economics. In *Handbook of behavioral economics: Applications and foundations* (Vol. 1, pp. 613–723). North-Holland.

DellaVigna, S., List, J. A., Malmendier, U., & Rao, G. (2016). Estimating social preferences and gift exchange at work. *NBER Working Paper Series*, *53*(9), 1689–1699.

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

de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, *108*(11), 3266–3302.

Doleac, J. L., & Stein, L. C. D. (2013b). The visible hand: Race and online market outcomes. *Economic Journal*, *123*(572), 469–492.

Eckel, C. C., & Petrie, R. (2011). Face value. *American Economic Review*, *101*(4), 1497–1513.

Fairlie, R. W., & Robb, A. M. (2007). Why are black-owned businesses less successful than white-owned businesses? The role of families, inheritances, and business human capital. *Journal of Labor Economics*, *25*(2), 289–323.

Fershtman, C., & Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *Quarterly Journal of Economics*, *116*(1), 351–377.

Glover, D., Pallais, A., & Pariente, W. (2017a). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. *Quarterly Journal of Economics*, 1219–1260.

Gneezy, U., & List, J. A. (2006). Putting behavioral economics to work: resting for gift exchange in labor markets using field experiments. *Econometrica*, *74*(5), 1365–1384.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*.

Grossman, P. J., Eckel, C. C., Komai, M., & Zhan, W. (2019). It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior and Organization*, *161*, 197–215.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*.

Hitlin, P. (2016). *Research in the Crowdsourcing Age: A Case Study* (Tech. Rep.). Pew Research Center.

Katz, L., & Krueger, A. (2019). *Understanding Trends in Alternative Work Arrangements in the United States.*

Kube, S., Maréchal, M. A., & Puppe, C. (2006). Putting reciprocity to work-positive versus negative responses in the field. *University of St. Gallen Economics Discussion Paper*.

Lahey, J. N., & Oxley, D. R. (2018). *Discrimination at the Intersection of Age, Race, and Gender: Evidence from a Lab in the field Experiment.*

Loury, G. C. (1998). Discrimination in the post civil rights era: Beyond market interactions. *Journal of Economic Perspectives*, *12*(2), 117–126.

Mummolo, J., & Peterson, E. (2019). Demand Effects in Survey Experiments: An Empirical Assessment. *American Political Science Review*, *113*(2), 517–529.

Neumark, D. (2018). Experimental Research on Labor. *Journal of Economic Literature*, *56*(3), 799–866.

Oh, S. (2019). *Does Identity Affect Labor Supply?*

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, *5*(5), 411–419.

Phelps, E. S. (1972). The Statistical theory of Racism and Sexism. *American Economic Review*, *62*(4), 659–661.

Rich, J. (2014). *What Do Field Experiments of Discrimination in Markets Tell Us? A Meta Analysis of Studies Conducted since 2000.*

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*.

Simon, H. A. (1993). Altruism and Economics. *American Economic Review: Papers & Proceedings*, *83*(2), 156–161.

Sundstrom, W. A. (1994). The Color Line: Racial Norms and Discrimination in Urban Labor Markets, 1910-1950. *Journal of Economic History*, *54*(2), 382–396.

Triplett, J. (2012). Racial bias and prosocial behavior. *Sociology Compass*, *6*(1), 86–96.

CHAPTER 4. DO NUDGES INDUCE SAFE DRIVING?

Modified from a manuscript to be submitted to the Americal Economic Journal: Applied.

Sher Afghan Asad Kevin D. Duncan
Iowa State University Iowa State University

4.1 Abstract

Behavioral economics has transformed the way we think about policy problems of our age. Governments all over the world are using nudges, one of the tools from behavioral economics, to direct people's behavior towards socially desirable outcomes. However, it is not clear what kind of nudges are most effective, if at all. In this research, we look at the impact of different types of nudges, adopted by various governments, on road safety behavior. Particularly, we look at the traffic-related messages such as "drive sober," "x deaths on roads this year," and "click it or ticket,", displayed on major highways, on reported near-to-sign traffic accidents. To estimate the causal effect of these nudges, we build a new high-frequency panel dataset using the information on the time and location of messages, traffic incidents, overall traffic levels, and weather conditions using the data of the state of Vermont. We estimate several models that control for endogeneity of these messages, allow for spillover effects from neighboring messages, and look at the impact as the function of distance from the sign. We find that behavioral nudges, such as "drive sober" and "wear seat belt", are at best ineffective in reducing the number of crashes while informational nudges, such as "slippery road" and "work zone", actually lead to causal reduction in number of crashes. Our findings are robust to many different specifications and assumptions.

JEL Codes: O18 R41 R42

4.2 Introduction

Thaler and Sunstein (2008) define nudges as "choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives." These benign behavioral interventions have become increasingly popular with the researchers and the governments all over the world to address various policy problems. There have been hundreds of studies which show that nudges are effective in influencing behavior ranging from donating organs (Johnson & Goldstein, 2003), reducing energy consumption (Allcott & Mullainathan, 2010), and saving more money (Thaler & Benartzi, 2004). While there is considerable evidence on the effectiveness of nudges, there is not much known about

what kind of nudges are most effective and if poorly designed nudges can backfire. In this paper, we study the effectiveness of different nudges on road safety behavior and examine whether all nudges are created equal and if not which ones are most effective in ensuring road safety.

Nudges have been used all over the world for decades as a part of road safety management and to give useful information to drivers. Recently governments are increasingly adopting intelligent transportation systems (ITS) to provide real time information to the drivers on the road. As a part of ITS, different state governments in the United States have installed dynamic message signs (DMS) on various highways. These dynamic signs have become a regular medium through which the government provide information, such as updated time to destination or road conditions, or behavioral nudges, such as reminders to wear a seat belt or how many individuals have died on the road this year, to the drivers. The purpose of DMS is to reduce driver anxieties related to commutes, and to encourage safer driving. These signs generally face broad public approval (Benson (1997) Tay and De Barros (2008)), however despite this popularity, whether or not these signs actively encourage safer driving is unknown. Even small changes to driver behavior can have large effects societal welfare through decreasing road injuries and deaths. In 2019, over 37,000 people died in road accidents in the United States, and another 2.35 million individuals were injured or disabled due to road incidents. Since 2010 motor vehicle accidents have ranked 11th overall as a cause of death, and 6th in terms of years of life lost.[1]

In this paper, we identify the causal impact of two types of nudges on near-to-sign reported traffic incidents. In particular we categorize each message on the DMS into two kinds of nudges, behavioral and informational. Behavioral Nudges constitute messages which are aimed at encouraging drivers to drive safe without any concrete information on driving conditions, such as "Buckle Up", "Drive Sober", and "Click It or Ticket." Informational Nudges consist of messages which provide concrete information on driving conditions, such as information on road conditions, weather conditions, upcoming events, and road diversions. The main focus of this paper is to examine whether behavioral nudges and/or informational nudges are effective at reducing traffic incidents.

The distinction between behavioral nudges and information nudges has its roots in the psychology literature, where different message types may trigger a different response from drivers. Behavioral nudges are supposed to encourage individuals to take precautionary measures and drive safer, however poorly thought out nudges can have opposite effects. Nudges can be interpreted as condescending and may invoke negative reaction (Dholakia, 2016). Individuals may feel that they don't like to be told what to do or how to drive and may indulge in over-speeding or more reckless driving. Informational nudges are more direct and are aimed at providing information relating to traffic, weather, or excess road risk. This may cause drivers to

[1] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812203

drive more carefully in light of information, or drive less carefully if the information changes their prior on riskiness in the opposite direction.

We estimate the causal impact of nudges on near to sign accidents on either the mile, or quarter mile, directly after a DMS using a Poisson fixed effects model. The main challenge in estimating the causal effect is to address the joint determination of nudges and crashes. To account for this we further include models with site specific trends, or nest a Difference-in-Difference approach that indexes post DMS treatment effects to be relative to the road area just before a DMS. Finally,, informational nudges (such as about weather and crashes on the road) are selectively displayed when conditions are precarious while behavioral nudges are selected in otherwise less risky conditions. To control for plausible contemporaneous assignment of messages to road conditions, we motivate a model a Poisson model with sequential exogeneity using generalized method of moments. Estimation is carried out using a new high frequency panel data set that covers the population of displayed DMS message signs, including exact display times, and reported traffic accidents from January 2016 to December 2018 in the state of Vermont. This further pools geocoded information on hourly traffic, weather, and temperature data. We map each reported crash to the potential DMS using information on road network incorporating driving distance, driving time, and number of turns between signs and the location of the crash. This allows pairing the location and the timing of the crash and the nudge to evaluate the treatment assignment of each driver before getting into a crash.

Our results show that behavioral nudges have either no to small negative effects on traffic accident rates on the road area just after a DMS. Across a variety of specifications message content of behavioral nudges do not meaningfully alter reported accident rates immediately after the sign. Our mainline specifications imply a decrease in accidents ranging from a 9-40% decrease in near-to-sign accidents thanks to behavioral nudges, and a 35-150% increase in accidents caused by Informational Nudges- though across both specifications some of these effects are indistinguishable from zero. In our preferred specification, Behavior Nudges have a 40% decrease in accidents, and Information Nudges do not impact near-to-sign accidents, which amounts to about 22 fewer traffic accidents a year.Over the three year period, this amounts to roughly 52 fewer property-only accidents, 12.5 fewer injuries, and .66 fewer deaths an additional hours worth of behavioral nudges are displayed every day. We run alternate specifications that decompose results into heterogeneous effects by message sub type, and show that there exists moderate heterogeneity between message types. Many of the behavioral messages remain statistically insignificant. Comparably, results for informational messages are shown to be a mix of a strong, endogenous response caused by "Crash Ahead" messages, but that "Other Caution" messages do provide information to drivers and decrease near-to-sign accidents. Across other informational messages, we find no effects.

