Essays in Behavioral Labor Economics

by

Andrea Kiss

Department of Economics
Duke University

Date: _____

Approved:

_____
Robert J. Garlick, Advisor

_____
Scott A. Huettel

_____
Seth G. Sanders

_____
Matthew Masten

_____
Vincent Joseph Hotz

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Economics
in the Graduate School of
Duke University

2021

ABSTRACT

Essays in Behavioral Labor Economics

by

Andrea Kiss

Department of Economics
Duke University

Date: _____

Approved:

_____
Robert J. Garlick, Advisor

_____
Scott A. Huettel

_____
Seth G. Sanders

_____
Matthew Masten

_____
Vincent Joseph Hotz

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Economics
in the Graduate School of
Duke University

2021

Abstract

This dissertation includes three chapters in behavioral labor economics. In the first chapter I study with Victoria Lee the effect of wage transparency policies on workplace interactions and workplace choice. We run an online experiment in which we manipulate whether wages are known or secret, and also control the wage allocations that participants faced in the teams. The results show that wage transparency can affect the level of hostility as well as the target of the hostility in the teams – as measured by a punishment game. These treatment effects are moderated by the level of wage inequality within the teams. We also find that wage transparency can alter which job offers participants accept. This result suggests that wage transparency policies can have general equilibrium effects due to workers changed sorting behavior. We argue that these two sets of results can be explained by the presence of social preferences and participants' inaccurate beliefs about the wage allocations that are corrected when wages are transparent.

In the second chapter Nayoung Rim, Roman Rivera, Bocar Ba and myself are studying racial bias within the police. Although there is substantial evidence showing racial bias in firms' hiring decisions, less is known about bias in career recognition. We construct a novel dataset of police award nominations to measure bias against minority employees. Exploiting quasi-random variation in supervisor assignment and randomized timing of annual evaluations, we find that white supervisors are less likely to nominate black officers than white and Hispanic officers leading up to and during the evaluation period. Further, the black-white nomination gap widens with the number of arrests. These patterns suggest that the disparity is not due to in-group favoritism towards white officers but rather bias against black officers. We conduct an online experiment to examine evaluator engagement and find that evaluators are less likely to engage with black officers vs. white officers. Our findings suggest bias

in career recognition may have important implications for the black-white promotion gap, the lack of diversity in upper-management positions, and, ultimately, the racial wage gap.

In the last chapter I estimate the effects of behavioral interventions on sleep and cognitive performance among undergraduate students. Three interventions are tested during a three-week long framed field experiment: bedtime text reminders, sleep hygiene tips and an educational meditation video. Throughout the experiment participants' sleep patterns are measured using fitness trackers. The results show that sleep quantity and quality may not be easy to move with mild nudges – although the estimates are noisy due to the small sample size. Estimates with panel data methods support the positive link between sleep and cognitive outcomes.

Acknowledgements

First and foremost I am thankful to my academic family who supported my journey at Duke. My dissertation committee – Robert Garlick, Scott Huettel, Seth Sanders, Matthew Masten and Joseph Hotz – made sure I do not get lost in the details for too long. This document reflects their tireless efforts in supporting my far-fetched research ideas. Thank you for your time and advice. Special thanks goes to Robert Garlick, who taught me how to be a better economist with great patience and guided me through the challenges of a pandemic. Thank you.

I am thankful for Scott Huettel for inviting me to his lab, the Huettel laboratory. I learned more than I could have imagined in that space and found a community that never ran out of jokes. Abby, Alex, Dianna, Ed, Kelsey, Khoi, Libby, Matthew, Mel, Nikki, Rosa, Sarah and Vicki, thank you for welcoming me with an open mind and heart.

Several chapters of this dissertation benefited from the work of my co-authors. Victoria Lee was pivotal at the conception of the first chapter. Bocar Ba, Nayoung Rim and Roman Rivera trusted me to design an experiment to amend their already excellent manuscript. Thank you all for working with me.

The financial support for the experiments was provided by Seth Sanders, Scott Huettel, the Interdisciplinary Behavioral Research Center, The Tobin Project, the US Naval Academy and the Center for Advanced Hindsight. I am grateful for their commitment to finance my research.

I also owe thanks to friends and family who provided listening ears and a great deal of empathy during the years. Drew, Jonathan, Rahul and Vivi; I thoroughly enjoyed your company and will miss our times together. Polina, Timi and Fruzsi thank you for keeping me grounded and checking-in often. Finally, I am indebted to my family the most. They supported me unconditionally to achieve my outlandish

goals, though deep inside they wished I would stay closer to them. I am deeply sorry for the missed graduations, birthdays and anniversaries.