This paper adds to the literature in a few meaningful ways. First, we provide new estimates of the impacts of DMS message content on near-to-sign traffic incidents outside of the initial reason for sign roll out. Many previous studies have focused on evaluating a single message type, while often deployed DMSs display a rotating array of messages. Secondly, our estimates are tied directly to when messages are displayed. By pairing over 600 reported traffic incidents in the mile after a DMS and known start and end times of different messages, we are able to more accurately tie treatment status at time of crash to the displayed message at the time. Finally, within the observational setting of our data, we discuss ways to accommodate the endogenous selection of displayed message to traffic and other time varying hazards that might impact accident rates immediately following a message board. Combined, these provide precise estimates of revealed driver behavior in response to the small nudges provided by DMS systems. Due to the proclivity of many states to set up, maintain, and use DMS networks to provide message content to drivers, there is strong interest in understanding the impacts of DMS nudges on driver behavior. Setting up, running, and maintaining these systems is not cost-free, and observational work on the effects of DMS networks often comes to conclusions that displayed messages do not improve driver safety. Hall and Madsen (2020) estimate the impacts of Death Toll reminders in Texas on near-to-sign traffic incidents. Using exogenous roll out of Death Toll reminders, they find strong increases in the number of traffic incidents even 10 kilometers after a DMS. Their identification strategy relies on the relative increase in Death Toll reminders as a share of total displayed messages, versus causally identified with times when Death Toll reminders are actually up. Secondly, the lack of effectiveness of behavioral nudges is consistent with at-least one previous study that indicate that message signs might cause additional collisions. Song, Wang, Cheung, and Keceli (2016) and Erke, Sagberg, and Hagman (2007) show reading messages on DMSs may lead to a slowing down and speeding up effect among drivers, potentially making roads more dangerous around the signs. Norouzi, Haghani, Hamedi, and Ghoseiri (2013) show no treatment effect using both on/off analysis, and comparing downstream traffic incidence to near-to-sign traffic incidents. Fallah Zavareh, Mamdoohi, and Nordfjærn (2017) examine how people respond to DMSs with road risk ratings. Risky behavioral adaptations were observed under low and medium risk messages during night time. The effects of high risk messages were consistently related to safety adaptations. The effects of messaging on rear-end collisions were significant only in the fast lane at night time. Overall, observational studies of the impacts of DMSs on near-to-sign traffic incidents often indicate that these signs have detritus effects on driver behavior and lead to an increase in such incidents.

The broader literature on the effectiveness of nudges in the transportation literature generally uses a mixture of simulation and stated preference surveys, or observational data comparing before and after using simulated traffic and accident data (see Mounce et al. (2007)). Compared to the observational work discussed above, simulation and stated preference surveys often find quite strong and positive evidence on message

boards on driver behavior (Benson (1997); Bonsall (1992); H. M. Hassan, Abdel-Aty, Choi, and Algadhi (2012); Peng, Guequierre, and Blakeman (2004); Tarry and Graham (1995); W. Xu, Zhao, Chen, Bian, and Li (2018)). Recent work by Choudhary, Shunko, Netessine, and Koo (2019) combines experimental design and revealed observational driving changes from interventions. They randomly given driving quality feedback messages on driver's smartphone, showing that personalized nudges generally improve driving performance compared to the control group.

Behavioral and informational responses on roads have been studied in many other contexts. Changing incentives for risky driving is common through examining how budget shortfalls and resulting decreases in police staffing impact safe driving (Makowsky and Stratmann (2011),DeAngelo and Hansen (2014)), that reduction in accidents following texting bans are short-lived Abouk and Adams (2013), and that scaling DUI punishments associated with how far over the legal limit a driver registers impact recidivism and future driving behavior Hansen (2015). Understanding and improving driver responsiveness to DMSs may lead to moderate reductions in traffic accidents, injuries, and fatalities. Alternatively, towns may be overstating belief in DMS value to drivers. De Borger and Proost (2013) show that the city government over-invest in externality reducing infrastructure whenever this infrastructure increases the generalized cost of through traffic. We can therefore expect an excessive number of speed bumps and traffic lights, but the right investment in noise barriers. In turn, we would expect higher rates of DMSs to exist along roads than socially optimal, and understanding these effects might help governments and public policy groups set more socially optimal levels of message signage. While in our study most of the DMSs are fixed, new hazards alternatively might cause additional road incidents, in this setting M. Xu and Xu (2020) evaluates how the introduction of new fracking wells is associated with near-to-well fatal car accidents.

The remainder of the paper proceeds as follows. Section 4.3 provides background information on the DMS system in Vermont and details on data and their sources. Section 4.4 describes the empirical strategy and series of models to be estimated. Section 4.5 provides the estimation results of our models. Section 4.6 provides various robustness check and finally Section 4.7 concludes.

4.3 Data

This section provides detailed information on how Dynamic Message Signs (DMSs) location and message content as decided upon by the Vermont Agency of Transportation (VTrans). This is collected from a series of primary documents and direct communication with the agency. We further describe our accident, traffic, and weather data, and how these data are combined. Finally, we provide basic descriptions of the final variables that we use in our estimation procedures.

4.3.1 Dynamic Message Sign Location

The installation of DMSs are a part of VTrans' effort under the Intelligent Transportation System (ITS) to facilitate drivers with updated and timely information on traffic and road conditions.[2] VTrans initially deployed these boards with portable installations with the aim to eventually phase in permanent installations. The message boards covered in this study are all portable installations - called portable variable message signs (PVMS).[3] The signs are typically mounted on trailers or pads, often with the wheels removed and secured in place for longer duration of use. Typically, PVMS run on solar power or battery. The PVMS have the ability for an adjustable display rate, which is typically set to allow for the message to be read at least twice at the posted speed limit.

The location of the message board is determined based on multiple factors including frequency of crashes and weather related incidents on a road segment. The detailed plan of location choice is provided in Vanasse Hangen Brustlin (2007). Broadly, the general location is determined by identifying areas where it warrants weather notification to the drivers of hazardous conditions, advanced notification of substandard roadway conditions and upcoming "chain up" areas can be provided to the truck drivers, notification of construction and planned events can be provided to avoid congestion on relevant roads, or notifications can complement counties' transportation management plans involving traffic and roadside safety. According to officials at VTrans, "*our goal was just to place them (DMSs) in high traffic areas and close to RWIS (Road Weather Information System). The placement of RWIS was based on high crash areas.... Going forward the goal was decided to place DMS before on/off ramps on interstates and close to major intersections on secondary highways.*" This suggests that these message boards are installed in the areas which are more susceptible to crashes.

The specific location of the message board is determined considering horizontal and vertical alignment of the message board. Typically, PVMS is visible from approximately 0.5 miles (or 2,500 feet) under both day and night conditions. The message is legible from a minimum distance of an 1/8th of a mile (or 650 feet). When possible, the PVMS signs are placed behind guardrail sections or outside the clear zone for errant vehicles. PVMS are mounted in such a way that the bottom of the message sign panel is minimum of seven feet above the roadway. Once the location of the DMS is determined, the next issue is about the content of messages that needs to be displayed on a particular DMS.

[2] In particular, the DMSs are primarily aimed at providing information on i) road conditions, ii) adverse weather notifications, iii) incident management, iv) in-route emergency evacuation information, iv) national missing and exploited children alert system - amber alerts, v) special events, vi) flight, train, and bus schedules in transportation terminals, vii) congestion management, viii) construction information/detours, ix) road closures, and x) special messages (such as variable speed limits, etc).

[3] Throughout our sample, the location of individual PVMS are fixed.

4.3.2 Message Data

Based on the conversations with officials at VTrans, the choice of message is determined based on risk factors such as road and weather conditions. For example, if the road conditions are more susceptible to accidents because of icy roads then drivers will be cautioned about the slippery conditions of the road. Behavioral nudges (such as nudging individuals to drive sober or notifying traffic death counts) are considered low priority messages and are only displayed when there is no other important information that needs to be conveyed. Some of the nudge messages such as "Click It and Ticket", "Drive Sober or Get Pulled Over", etc are based on national campaigns run by National Highway and Traffic Safety Administration (NHTSA). NHTSA run regular campaigns countrywide to raise awareness on drunk driving, seat belts etc. Just like other nudge messages, the campaign messages are displayed if the message boards are not being used for more important informational messages such as construction, crashes, winter weather, etc.

The data on messages is obtained from VTrans from June 2016 to December 2018. Messages were displayed on 67 unique sites during this time period. Table 4.1 presents the number and duration of various messages during the time period. During this period, there were total of 10,409 messages spanning 308,800 hours. Figure 4.3.1 shows the durations during which the message boards were active. It's clear that there is considerable heterogeneity in the activity and duration of messages across these message boards.

As shown in Table 4.1 and Figure 4.3.1 we categorize each message into different groups. "Death toll reminder" provide information about the number of people who have died that year from traffic accidents. "Seat belt reminders," "Texting reminders," "Speed reminders," and "Drinking reminders" aim to encourage seat belt use, no texting, staying under the speed limit, and sober driving, respectively. "Road condition message" displays information about the road characteristics such as gravel road. "Weather message" display the current weather conditions, and are most frequent in winter season due to snow and icy roads along with precipitation levels. "Traffic message" display information about traffic congestion or delays. Comparably, "Work zone message," "Road closure message," and "Crash message" inform about the upcoming work areas, upcoming road closures, or if there is a crash ahead. Finally, "Other caution message" is any other message which does not fall into the above category such as "Drive Safe" or "Better late than sorry". Some messages can have overlapping content between these types, for example, "40 Lives Lost in 2017 Buckle Up" is categorized in both "Death related message" and "Seat belt related message."