Contents

List of Figures

List of Tables

Introduction

Behavioral labor economics uses behavioral economics tools to study labor topics. It deviates from traditional labor economics in terms of its assumptions; it allows for non-standard preferences, non-standard beliefs and non-standard decision making (DellaVigna, 2009). Its premise is that with a more realistic model for individual decision making, it is possible to understand the labor markets deeper, which would lead to more accurate policy recommendations.

Behavioral labor economics has grown over the past decades and is expected to contribute to labor economics in the future (Dohmen, 2014). One of its important contributions so far is the diffusion of experiments in labor economics. Both field experiments and laboratory experiments have become common contributing to labor theories and empirical work.[1]

This dissertation has three chapters in behavioral labor economics. Each chapter relies on a specific deviation from standard preferences, such as time-inconsistent preferences, homophily or social preferences. In addition, the chapters are tied together by experimentation. Two of the three chapters include online experiments, while one of them is about a framed field experiment.

The first chapter, co-authored with Victoria Lee, studies the effect of the wage transparency policies. Wage transparency policies are popular policies to reduce the wage gap in the labor market and are currently in the public debate. With an online experiment, we show that wage transparency can have unintended consequences; it can increase workplace hostility and can change how work-seekers sort to workplaces depending on the wage allocations that are revealed due to the transparency. These results are consistent with social preferences and inaccurate beliefs.

[1]See List and Rasul (2011) for a review about field experiments and Charness and Kuhn (2011) for laboratory experiments in labor economics.

The second chapter – co-authored with Nayoung Rim, Roman Rivera and Bocar Ba – studies whether there is racial bias among the police. We analyze observational data from the Chicago Police Department (CPD) to show that black police officers are less likely to be nominated for an award if they have a white supervisor. In search for an explanation, we ran an online experiment that revealed that while the racial disparities are partially due to homophilious interactions, part of the racial gap remains unexplained.

The third chapter is on sleep and cognitive performance. I ran a framed field experiment to test if various light touch interventions can affect undergraduate students' sleep patterns, and cognitive outcomes. Participants were randomized into a control group or one of the three intervention groups: one group received bedtime reminders, another received tips about how to improve their sleep, and the final group viewed a video about a breathing technique. The intervention experiment was not successful in changing participants' sleep patterns, as measured by fitness trackers. Yet, I find suggestive evidence for the positive relationship between sleep and cognitive performance.

Overall, this dissertation demonstrates that the tools of behavioral economics are useful for studying the labor market.

Chapter 1

When I make less than you: Experimental analysis of punishment and sorting in hierarchical groups (with Victoria Lee)

1.1 Introduction[1]

Wage transparency has been viewed as a tool to eliminate discrimination in wage setting practices. In the United States wages are already transparent in the public sector, but are typically not in the private sector[2]. To increase wage transparency in the private sector and hence limit wage discrimination, the US House of Representatives passed the Paycheck Fairness Act in 2019 (Paycheck Fairness Act, 2019).

The allure of wage transparency is that it would expose previously hidden, illegal wage discrimination cases[3]. As a byproduct, the policy may also increase the bargaining power of (minority) workers and allow them to make more informed decisions. Cullen and Perez-Truglia (2018a) show that in line with this argument, although workers know their own salary, they have inaccurate beliefs about their peers' wages when wages are secret. These inaccurate beliefs are not due to ignorance – workers

[1]This chapter is co-authored with Victoria Lee (Duke University, Psychology and Neuroscience Department). Lee assisted in the design and the data collection plan. Kiss was responsible for securing funds, developing the experimental instruments, obtaining ethical board approval, piloting, developing the pre-analysis plan, carrying out the data collection and analyzing the data. The project was pre-registered on the AEA's RCT Trial Registry (AEARCTR-0006252) and received IRB approval from Duke University (2021-0024). The experiment was funded by Duke University and The Tobin Project. We are grateful for both parties.

[2]Some notable early adopters are Whole Foods and a handful of tech start-ups (Loudenback, 2017).

[3]Wage discrimination is illegal on the basis of race and sex in the United States. See the Equal Pay Act of 1963 (1963) and the Civil Rights Act of 1964 (1964). If workers have different level of productivity or their job content is different, it is not illegal to compensate them for their work differently.

have a positive willingness to pay for their peers' wage information –, but rather due to a taboo around salaries (Cullen and Perez-Truglia, 2018b). Therefore, wage transparency policies could be valuable from the perspective of the workers.