For our analysis we further categorize each message type into two kinds of nudges, behavioral and informational. Behavioral Nudges are nudges which are aimed at encouraging drivers to drive safe without any concrete information on driving conditions. Death, seat belt, texting, drinking, and speed are categorized as Behavior Nudges. Information Nudges are nudges which provide concrete information on driving conditions

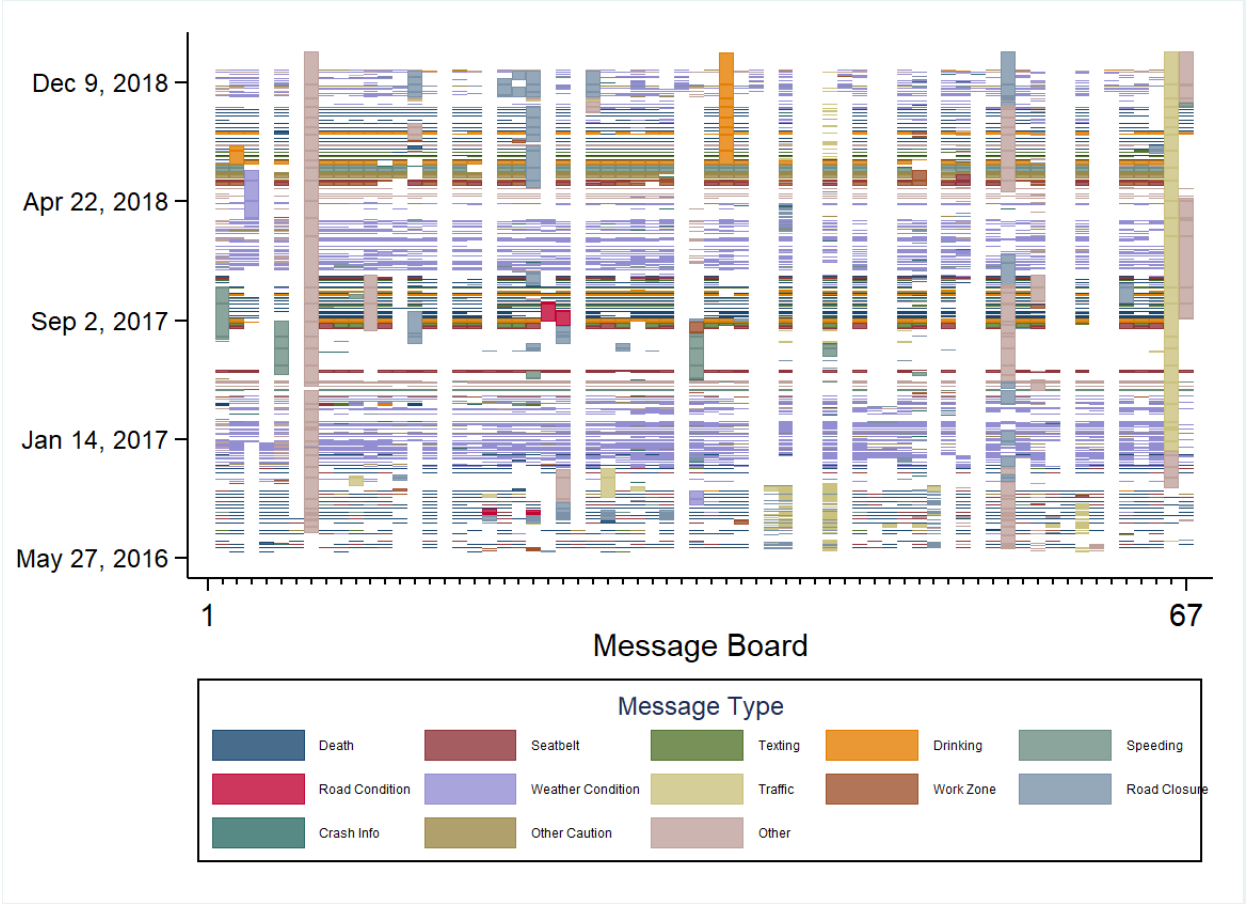


Figure 4.3.1: Message Boards Activity
Notes: The figure shows the time periods during which any message was displayed on the message board. Each bar represents a message board, with white areas indicating no message during the time period.

to the drivers. Road condition, weather, traffic, work zone, road closure, crash ahead, other caution, and other messages are categorized as Information Nudges, which broadly categorizes each of the message types into either reminders, or informational messages.

4.3.3 Crash Data

Data on crashes between the periods January 2016 and December 2018 is also obtained from VTrans. This data set reports wide set of details from the police reports about the crash including location, time and date, road conditions, weather conditions, driver details and condition, vehicle details, number and nature of injuries, and number of passengers involved. There were total of 35,554 police reports, that involved 64,027 vehicles, of crashes during this time period in the state of Vermont.

Since the data on crashes come from police reports in a busy field setting of a crash site, the spatial location of each crash may not always be accurate. For our purposes, the exact location of a crash is crucial

Table 4.1: Summary of Messages

	(1) Proportion of Number of Messages	(2) Proportion of Duration of Messages (in hours)
Death related message	0.19	0.11
Seatbelt related message	0.03	0.07
Phone related message	0.02	0.04
Drinking related message	0.02	0.09
Speed related message	0.16	0.14
Road condition message	0.05	0.03
Weather message	0.48	0.17
Traffic message	0.06	0.08
Work zone message	0.02	0.03
Road closure message	0.05	0.09
Crash message	0.01	0.00
Other caution message	0.03	0.04
Other message	0.09	0.25
Total	10409.00	308799.85

Notes: The table presents the number and duration of various messages during the time period from June 2016 to December 2018 in Vermont.

to be able to map the crash to a potential message that may have been seen by the driver before getting into the crash. Here we describe the measures that we have taken to validate the spatial location of each crash.

First, VTrans has taken steps to geocode precise crash location for the recent data in their efforts to improve the quality of data for traffic safety and analysis purposes.[4] We use this data to get precise spatial location of most (35,202) crash sites during the said time period. Second, we are able to update geographic location of 5,999 police reports using spatial location of overlapping subset of crash data provided on VTrans Public Query Tool.[5] Third, few of the geographic coordinates in our data are using State Plane Coordinate System (rather than GPS coordinate system), we convert those to GPS coordinates and are able to update spatial location of 31 crash sites. Forth, there are cases for which crash location is provided in the text fields but with missing coordinates. We use ArcGIS® to geocode these addresses and are able to update spatial location of 140 crash sites.

To check for the validity of the coordinates from the above sources we reverse geocode the GPS coordinates using ArcGIS® and find that coordinates of 94 crash sites are either not street addresses or fall outside the county (within Vermont) in which they are supposed to lie (as determined on the basis of county of crash site). We then, once again, geocode the addresses for which either GPS coordinate is missing, not a street address, or found to lie outside the respective county and are able to find locations of 42 crash sites. The above measures leave us with missing or incorrect location for 347 crash sites which are manually looked at

[4]The data set is available at https://geodata.vermont.gov/datasets/
[5]This data set can be extracted from http://apps.vtrans.vermont.gov/CrashPublicQueryTool/

Table 4.2: Contributing Circumstances to the Crash

	(1) Proportion
No improper driving	35.62
Inattention	12.47
Other improper action	11.06
Driving too fast for conditions	9.42
Failed to yield right of way	8.23
Failure to keep in proper lane	7.77
Other	5.97
Followed too closely	5.74
Under the influence of medication/drugs/alcohol	1.86
Visibility obstructed	1.52
Other Activity- Electronic Device	0.33
Distracted	0.02
Total	100.00

Notes: This table presents the share of each contributing factor in a crash as recorded by the VTrans.

using information on various address fields.[6] Given the information, we remain unable to manually locate 82 crash sites, i.e., overall, we are able to locate 35,472 crashes (99.9 percent) with reasonable degree of accuracy.

Out of the located crash sites, 8 police reports lack information on the date and time of crash and therefore we drop them. Our final data set of crashes has 35,472 crash reports majority of which (approximately 79%) involved "property damage" only while the rest of them constitute injuries (approximately 20%) and fatalities (approximately 1%). The geographical location of each crash along with message board location in visually presented in Figure 4.3.2. As is typical of the collision data, the crashes are clustered around each other. The value of the nearest neighbor index is 0.11 ($z = -426.66$) which represents high degree of clustering of crashes around each other (Clark & Evans, 1954).

The contributing circumstances for the crashes as recorded by the VTrans are presented in Table 4.2. Factors such as fast driving, failure to yield, failure to keep in proper lane, following too closely, and inattention are some of the major factors contributing to the crashes.

4.3.4 Combining Message and Crash Data

The purpose of this study is to assess the impact of a particular nudge on the probability of a crash. To analyze that, we restrict our analysis to the regions where a message board is installed for at-least some

[6]We use information on street address and distance from intersecting street to manually locate the crash location on the Google Maps. In case of missing information about the street address the nearest intersecting street information is used to approximate the location of the crash. We use Google Map's measurement feature to measure the offset from the intersection based on the information provided, for example, 100 feet south of 1st St. and 1st Ave. We also use the measurement feature to locate addressed based on mile markers, such as, I-89 South, Mile Marker 65.3. We find base mile markers by using a map of Vermont's interstate exits and rest areas which is then located on Google Maps to get a reference mile marker and a measure of specified distance to the target mile marker.