The literature further finds that wage transparency policies are successful in reducing wage dispersion (Cullen and Pakzad-Hurson, 2019; Mas, 2017; Baker et al., 2019; Obloj and Zenger, 2020) and decreasing the gender pay gap (Baker et al., 2019; Kim, 2015; Bennedsen et al., 2019)[4]. Researchers, however, also uncovered that wage transparency policies have several unintended consequences. The policy increases quit rates (Mas, 2017; Cullen and Perez-Truglia, 2018a), reduces low earners' job satisfaction (Card et al., 2012) and effort (Cullen and Perez-Truglia, 2018a; Cohn et al., 2011; Nosenzo, 2013). Thus the overall welfare effect of the policy is ambiguous.

In this paper we show that wage transparency policies have further possible unintended consequences: workplace hostility may increase and job seekers may accept a different job offer than under wage secrecy, leading to systematic sorting across workplaces.

We ran an online experiment in which we control not only whether wages are transparent or secret, but the underlying wage allocations as well. Participants complete two incentivized tasks: a costly punishment game and a workplace choice game. The first one models hostility at the workplace in a hierarchical set-up, while the second one models the job acceptance decision of job seekers. In the first game, the punishment game, participants are assigned to a team of four. One of the players is in the role of the supervisor, while the other three are workers. In the first stage of the game the supervisor decides on the wage allocation, then we randomly reveal either the entire wage allocation in the team for a selected worker (wage transparency condition) or only their own wage (wage secrecy condition). In the second stage, the

[4]However, Gulyas et al. (2020) does not find that wage transparency in Austria affect the gender pay gap.

previously selected worker may subtract money from their team members – the supervisor and the two coworkers. We refer to this act as punishment and will use it to measure hostility towards others at the workplace.

We find in the punishment game that the treatment effect of wage transparency depends on the wage allocation in the teams, in particular, on the wage inequality among workers. At low levels of inequality wage transparency has no effect on hostility, but at high levels of inequality the treatment effect can be positive or negative depending on the wage of the participant who can punish. We also see that participants change their punishment pattern. When wages are secret, participants punish predominantly the supervisor, who is responsible for the wage allocation; when wages are transparent, participants increase the punishment they give to their relatively richer coworker. This result is surprising because the coworkers are by-standers and had no impact on the wage allocation.

The second task in the experiment measures the effect of wage transparency on job acceptance. The game is a choice task in which participants choose between two teams that they want to join under the two conditions: when they only know their own wage offers, and when they also know what others make in the two teams.

In this second task we document that the wage allocations in the two teams affect participants' workplace choice, even conditional on their own offered wage. We also see that in about 14% of the cases, participants reversed their workplace choice when they learned the others' wages in the teams. We interpret this finding as a sign for likely general equilibrium effects: if wages are transparent, workers may systematically sort to different workplaces by voting with their feet. This sorting behavior in turn could initiate changes in the wage setting behavior of the firms.

Both sets of our results can be explained through the lens of social preferences. Our theoretical framework implies that wage transparency can change optimal decisions if

two conditions are simultaneously true: 1) participants have inaccurate beliefs about the wage allocations when wages are secret, and 2) participants' utility functions depend on the other players' wages (i.e. participants have social preferences). We find empirical evidence for both conditions and argue that our treatment effects are mainly driven by inequality averse and reciprocal participants.

This paper contributes to two branches of the existing literature. First, the paper speaks to the labor literature by providing experimental evidence on the effect of wage transparency on hostility and sorting. Three papers – Cullen and Perez-Truglia (2018a), Breza et al. (2017) and Huet-Vaughn (2015) – have particular relevance to our research. Cullen and Perez-Truglia (2018a) randomized wage transparency in a field experiment and measured its effects on effort measures, while Breza et al. (2017) randomized wage inequality within teams and estimated the effect on attendance, output and cooperation in a field experiment. Relative to these two papers we control wage transparency and the wage allocations at the same time, which allows us to trace the treatment effect of wage transparency as a function of inequality in our experiment. This finding echoes Huet-Vaughn (2015), who shows that wage transparency has a different effect on effort depending on the level of wage inequality. Our study focuses on new outcomes, hostility and workplace sorting, and measures the effect of wage transparency across numerous levels of inequality for each participant.

Second, this paper also contributes to the laboratory experimental literature. In the first game, we study the effect of wage transparency on punishment. To the best of our knowledge so far only Hauser et al. (2019) linked punishment games to wage transparency using a public good game. We contribute by studying the effect of wage transparency in a hierarchical punishment game that follows an allocation stage. Our second game, the workplace choice game, speaks to other choice games that elicit distributional preferences, such as Fisman et al. (2018). We elicit distributional

preferences under both wage secrecy and transparency in a within-subject design and show that workers' choice can be reversed under the two regimes.