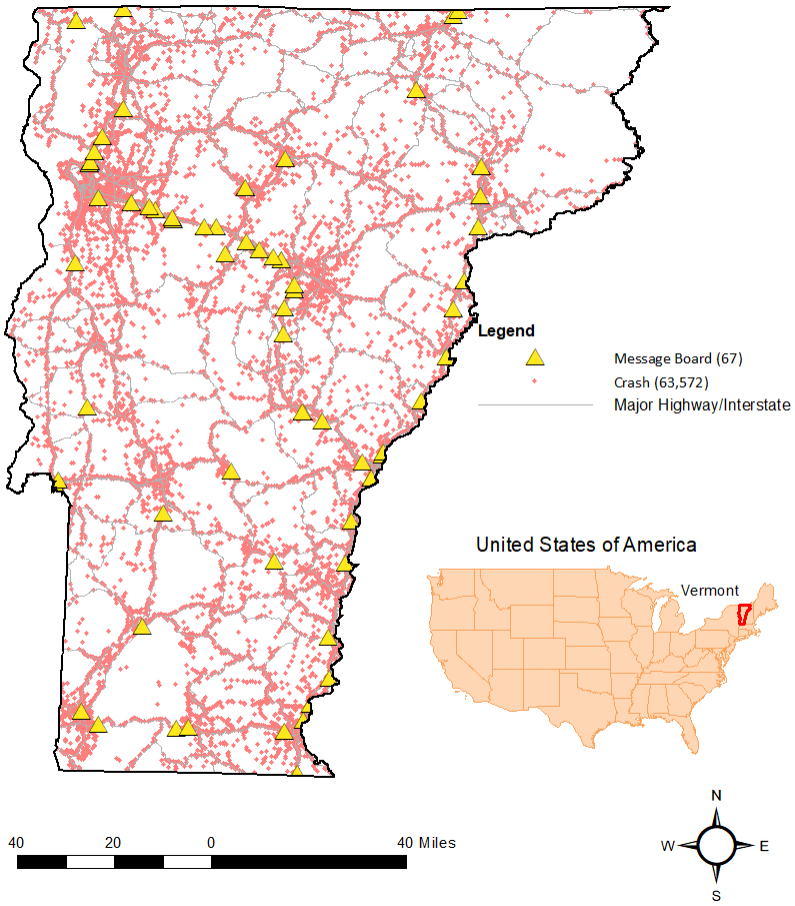


Figure 4.3.2: Map of Message Boards and Crashes

Notes: This map of Vermont represents crashes and message boards throughout the period between June 2016 and December 2018.

duration during the study time period. This implies that we use 67 locations/message-boards around which we focus our analysis. We map crashes to these 67 locations using ArcGIS® 'Find Closest Facility' tool. This tool finds one or more crashes that are closest from a message board based on travel distance (and travel time), and outputs the driving directions between the message board and the crash. When finding closest crashes, we specify to find closest crashes within a 10 mile distance to or from a message board and then restrict to crashes which are maximum of two turns away from the message board. We restrict to a maximum of two turns to be reasonably confident of a driver having read the message before getting into the crash. We also adjust for the message read time by adjusting the time of crash by the travel time from the message board to the location of the crash. We also assign the status of "pre" or "post" to each crash to determine whether the crashed occurred before the mapped message board or after. We use driving directions along with direction of travel of a vehicle to determine the pre/post status. Approximately 1,700 traffic accidents are mapped across the mile bandwidth directly before, and after all of the message boards in the sample. The mile just prior to all of our DMSs experiences 1,034 accidents, while the mile after has 627 reported traffic accidents over the sample time period.

The average number of crashed vehicles by different message types is presented in Figure 4.3.3. Of note, a clear case of endogeneity can be seen using crash info message which is displayed when there is a crash ahead Secondly, across the remaining message types, there remains large heterogeneity in the number of crashes naively correlated when those messages are displayed. Death, drinking, and speeding reminders all seem to have about equivalent average crashes/hour associated with them. Of the Behavior Nudges, Texting and Seat belt Reminders have about half of the average crashes. Similar stratification exists among Information Nudges. Multiple, Other, Other Caution, Road Closure, and Work Zone messages similarly have equivalently naive road hazard associated with them. Displayed Weather messages have the highest, likely related to being displayed when weather related hazards on the road are the most perilous. Comparably, Road Condition and Traffic messages have very low perceived correlation between being displayed, and number of accidents at the time.

Generally these signs go up endogenously in response to other perceived, or lack of, road hazards. Therefore, comparison of means is not a sufficient measure to judge their impacts on near-to-sign crashes, and more robust models and estimators need to be employed.

4.3.5 Traffic and Weather Data

Traffic on a particular road is considered to be one of the crucial factors that can effect the probability of a crash. The traffic data is obtained from the VTrans which has installed traffic counters on various

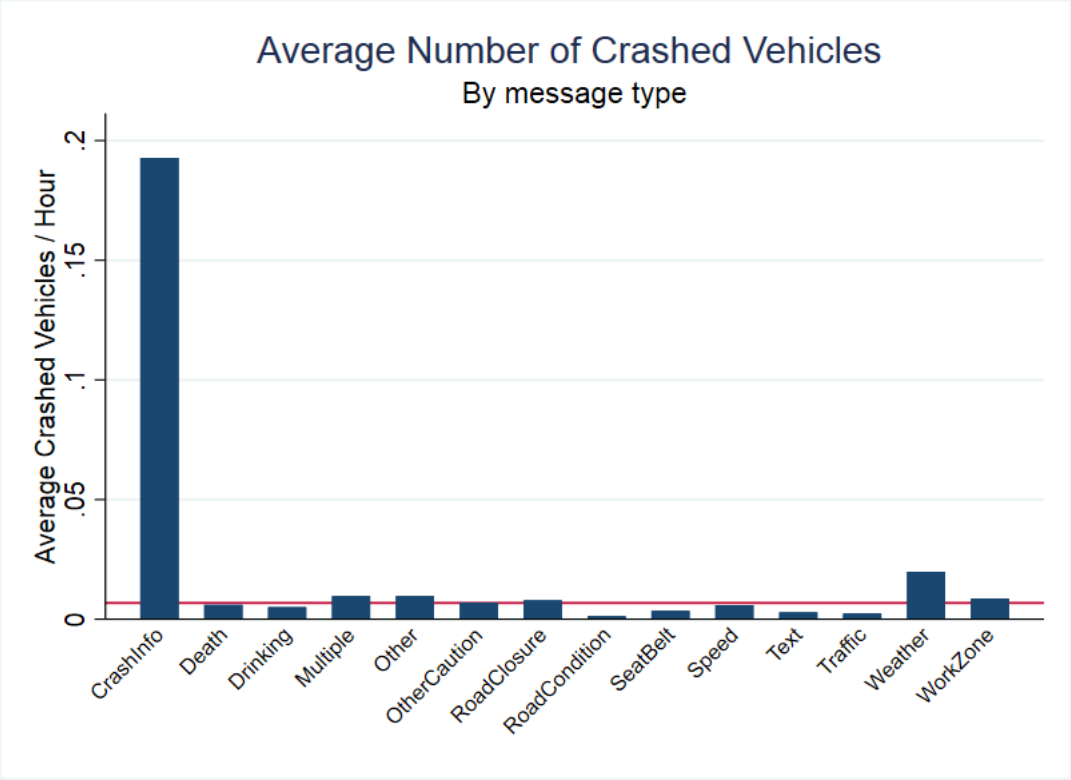


Figure 4.3.3: Average Number of Crashed Vehicles by Message Type

Notes: This figure presents the average number of crashed vehicles by message type within 10 miles post the message sign. Red line indicate the average crashed vehicles per hour when there was no message displayed at the time the driver may have passed the location of the message board.

highways in the state of Vermont. This data covers hourly road volume counts across 86 sites in Vermont over the entire duration of our traffic and message board data. The traffic volume on an average day follows the typical seasonality with traffic peaking during rush hours and returning to low volume during off-peak hours.

We map the traffic information to the message boards by once again using ArcGIS® 'Find Closest Facility' tool. In most instances, the closest traffic monitoring station is found on the same road as the message board, and when there is no traffic monitoring station on the road of the message board, we use the nearest traffic monitoring station to map traffic information to the message board. We also account for the direction bound of the road in mapping the stations to the message boards. This gives a local approximation to local traffic trends, and is generally a good approximation as both volume counters and DMSs tend to be placed on busier roads.

We further obtain hourly weather data from National Oceanic and Atmospheric Administration's National Centers for Environmental Information's Local Climatological Data, which provides daily and hourly summaries for approximately 11 Vermont locations, including Automated Surface Observing System and Automated Weather Observing System stations. They provide us daily data on snowfall and snow depth, as well as hourly data on dew point temperature, precipitation, wind condition, sky condition (cloudy, overcast, etc), weather condition (snowing, raining, drizzle, hail etc), and visibility.[7]

The exact definition of variables used is provided in Table 4.3. As a final note, our sample does not include road construction data which might be a relevant piece of information that influences crashes around the DMS.Correspondence with VTrans concluded that obtaining this data would be costly, and empirical design is ideally robust to this missing information due to the overall fixed location of signs during our period of observation, and overall methods described in Section 4.4.

4.4 Empirical Model

The question we are answering is as follows: is the change in the number of crashes happening because of nudges, or would the crashes be lower (higher) anyway at the time these nudges are displayed, perhaps because these are displayed during times when conditions are relatively safe (unsafe) for driving. To answer this question, we model the number of crashes as a count process following a Poisson distribution. We take this modeling approach for several reasons. First, individual probabilities of accidents on a given road segment at a given time are Bernoulli trials, so the probability of observing a certain number of crashes on a given location follows a Binomial distribution, for which as traffic volume gets large, converges to

[7] Link to data.

Table 4.3: List of Variables

Variable	Source	Description
Dependent Variable (Y_{it}^r)		
Crashes	VTrans	Number of crashes r miles from site i at time t
Time Varying Covariates (X_{it})		
Traffic	VTrans	Hourly traffic volume from 24 traffic counters in Vermont
Dew Point Temperature	NOAA	Hourly dew point temperature in Fahrenheit.
Precipitation	NOAA	Hourly amount of precipitation in inches to hundredths.
Humidity	NOAA	Hourly relative humidity given to the nearest whole percentage
Visibility	NOAA	Hourly horizontal distance an object can be seen and identified given in whole miles.
Sky Conditions	NOAA	Hourly report of cloud layer with options clear, partly cloudy, and mostly cloudy.
Wind Speed	NOAA	Hourly speed of the wind at the time of observation given in miles per hour (mph).
Snow Depth	NOAA	Daily amount snow depth in inches.
Snowfall	NOAA	Daily amount of snowfall in inches
Message Data (T_{it})		
Behavioral Nudge	VTrans	Dummy variable which takes a value 1 if either of the death, seat belt, texting, drinking, speeding, or other caution was active for at-least some time during the hour t on site i.
Informational Nudge	VTrans	Dummy variable which takes a value 1 if either of the road condition, weather condition, traffic, work zone, road closure, crash info, or other message was active for at-least some time during the hour t on site i.

the Poisson distribution. Secondly, as shown in Chamberlain (1987); Hausman, Hall, and Griliches (1984); J. M. Wooldridge (1999) location specific fixed effects fall out of the Poisson distribution. Generally though in our setting this is not a problem, since we have roughly 62 different message boards, but almost 27,000 observations per site. To get around an incidental parameters problem with respect to the time dimension, our fixed effects follow a Year-by-Month structure. Specifically, we adopt the following baseline specification

$$E[Y_{it}|\alpha_i, \lambda_t, \boldsymbol{X_i}, \boldsymbol{T_i}] = exp(\alpha_i + \lambda_t + X'_{it}\beta + T'_{it}\rho) \tag{4.4.1}$$

where Y_{it} equal the number of crashes on road segment i at time t, X_{it} is a vector of traffic and weather conditions, T_{it} is a vector of Behavior and Information Nudge treatment status on road segment i at time t,and α_i is a vector of unobserved but fixed confounders that influence near to sign road hazard. For instance, road segments with specific features (e.g. rough road, curved road, junctions, merging roads) are more probable to have one nudge or the other. λ_t represents time fixed effects to control for year and seasonal effects that impact road hazard.. Similar to the canonical Within-Transform of the linear additive fixed effects model, this transformation removes both individual effects, as well as other time-invariant factors from Equation 4.4.1 (J. M. Wooldridge (2010) Section 18.7.4), such that much of the remaining variation in accidents is coming from time varying covariates such as traffic, weather, and message board status.