The paper is organized as follows. Section 1.2 introduces our theoretical framework. Section 1.3 describes the experimental design, and Section 1.4 summarizes our data set. Next, in Section 1.5 we will show our analyses. We discuss how our game relates to workplaces in Section 1.6 and conclude in Section 1.7.

1.2 Theoretical Framework

In this section we describe the theoretical underpinnings of our experiment. The framework implies that wage transparency can have treatment effects under two necessary conditions: 1) the agent has inaccurate beliefs about the wage allocation when wages are secret, and 2) the agent has social preferences. Condition 1) ensures that wage transparency reveals new information for the agent. Condition 2) establishes that the agent's utility function depends on the wage distribution and so a shock may affect the utility maximizing problem of the agent.

1.2.1 Utility Function

We assume that the agent has a utility function that depends on their own wage (w_i) and potentially on a vector of their teammates' wages ($\mathbf{w}_{-i}$). In particular, we consider the following utility function, inspired by Charness and Rabin (2002):

$$U_i(w_i, \mathbf{w}_{-i}) \quad = \quad \alpha * f(w_i) \quad + \quad (1-\alpha) * g(w_i, \mathbf{w}_{-i}) \tag{1.1}$$

, where α is a weighting parameter ($0 \leq \alpha \leq 1$), $f(\cdot)$ is a function that depends only on the agent's own wage, and $g(\cdot)$ is a function that has two arguments, the

agent's own wage and the vector of other players' wages, $\mathbf{w}_{-i}$. We refer to $f(w_i)$ as the self-regarding part, and to $g(w_i, \mathbf{w}_{-i})$ as the other-regarding part of the utility function.

We are going to be agnostic about the particular functional form of the other-regarding part of the utility function ($g(w_i, \mathbf{w}_{-i})$) because our argument holds without functional form assumptions. At the same time, it may be useful to provide examples for specifications used in the literature. Charness and Rabin (2002) define the following utility specification:

$$U_i(w_i, \mathbf{w}_{-i}) \quad = \quad \alpha * w_i \quad + \quad (1-\alpha) * \underbrace{\left[\delta * \min\{\mathbf{w}\} + (1-\delta) * \left(w_i + \sum \mathbf{w}_{-i}\right)\right]}_{g(w_i, \mathbf{w}_{-i})}. \tag{1.2}$$

The expression in the brackets is the equivalent to our $g(w_i, \mathbf{w}_{-i})$, and is the weighted sum of the lowest wage in the entire team ($\min\{\mathbf{w}\}$) and the sum of the total wealth ($w_i + \sum \mathbf{w}_{-i}$). This specification incorporates two common social preference functions: if δ is 1, the agent has Rawlsian social preferences, if on the other hand, $\delta = 1$, the agent is a social surplus maximizer. As another example for specifying the other regarding part of the utility function, $g(w_i, \mathbf{w}_{-i})$ can also be set to incorporate inequality aversion. A simple way would be to let $g(\cdot)$ be the sum of the absolute differences between each players' wages ($g(w_i, \mathbf{w}_{-i}) = \sum |w_i - w_{-i}|$), or to let $g(\cdot)$ be a negative function of the variance of the wages in the team ($Var(w_i, \mathbf{w}_{-i})$). More complex models of inequality aversion can be adapted to the form of Equation 1.1 as well, such as the Fehr-Schmidt model (Fehr and Schmidt, 1999).[5]

After providing examples for utility functions that are encapsulated in Equation

[5]The functional form of the Fehr-Schmidt preference is: $U_i(w_i, \mathbf{w}_{-i}) = w_i + [-\lambda_i \max\{\mathbf{w}_{-i} - w_i, 0\} - \nu_i \max\{w_i - \mathbf{w}_{-i}, 0\}]$, adapting the notation to ours. The first term of the utility function is the self-regarding part of the utility function, while the bracketed term is the other-regarding part of the utility function. The two parts of the utility functions are weighted equally and after a normalization of the weights, their sum can be set to 1, to match the form of Equation 1.1.

1.1, we continue our theoretical explanation using the general specification. Based on Equation 1.1, we can define the marginal utility of another player's wage (w_{-i}) for player i as:

$$\frac{\partial U_i(w_i, \mathbf{w}_{-i})}{\partial w_{-i}} = (1-\alpha)\frac{\partial g(\cdot)}{\partial w_{-i}} \tag{1.3}$$

, where w_{-i} is an element of $\mathbf{w}_{-i}$ vector.