Endogeneity arises when $E[\epsilon_{it}|\alpha_i, \lambda_t, \boldsymbol{X_i}, \boldsymbol{T_i}] \neq 0$. This may arise from three sources. First, via location choice of message boards. As the message boards are strategically installed on high risk roads, it is likely that the number of crashes are higher on these sites as compared to sites without message boards. Second, when the unobserved effects that may influence crashes on the site are varying time (i.e. α_i is actually time varying). In that case, ρ will capture the wrong effect when T_{it} is correlated with changes in these unobserved factors. Finally, if the nudges are jointly determined with the crashes. This is likely to occur given that the messages are not displayed randomly and therefore there is a selection bias from message choice. Most importantly, Informational Nudges include displayed warnings of crashes ahead, which respond to accidents either contemporaneously or that just happened. Ignoring these potential sources of endogeneity may lead to inconsistent estimates in the above specification.

To deal with endogeneity from the location choice of message boards, we restrict the sample to only sites with message boards. This is possible because we know the precise location of each crash and the message board. The assumption here is that sites with message boards during the study period are quite similar to each other. To control for unobserved time varying factors we follow Angrist and Pischke (2009) and

introduce site-specific time trends to the list of controls in 4.4.1, i.e., we estimate

$$E[Y_{it}|\alpha_i, \lambda_t, \boldsymbol{X_i}, \boldsymbol{T_i}] = exp(\alpha_o + \alpha_i t + \lambda_t + X'_{it}\beta + T'_{it}\rho) \tag{4.4.2}$$

where α_i is the site-specific trend coefficient multiplying the time trend variable, t. This allows treatment and control sites to follow different trends. Our interest is in examining whether the estimated effect ρ changes by the inclusion of these site specific trends.

A remaining concern is that messages are displayed when even conditional on traffic and weather data there might already be additional average hazards on the road on the entire segment around the sign. Implicitly this accommodates variants of our mainline specification where α_i is changing, or, perhaps switching states, or accommodating unobserved heterogeneity in when different message types go as decided by VTrans. Under this setting the entire road segment, both immediately before and after the sign, might face excess, or lower, probabilities of individual drivers getting into accidents. To address this concern we estimate the following fixed effects specification:

$$E[Y_{itr}|\alpha_{ir}, \lambda_t, \boldsymbol{X_i}, \boldsymbol{T_i}] = \exp(\alpha_{ir} + \lambda_t + X_{it}\beta_r + T'_{it}\rho + 1\{r = 1\}T'_{it}\rho_1) \tag{4.4.3}$$

Here r indexes relative distance to a DMS. The accidents for both $r = -1$ (the mile before a DMS) and $r = 1$ (the mile after a DMS) are combined together, allowing for estimation of level shifts in mean road hazards when messages of a given time are displayed. α_{ir} is now a fixed effect for each tranche relative to a DMS, such that $\alpha_{i1} = \alpha_i$ as in Equation 4.4.1, but allowing the mile immediately before a sign to have it's own fixed effects, and allowing identification of whether or not there is excess road hazard locally around a DMS when signs of a particular type are up. For this model, the key parameters of interest is ρ_1. Similar to traditional Differences-in-Differences, ρ is the pre-treatment effect when a sign is active in the mile before a sign. This tries to capture level differences in near-to-sign hazards that exist on average when messages of a given type are displayed. Indexing in this fashion implies the causal interpretation that, ρ_1 is the impact of a particular message type on the mile wide bin after a sign while controlling for excess hazard in the full region.[8]

As with Equation 4.4.2, we can use this stacked structure to estimate a model with time varying location specific factors by creating an ID variable that indexes observations by DMS ID and time, and has a creates a

[8]This can be thought of as a pseudo Regression Discontinuity Design. Since we do not observe actual traffic data at the DMS, and traffic accidents are very rare, canonical Regression Discontinuity Design estimation strategies are unavailable to us. This approach enables two different ways of trying to proxy for the at-sign hazard rate. Allowing for different hazard rates on either side of the DMS through relative distance level fixed effects controls time invariant factors that might be correlated with the initial sign placement. The variable ρ reflects the mean hazard in the mile immediately before and after a DMS under each of the different signs.

secondary index based on relative distance to the DMS. This model has multiple desirable features. First, it removes all covariates that are invariant over the two distances, including our traffic and weather variables, which due to matching might have meaningful measurement error. Secondly, it allows for temporary or other time varying changes to near-to-sign road hazard, such as temporary construction work, or seasonal variation in mean road hazard that might not just be captured by weather variables that do not fit into the linear trend presented in Equation (4.4.2).

$$E[Y_{itr}|\alpha_{it}, \lambda_r, \boldsymbol{X_i}, \boldsymbol{T_i}] = \exp(\alpha_{it} + \lambda_r + T'_{it}\rho + 1\{r = 1\}T'_{it}\rho_1) \tag{4.4.4}$$

The above models is robust to initial placement of DMSs, time varying location specific effects, and unobserved risk factors on the entire road segment around the DMS, they do not account for concerns about endogenous response of messages, in particular informational nudges, to crashes happening after a DMS. The fixed effect estimation require a core assumption of strict exogeneity:

$$E[\epsilon_{is}|\alpha_i, \lambda_t, X_{it}, T_{it}] = 0 \ \forall \ s, t \tag{4.4.5}$$

This assumption forbids current value of ϵ_{is} to be correlated with past, present, and future values of T_{it}. Comparably, in the presence of reverse causality this assumption is necessarily violated, i.e., if crashes in current time period (Y_{it}) effect choice of nudge in the next time period (T_{it+1}), then ϵ_{it} is correlated with T_{it+1}. This violation leads to biased estimates using the fixed effect estimation. We relax 4.4.5 to instead accommodate the data generating process,

$$E[Y_{it}|\alpha_i, X_{i1}, ..., X_{it}, T_{i1}, ..., T_{it}, \lambda_1, ..., \lambda_t] = exp(\alpha_i + \lambda_t + X'_{it}\beta + T'_{it}\rho) \tag{4.4.6}$$

Under this setting future displayed messages can be correlated with past levels of realized crashes. To accommodate for this we follow Chamberlain (1992); Windmeijer (2000); J. M. Wooldridge (1997) to estimate fixed effect Poisson models that exhibits sequential exogeneity. We take quasi first-difference to eliminate the fixed effects by using the following transformation,[9]

$$\Delta Y_{it} = \frac{\mu_{i,t-1}}{\mu_{it}} Y_{it} - Y_{i,t-1}$$

[9]The difference appears as under 4.4.6 $Y_{it} = exp(\lambda_t + X'_{it}\beta + T'_{it}\rho)u_{it} = exp(\lambda_t + X'_{it}\beta + T'_{it}\rho)\phi_i\epsilon_{it}$. From this we get $\phi_i\epsilon_{it} = \frac{Y_{it}}{\exp(\lambda_t + X'_{it}\beta + T'_{it}\rho)} = \frac{Y_{it}}{\mu_{it}}$.

where $\mu_{it} = exp(\lambda_t + X'_{it}\beta + T'_{it}\rho)$. Iterated expectations shows that $E[X_{it-s}\Delta Y_{it}|X_{i1}, ..., X_{it}, T_{i1}, ..., T_{it}, \lambda_1, ..., \lambda_t] = 0$ for all $s \geq 2$ Windmeijer (2000),[10] allowing for estimation using generalized method of moments.[11] This model offers protection both unobserved time-invariant heterogeneity but also from reverse causality, but it does not offer protection against time varying heterogeneity of discussed in models 4.4.2 and 4.4.3. Similar methods are further discussed in Allison (2012); Colin and Pravin (2013).

4.5 Main Results

In this section we present our main results. First, we present the baseline estimates on the effect of nudges within one mile from the legibility of the message boards. We then put these estimates through various specifications (as outlined in Section 4.4) to address the endogeneity of nudges and crashes. Finally, we examine the effect of these nudges as the distance from the message board location decreases.Throughout we report Incident Rate Ratios. These provide a standardize measure of understanding the relative rates of accidents occurring given a 1 unit (1 hour) increase in a given message type, which is equivalent to reporting $\exp(\beta)$ for a given effect of interest. Comparably, the percentage increase in accidents can be calculated as $(\exp(\beta) - 1)x100$ is the percent change in crashes caused by a one hour increase in a given message type's exposure time. In turn, we report the incidence rate ratio, along with the mean number of accidents for the a given tranche side following a DMS. The benchmark values for the number of accidents are 0.0004 accidents happen on average in a given hour, and 0.0000387 crashes happen on average in the quarter mile after a sign.

To get an idea of how rare crashes of any type are, across our 35,000 accidents, we have roughly 67 message boards and 26,208 hourly observations per board. Across the entire state, the total number of accidents is about 1.35 per hour. Thus, even across the whole state of Vermont, accidents are rare, let alone in just the mile long tranche immediately following a DMS. As a result, even a 100% increase in the rate of accidents implies the mean number of accidents happening in the mile after a DMS rises to only 0.0008 accidents per hour, or about 7 accidents per a year. Similarly, a 100% increase in the mean number of accidents happening in the mile after a DMS rises to 0.0000774 accidents per hour, or about 0.678 accidents per year. The multiplicative effect of the incidence rate ratio makes the actual impacts on number of crashes dependent on the tranche size, and likely that otherwise "large" effects will exist.

Table 4.4 presents the set of estimates that captures the effect of nudges within one mile from the installation of message boards. In the baseline specification (column 1), the parameter estimate for behavioral

[10] A major concern is that due to the pooling information at the hour level, X_{it} dependent on ϵ_{it}, to get around this we use a twice lagged set of covariates as instruments.

[11] This moment condition is equivalent to $E[X_{t-s}\mu_{it-s}\Delta u_{it}] = 0 \; \forall \; s \geq 2$.

nudge is statistically zero while information nudge is positive and significant. Controlling for site-specific time trends (column 2) do not change the parameter estimates much. In both cases though the implied increase in crashes caused by the presence of informational nudges is about 145%. After conditioning on road hazard correlated with the presence of signs in the mile before a DMS when a given message type goes up, implies either only a 30% increase in near-to-sign accidents, or a result that is statistically indistinguishable from zero.

Similarly, when looking at just the Post Mile results, Behavior Nudges have no impact on near-to-sign road accidents. After conditioning on pre-trends, Behavioral Nudges causally decrease near-to-sign traffic incidents by about a 40%. These results indicate that both Behavior and Information Nudge messages go up when there is already excess risk on the whole region around the DMS relative to time periods without a displayed message.

Table 4.4: Effect of nudges on crashes within 1 mile from message board

	Baseline	Site Trends	IDxDist	IDxTime	Sequential Exog
	(1)	(2)	(3)	(4)	(5)
BehaviorNudge	0.826	0.910	0.585**	0.611*	0.277
	(0.191)	(0.218)	(0.106)	(0.133)	(2.839)
InformationNudge	2.316***	2.451***	1.509	1.349*	1.906
	(0.498)	(0.580)	(0.385)	(0.203)	(74.48)
Weather Controls	Yes	Yes	Yes	No	Yes
Traffic Controls	Yes	Yes	Yes	No	Yes
Site-Specific Trends	No	Yes	No	No	No
Observations	1473017	1473017	2919730	2386	1472891

Notes: The table presents the estimates for the effect of nudges on crashes within one mile from the legibility of message board. Column 1 presents the result of baseline fixed-effect Poisson regression. Column 2 control for the site-specific time trends in the baseline specification. Column 3 presents an event study style estimator indexed to the mile before a DMS and an indexing variable of Message Board ID and relative distance. Column 4 presents an event study style estimator indexed to the mile before a DMS with an indexing variable of Message Board ID and Date. Column 5 presents results of a Poisson fixed effects regression under sequential exogeneity. Clustered robust standard errors in parenthesis. ** for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$*

A remaining concern is that the mile-long bandwidths before and after a DMS capture too much road surface area not attributable to a particular sign. To test this hypothesis we estimate our preferred specifications using data on accidents that occurred in just the quarter mile before, and the quarter mile after, a DMS. Ideally this better captures near to sign determinants of sign content, as well as a stronger share of drivers who actively saw the sign's message. Throughout the measures of informational nudges increase by 390-580% increase in accidents, but as before, results that condition on road hazard immediately around a DMS removes those effects. Compared to above, across models the coefficients related to behavioral nudges remain statistically insignificant.

Table 4.5: Effect of nudges on crashes within 1/4 mile from message board

	Baseline (1)	Site Trends (2)	IDxDist (3)	IDxTime (4)
BehaviorNudge	0.791	2.058	0.461	0.447
	(0.634)	(1.617)	(0.361)	(0.370)
InformationNudge	4.941***	6.818***	0.767	1.640
	(1.796)	(2.417)	(0.388)	(0.807)
Weather Controls	Yes	Yes	No	Yes
Traffic Controls	Yes	Yes	No	Yes
Site-Specific Trends	No	Yes	No	No
Observations	447166	447166	516	1525627

Notes: The table presents the estimates for the effect of nudges on crashes within a quarter mile from the legibility of message board. Column 1 presents the result of baseline fixed-effect Poisson regression. Column 2 control for the site-specific time trends in the baseline specification. Column 3 presents an event study style estimator indexed to the quarter mile before a DMS and an indexing variable of Message Board ID and relative distance. Column 4 presents an event study style estimator indexed to the quarter mile before a DMS with an indexing variable of Message Board ID and Date. Clustered robust standard errors in parenthesis. ** for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$*

Both these models share fundamentally similar results. Across models, behavioral nudges have no impact on near-to-sign accidents. Poisson regressions on just the mile or quarter-mile tranche after a DMS, shows strong effects of informational nudges increase in the relative rate of accidents. By controlling for road hazard associated with signs of a given type being displayed, or controlling for endogeneity in displayed message, reduce these effects to zero.

4.6 Robustness Checks

In this section we provide two sets of robustness checks over our mainline specification in Section 4.5. .Each robustness check estimates variants of Equations 4.4.1, 4.4.2, 4.4.3, and 4.4.4. Throughout they focus on dis-aggregated estimates of DMS message content, breaking behavioral nudges into 5 categories- death toll, drinking and driving, seat belt, speeding, and texting reminders- and breaking informational nudges into eight different reminders- road, weather, traffic, work zone, road closure, crash info, other caution, or other message conditions. The first test re-estimates our mainline set of models with these dis-aggregated displayed message content. The second robustness check further tries to control for whether or not spatial spillovers might be biasing results. This concern comes from two points of view. The first is an implicit test for whether or not DMSs might be creating long-run changes in driving behavior. If spatial spillovers appear in the our results, the induced better, or worse, driving behavior might continue throughout the DMS network. Secondly, many DMSs are placed relatively close to each other, and results from Section 4.5 might

be confounded by the presence of upstream messages. This check for spatial spillovers is provided through estimating two different models. The first includes whether or not there was an upstream behavioral or informational nudge within 5 miles, and the second conditions on signs that are at least 5 miles away from any upstream neighbor.

4.6.1 Heterogeneous Message Type Effects

As noted above, we estimate impacts of heterogeneous message content. Using detailed message content provided by VTrans, we split the messages into 13 different categories. The aim here is to understand where plausible sources of endogeneity are coming from, for example Crash Ahead should always be contemporaneously correlated with an accident occurring ahead at some interval. Secondly, drivers might be responding to different message types in a heterogeneous fashion. Seat belt reminders might not elicit a response for safer driving since individuals are already often buckled up, while death toll reminders might elicit momentary feelings of remorse and changes to safer driving behavior. The model that controls for plausible sequential exogeneity of regressors is omitted, since the likelihood is too flat with respect to our parameters of interest.

Effects for the mile wide tranche are discussed in Table 4.6. All of the dis-aggregated measures of behavioral message content are again statistically insignificant from zero, but many indicate a decrease in the number of near-to-sign accidents by about 30-50%. One concern is that the point estimates across the two class of models vary wildly, to the point where one much choose which class of specification they believe in. Comparably there are large heterogeneous effects across informational content provided by DMSs. As expected of our indexing approach, road condition and weather messages disappear from being significant due to these being reflected in excess road hazard both before and after the sign, and can only be displayed due to fixed nature of the signs in the sample. Moreover, Crash Ahead messages lead to huge increases in the probability of a crash, but are also clearly endogenous to the displaying of such messages- a reason why we estimated a model with sequential exogeneity in our mainline specifications.

Changing to the quarter mile bandwidth paints and entirely different story. Under this specification there are many signs that now have strong, statistically significant, negative impacts on near-to-sign accidents. Drinking, Speed, Work zone, and Other Caution Messages share sign, and often magnitudes, across specifications. These effects range from a 30-100% decrease in the number of near to sign crashes. Texting, Traffic, and Crash Info messages do not have crashes associated with them on this interval, so are dropped from the IDxTime fixed effects model.

Table 4.6: Effect of Heterogeneous Message Types on crashes within 1 mile from message board

	Baseline (1)	Site Trends (2)	IDxDist (3)	IDxTime (4)
Death Toll	1.253	1.325	0.718	0.693
	(0.383)	(0.419)	(0.327)	(0.316)
Seatbelt Reminder	0.762	0.775	2.259	2.440
	(0.423)	(0.435)	(1.859)	(1.705)
Texting Reminder	0.434	0.436	1.995	2.181
	(0.292)	(0.298)	(1.903)	(1.903)
Anti Drinking Reminder	1.021	1.118	0.440	0.452
	(0.354)	(0.393)	(0.194)	(0.243)
Speeding Reminder	0.918	0.996	0.574	0.615
	(0.292)	(0.326)	(0.211)	(0.215)
Road Condition	0.444*	0.430*	1.140	1.006
	(0.145)	(0.142)	(0.705)	(0.518)
Weather Condition	3.211***	3.317***	1.302	1.281
	(0.907)	(1.004)	(0.427)	(0.270)
Traffic Condition	1.513	1.461	3.949	5.907
	(1.707)	(1.599)	(5.977)	(8.302)
Work Zone	0.982	0.972	0.429	0.356
	(0.575)	(0.580)	(0.333)	(0.282)
Road Closure	1.131	0.996	1.714	1.475
	(0.364)	(0.454)	(0.603)	(0.599)
Crash Info	24.47***	24.70***	1.2950e+157***	7.81365e+18***
	(19.65)	(20.24)	(5.4940e+158)	(2.58189e+19)
Other Caution	0.177	0.209	0.0571**	0.0493**
	(0.188)	(0.222)	(0.0620)	(0.0542)
Other Message	1.357	1.375	1.503	1.176
	(0.736)	(0.845)	(0.888)	(0.317)
Weather Controls	Yes	Yes	Yes	No
Traffic Controls	Yes	Yes	Yes	No
Site-Specific Trends	No	Yes	No	No
Observations	1473017	1473017	2919730	2386

Notes: The table presents the estimates for the effect of nudges on crashes within one mile from the legibility of message board. Column 1 presents the result of baseline fixed-effect Poisson regression. Column 2 control for the site-specific time trends in the baseline specification. Column 3 presents an event study style estimator indexed to the mile before a DMS and an indexing variable of Message Board ID and relative distance. Column 4 presents an event study style estimator indexed to the mile before a DMS with an indexing variable of Message Board ID and Date. ** for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$*

Table 4.7: Effect of Heterogeneous Message Types on crashes within 1/4 mile from message board

	Baseline (1)	Site Trends (2)	IDxDist (3)	IDxTime (4)
Death Toll	4.741	5.617*	2.109	1.888
	(4.018)	(4.873)	(1.862)	(1.989)
Seatbelt Reminder	1.489	2.185	1.433	0.913
	(1.601)	(2.517)	(1.870)	(0.835)
Texting Reminder	6.43e-113***	6.31e-100***	1.51e-15	1
	(3.56e-111)	(2.08e-98)	(1.29e-13)	(.)
Anti Drinking Reminder	4.71e-54***	5.23e-50***	7.66e-62***	0.000000114***
	(3.89e-53)	(5.45e-49)	(7.07e-61)	(8.56e-08)
Speeding Reminder	8.47e-195***	3.14e-193***	7.63e-233***	1.90e-09***
	(3.79e-193)	(1.37e-191)	(3.22e-231)	(3.52e-09)
Road Condition	0.133	0.158	0.117	0.107*
	(0.166)	(0.191)	(0.170)	(0.122)
Weather Condition	3.280	4.662*	0.766	0.459
	(2.068)	(3.110)	(0.589)	(0.348)
Traffic Condition	6.54e-195***	4.93e-166***	4.97536e+38	1
	(4.11e-193)	(2.69e-164)	(3.55679e+40)	(.)
Work Zone	7.21e-36***	7.20e-24***	5.59e-40***	0.000000320***
	(6.95e-35)	(3.17e-23)	(5.04e-39)	(0.000000326)
Road Closure	9.607***	6.033***	10.09***	1.978
	(2.799)	(1.514)	(3.603)	(2.265)
Crash Info	2.53e-77***	8.31e-90***	4.7396e+128***	1
	(3.78e-76)	(1.43e-88)	(3.1588e+130)	(.)
Other Caution	1.91e-186***	3.68e-175***	4.59e-220***	8.62e-17***
	(6.11e-185)	(1.29e-173)	(1.34e-218)	(2.60e-16)
Other Message	13.82***	22.03***	18.53***	0.900
	(9.436)	(18.74)	(13.20)	(0.732)
Weather Controls	Yes	Yes	Yes	No
Traffic Controls	Yes	Yes	Yes	No
Site-Specific Trends	No	Yes	No	No
Observations	447166	447166	1525627	516

Notes: The table presents the estimates for the effect of nudges on crashes within a quarter mile from the legibility of message board. Column 1 presents the result of baseline fixed-effect Poisson regression. Column 2 control for the site-specific time trends in the baseline specification. Column 3 presents an event study style estimator indexed to the quarter mile before a DMS and an indexing variable of Message Board ID and relative distance. Column 4 presents an event study style estimator indexed to the quarter mile before a DMS with an indexing variable of Message Board ID and Date. Clustered robust standard errors in parenthesis. ** for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$*

4.6.2 Spillover from neighboring signs

A remaining concern is that upstream signs might have impacts on downstream driving behavior. Our final set of robustness checks explicitly models this in two ways, first we include an indicator for a behavioral or informational nudge from upstream signs, and secondly we subset our sample to include signs that have at least 5 miles between them and any upstream neighbor. This is important since in the presence of non-zero treatment effects from upstream signs, drivers might be already driving differently than they would have in the presence of no prior treatment assignment.

To model this effect we condition on upstream neighbor sign status within miles of each subject sign. To do so, define variables

$$SpillBehavior_{it} = 1\{\text{Upstream sign within 5 miles of i has an Behavioral Nudge message up}\}$$

$$SpillInformation_{it} = 1\{\text{Upstream sign within 5 miles of i has an Informative Nudge message up}\}$$

In the case where there is no spatial dependence, the coefficients on these terms converges to zero and this becomes a regular Poisson fixed effects model presented in Equation 4.4.1, while if they're non-zero incorporates downstream sign information. The coefficients for sign effects are almost identical to Table 4.4, indicating that while in some models these spillover effects are meaningful, they are seemingly independent of drivers response to new message content even a few miles downstream. While for the Poisson fixed effects estimator on just the mile after has positive effect, after accounting for road hazard in the entire region, these upstream messages have no impact on downstream driving.

Alternatively, we create a sub sample of signs with no upstream neighbor within 5 miles. Under local effects this implies that plausible spillover effects of neighboring signs is zero, and might alternatively remove impacts of upstream message content on downstream signs if upstream response is also heterogeneous as shown in Section 4.6.1. The downside to this approach is that it further restricts the sample to rural or otherwise isolated areas, and away from urban and high traffic areas. Even after carrying out this sub setting, the story remains identical.

4.7 Conclusion

In this paper we study the impact of behavioral and informational nudges on near-to-sign traffic incidents. We generate a large geospatial panel data set using hour level information on weather, traffic, crashes, and

Table 4.8: Effect of Heterogeneous Message Types on crashes within 1 mile from message board with message spillovers

	Baseline (1)	Site Trends (2)	IDxDist (3)	IDxTime (4)
Death Toll	1.215	1.285	1.513**	0.667
	(0.366)	(0.392)	(0.198)	(0.307)
Seatbelt Reminder	0.771	0.785	0.700	2.484
	(0.426)	(0.439)	(0.307)	(1.730)
Texting Reminder	0.462	0.463	2.253	2.182
	(0.314)	(0.318)	(1.866)	(1.936)
Anti Drinking Reminder	0.965	1.034	1.992	0.428
	(0.348)	(0.378)	(1.952)	(0.235)
Speeding Reminder	1.012	1.088	0.424	0.605
	(0.362)	(0.392)	(0.199)	(0.222)
Road Condition	0.451**	0.433**	0.592	1.048
	(0.137)	(0.133)	(0.250)	(0.550)
Weather Condition	2.433**	2.526**	1.129	1.210
	(0.796)	(0.856)	(0.701)	(0.267)
Traffic Condition	1.392	1.361	1.212	5.605
	(1.551)	(1.470)	(0.447)	(7.755)
Work Zone	0.976	0.946	3.975	0.361
	(0.580)	(0.578)	(6.045)	(0.287)
Road Closure	1.054	0.971	0.429	1.424
	(0.347)	(0.416)	(0.339)	(0.579)
Crash Info	22.03***	22.10***	1.684	7.10121e+18***
	(18.68)	(19.20)	(0.610)	(2.34758e+19)
Other Caution	0.131	0.151	1.4289e+156***	0.0449**
	(0.143)	(0.165)	(5.9744e+157)	(0.0495)
Other Message	1.132	1.171	0.0522**	1.112
	(0.564)	(0.658)	(0.0577)	(0.305)
SpillBehavior	1.002	0.977	1.408	1.066
	(0.164)	(0.158)	(0.800)	(0.214)
SpillInformation	1.706**	1.705**	1.002	1.122
	(0.340)	(0.341)	(0.199)	(0.157)
Weather Controls	Yes	Yes	Yes	No
Traffic Controls	Yes	Yes	Yes	No
Site-Specific Trends	No	Yes	No	No
Observations	1473017	1473017	2919730	2386

Notes: This table presents the effect of nudges on crashes within one mile of a message board. Column 1 presents the result of baseline fixed-effect Poisson regression. Column 2 control for the site-specific time trends in the baseline specification. Column 3 and 4 presents an event study style estimator indexed to the mile before a DMS and an indexing variable of Message Board ID and relative distance, or Message Board ID and Date respectively. ** for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$*

Table 4.9: Effect of Heterogeneous Message Types on crashes within 1 mile from message board with no upstream neighbor in 5 miles

	Baseline	Site Trends	IDxDist	IDxTime
	(1)	(2)	(3)	(4)
Death Toll	1.726	1.845	0.930	0.958
	(0.708)	(0.773)	(0.621)	(0.576)
Seatbelt Reminder	0.214	0.231	0.396	0.399
	(0.208)	(0.225)	(0.464)	(0.285)
Texting Reminder	0.756	0.762	1.567	1.604
	(0.498)	(0.516)	(1.505)	(1.399)
Anti Drinking Reminder	1.530	1.661	0.464	0.444
	(0.389)	(0.457)	(0.246)	(0.305)
Speeding Reminder	0.722	0.801	0.421	0.457
	(0.254)	(0.282)	(0.187)	(0.232)
Road Condition	0.517*	0.500*	2.067	1.827
	(0.165)	(0.157)	(1.432)	(1.114)
Weather Condition	3.356***	3.594***	1.325	1.354
	(0.982)	(1.074)	(0.427)	(0.376)
Traffic Condition	2.31e-80***	7.85e-68***	4.63535e+09	1
	(1.15e-78)	(1.44e-66)	(3.91606e+11)	(.)
Work Zone	0.762	0.740	0.268	0.210
	(0.577)	(0.584)	(0.273)	(0.225)
Road Closure	0.759	0.892	2.301	2.231
	(0.419)	(0.516)	(2.132)	(1.490)
Crash Info	1.36e-207***	3.95e-171***	4.64e-17	1
	(8.97e-206)	(2.54e-169)	(4.16e-15)	(.)
Other Caution	0.195	0.236	0.226	0.169
	(0.212)	(0.256)	(0.251)	(0.219)
Other Message	1.043	1.201	0.777	1.044
	(1.171)	(1.694)	(0.932)	(0.360)
Weather Controls	Yes	Yes	Yes	No
Traffic Controls	Yes	Yes	Yes	No
Site-Specific Trends	No	Yes	No	No
Observations	868025	868025	1709746	1228

Notes: This table presents the effect of nudges on crashes within one mile of a message board. Column 1 presents the result of baseline fixed-effect Poisson regression. Column 2 control for the site-specific time trends in the baseline specification. Column 3 and 4 presents an event study style estimator indexed to the mile before a DMS and an indexing variable of Message Board ID and relative distance, or Message Board ID and Date respectively. ** for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$*

the content of messages displayed on dynamic message signs. We estimate several variations of Poisson fixed effects models, including a baseline model on just the (quarter) mile after a DMS, allowing for cite specific trends, two difference-in-differences model with Message Board ID by Time or Message Board ID by Distance fixed effects, and a model with sequential exogeneity to identify the effect of the nudges.

Our results show that without conditioning on near-to-sign excess road hazard that is correlated with variant message signs going up, or the causal relationship between different message types and previous periods or contemporaneous accidents, can lead to spuriously believe that informational nudges have a causal relationship on near-to-sign accidents.After accounting for these relationships, there exists minor evidence that behavioral nudges might improve driving behavior, while informational nudges have no impact on near-to-sign accidents. After dis-aggregated results there exists large heterogeneity in driver response to individual signs. Many of our behavioral nudges have economically meaningful, but statistically insignificant, negative impacts on near-to-sign traffic incidents, while reductions in accidents caused by Other Caution Messages are meaningfully swamped out by the existence of endogeneity between Crash Ahead and contemporaneous traffic behavior. Models that try to explicitly control for spillover effects further cast doubt on claims about short lived duration of positive or negative effects of DMSs on near-to-sign traffic incidents.

From the policy perspective, our results indicate that behavioral nudges are not an effective way to reduce the number of traffic incidents. Drivers seem to ignore these signs messages, or not change their driving behavior in response to them. On the other hand, informational nudges, being direct useful information, may causally reduce the number of crashes, but continued issues with Crash Ahead and messages muddle this response in many models. Therefore, we suggest the use of informational nudges while avoiding behavioral nudges. More research is needed to examine which exact kind of behavioral nudges are effective and which may lead to negative reaction from the drivers.

The effects that we study in the paper are local to the sign location. We acknowledge that these nudges might be causing long run changes in overall driving behavior which are not captured in the immediate vicinity of the message board location. It is possible that these nudges have time varying effects, for example, people may react to these nudges when they are first displayed but over time drivers become immune to them and start ignoring them, or worse start to get annoyed by them, thereby petering out the effect. How rare crashes are in our matched message sign and near to sign crashes make such models hard to estimate. Instead we restrict ourselves to models where drivers are effectively naive, and do not experience long run changes in driving behavior caused by displayed message treatment.

References

Abouk, R., & Adams, S. (2013). Texting bans and fatal accidents on roadways: Do they work? Or do drivers just react to announcements of bans? *American Economic Journal: Applied Economics*, *5*(2), 179–199. doi: 10.1257/app.5.2.179

Allcott, H., & Mullainathan, S. (2010). Behavior and energy policy. *Science*, *327*(5970), 1204–1205.

Allison, P. (2012). *Fixed Effects Regression Models*. SAGE Publications. doi: 10.4135/9781412993869

Angrist, J. D., & Pischke, J.-S. (2009). Parallel Worlds: Fixed Effects, Differences-in-Differences, and Panel Data. In *Mostly harmless econometrics* (pp. 221–247).

Benson, B. G. (1997, jan). Motorist attitudes about content of variable-message signs. *Transportation Research Record*, *1550*(1550), 48–57. doi: 10.1177/0361198196155000107

Bonsall, P. (1992, feb). The influence of route guidance advice on route choice in urban networks. *Transportation*, *19*(1), 1–23. doi: 10.1007/BF01130771

Chamberlain, G. (1987, mar). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*(3), 305–334. doi: 10.1016/0304-4076(87)90015-7

Chamberlain, G. (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business and Economic Statistics*, *10*(1), 20–26. doi: 10.1080/07350015.1992.10509881

Choudhary, V., Shunko, M., Netessine, S., & Koo, S. (2019). *Nudging Drivers to Safety: Evidence from a Field Experiment.* Retrieved from `https://www.ssrn.com/abstract=3491302` doi: 10.2139/ssrn.3491302

Clark, P. J., & Evans, F. C. (1954). Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology*, *35*(4), 445–453. doi: 10.2307/1931034

Colin, C. A., & Pravin, T. (2013). *Regression analysis of count data, Second edition.* Cambridge University Press. doi: 10.1017/CBO9781139013567

De Borger, B., & Proost, S. (2013, jul). Traffic externalities in cities: The economics of speed bumps, low emission zones and city bypasses. *Journal of Urban Economics*, *76*(1), 53–70. doi: 10.1016/j.jue.2013.02.004

DeAngelo, G., & Hansen, B. (2014). Life and death in the fast lane: Police enforcement and roadway safety. *American Economic Journal: Economic Policy*, *6*(2), 231–257.

Dholakia, U. M. (2016). Why Nudging Your Customers Can Backfire. *Harvard Business Review*.

Erke, A., Sagberg, F., & Hagman, R. (2007, nov). Effects of route guidance variable message signs (VMS) on driver behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(6), 447–457.

Fallah Zavareh, M., Mamdoohi, A. R., & Nordfjærn, T. (2017, apr). The effects of indicating rear-end collision risk via variable message signs on traffic behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour*, *46*, 524–536. doi: 10.1016/j.trf.2016.09.019

Hall, J. D., & Madsen, J. (2020, jun). *Can behavioral interventions be too salient? Evidence from traffic safety messages.* Retrieved from `https://papers.ssrn.com/abstract=3633014`

Hansen, B. (2015). Punishment and deterrence: Evidence from drunk driving. *American Economic Review*, *105*(4), 1581–1617. doi: 10.1257/aer.20130189

Hassan, H. M., Abdel-Aty, M. A., Choi, K., & Algadhi, S. A. (2012). Driver behavior and preferences for changeable message signs and variable speed limits in reduced visibility conditions. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *16*(3), 132–146. doi: 10 .1080/15472450.2012.691842

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric Models for Count Data with an Application to the Patents-R \& D Relationship. *Econometrica*, *52*(4), 909. doi: 10.2307/1911191

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *5649*(302), 1338–1339.

Makowsky, M. D., & Stratmann, T. (2011, nov). More tickets, fewer accidents: How cash-strapped towns make for safer roads. *Journal of Law and Economics*, *54*(4), 863–888. doi: 10.1086/659260

Mounce, J. M., Ullman, G. L., Pesti, G., Pezoldt, V., Institute, T. T., of Transportation, T. D., & Administration, F. H. (2007). Guidelines for the Evaluation of Dynamic Message Sign Performance. , *7*(2), 252p.

Norouzi, A., Haghani, A., Hamedi, M., & Ghoseiri, K. (2013). *Impact of Dynamic Message Signs on occurrence of road accidents.*

Peng, Z. R., Guequierre, N., & Blakeman, J. C. (2004, jan). Motorist response to arterial variable message signs. *Transportation Research Record*, *1899*(1899), 55–63. Retrieved from `http://journals.sagepub.com/doi/10.3141/1899-07` doi: 10.3141/1899-07

Song, M., Wang, J.-H., Cheung, S., & Keceli, M. (2016). Assessing and Mitigating the Impacts of Dynamic Message Signs on Highway Traffic. *International Journal for Traffic and Transport Engineering*, *6*(1), 1–12. doi: 10.7708/ijtte.2016.6(1).01

Tarry, S., & Graham, A. (1995). The role of evaluation in ATT development. IV: Evaluation of ATT systems. *Traffic Engineering and Control*, *36*(12), 688–693.

Tay, R., & De Barros, A. G. (2008, jan). Public perceptions of the use of dynamic message signs. *Journal of Advanced Transportation*, *42*(1), 95–110. Retrieved from `http://doi.wiley.com/10.1002/atr.5670420107` doi: 10.1002/atr.5670420107

Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy*, *112*(1), S164–S187. doi: 10.1086/380085

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, And Happiness.*

Vanasse Hangen Brustlin, I. (2007). *Dynamic Message Sign Study* (Tech. Rep.). Vermont Agency of Transportation.

Windmeijer, F. (2000, jul). Moment conditions for fixed effects count data models with endogenous regressors. *Economics Letters*, *68*(1), 21–24. doi: 10.1016/s0165-1765(00)00228-7

Wooldridge, J. M. (1997, oct). Multiplicative Panel Data Models Without the Strict Exogeneity Assumption. *Econometric Theory*, *13*(5), 667–678. doi: 10.1017/s0266466600006125

Wooldridge, J. M. (1999, may). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, *90*(1), 77–97. doi: 10.1016/S0304-4076(98)00033-5

Wooldridge, J. M. (2010). *The Econometrics of Cross-section and Panel Data* (2nd ed.). Cambridge: MIT Press.

Xu, M., & Xu, Y. (2020, may). Fraccidents: The impact of fracking on road traffic deaths. *Journal of Environmental Economics and Management*, *101*, 102303. doi: 10.1016/j.jeem.2020.102303

Xu, W., Zhao, X., Chen, Y., Bian, Y., & Li, H. (2018). Research on the Relationship between Dynamic Message Sign Control Strategies and Driving Safety in Freeway Work Zones. *Journal of Advanced Transportation*, *2018*, 1–19. doi: 10.1155/2018/9593084

CHAPTER 5. GENERAL CONCLUSION

In this dissertation I have outlined three essay that use insights from behavioral economics to better understand labor market discrimination and road-safety behavior. I have learned that it is important to understand the worker side of the market to better understand labor market discrimination and come up with effective affirmative action policies. From the road-safety work I learnt that nudges are not always effective and some nudges might work better than others.

Going forward, I intend to work on a project which tests the theory of worker-to-employer discrimination in the field setting of the Indian labor market in the context of religious discrimination (Hindus and Muslims). In the Indian setting, where social traditions exercise much more influence on economic behavior, this channel of discrimination may be even more salient.

Building on from my work on nudges, I am also interested in understanding the welfare consequences of nudges. For example, when a person drives through the highway and sees the information nudge of "x deaths this year," this kind of nudge may evoke negative emotions and hence may be welfare reducing (sometimes less information is welfare improving).