If $\alpha = 1$, the agent does not assign any weight on social preferences and hence their marginal utility of any other player's wage is zero. In other cases, when $\alpha \neq 1$, the agent cares about the entire wage distribution, and their marginal utility defined in Equation 1.3 is not zero if $\frac{\partial g(\cdot)}{\partial w_{-i}} \neq 0$. Because we formalized $g(\cdot)$ with the argument of $\mathbf{w}_{-i}$ is one of them, we implicitly assumed that $\frac{\partial g(\cdot)}{\partial w_{-i}} \neq 0$. For this reason, the marginal utility can be zero only if $\alpha = 1$.

The marginal utility of others' wages, defined in Equation 1.3, is important to understand the possible patterns of behavior in our two tasks. First, in the punishment task participants are offered to reduce the wages of the other players in their team at a cost. If $\alpha = 1$, the marginal utility of all the other players' wage for the participant is zero, so the participant has no incentive to change the wages of the others at a cost. If $\alpha \neq 1$, so the marginal utility is not zero, the participant has an incentive to change the wages of the others. Therefore if a participant decides to change the wages of the other players, it is consistent only with utility functions that have social preferences. In the second task participants are choosing between two job offers and the marginal utility will again show some insights. The interesting case is when the two offers are the same in terms of the offered wage for the participant, but differ in the wage allocation for the other players. An agent with $\alpha = 1$ would be indifferent between the two offers, because for them the others' wages does not matter in their maximization problem. But an agent with $\alpha \neq 1$ may not be indifferent because different wage distributions could lead to different levels of utility for the agent in the

presence of social preferences even if their own wage is the same.

Using a utility function that embeds social preferences makes it possible that the participants in our games will care about the entire wage allocation and not just their own wage. Next, we will discuss how the wage transparency treatment can affect the utility maximization problem of the agent and under what conditions can we expect different optimal decisions under the two regimes.

1.2.2 Wage Transparency Treatment and Optimization

What would happen with agents who have utility functions as in in Equation 1.1 under wage transparency and wage secrecy? Under wage transparency, the agent's utility function is as shown in Equation 1.1; the functional form and all inputs of the utility function are known to the agent. The difference under wage secrecy is that some of the inputs of the utility function – the other players' wages – are unknown to the agent.[6] So to maximize utility, the agent needs to form expectations over the unknown inputs of the utility function conditional on their information set ($\mathcal{I}$). The maximand, the expected utility of the agent, can be written as:

$$\mathbb{E}\Big[U_i(w_i, \mathbf{w}_{-i})\Big|\mathcal{I}\Big] \quad = \quad \alpha * f(w_i) \quad + \quad (1-\alpha) * \mathbb{E}\Big[g(w_i, \mathbf{w}_{-i})\Big|\mathcal{I}\Big] \tag{1.4}$$

, where the agent's information set ($\mathcal{I}$) contains their own wage (w_i) and the rules of the game.

Under this set up, we would expect different optimal decisions under wage transparency and wage secrecy only if the maximands in the two conditions are different. That is, if $U_i(w_i, \mathbf{w}_{-i}) \neq \mathbb{E}\Big[U_i(w_i, \mathbf{w}_{-i})\Big|\mathcal{I}\Big]$. Because the self-regarding part of the

[6] We assume that the agent faces uncertainty only through the inputs of the utility function when wages are secret, and thus we impose that the agent always knows the functional form of their utility function.

utility function is known under both wage transparency and wage secrecy, the difference could only stem from the second, the other-regarding segment of the utility function.

There are two necessary conditions for the other-regarding segment of the utility function to be different under wage secrecy and wage transparency. First, it needs to be the case that $\mathbb{E}\left[g(w_i, \mathbf{w}_{-i})\middle|\mathcal{I}\right] \neq g(w_i, \mathbf{w}_{-i})$, meaning that the other-regarding preference segment of the utility function, $g(\cdot)$, evaluated at the expected wage allocation has to be different than when it is evaluated at the realized wage allocation. Because we assume that the only uncertainty for the agent arises from the inputs of the utility function, this condition is met if the expected wage distribution is not the same as the realized one. The second necessary condition is that this difference between the expected and realized $g(\cdot)$ function should matter in the optimization process of the agent; they should have a non-zero weight on the other-regarding segment of the utility function ($\alpha \neq 1$). The intuition is that if the agent has inaccurate beliefs about the wage allocation but has no social preferences ($\alpha = 1$), the wage transparency treatment will not change their optimal decision because correcting the error in their beliefs have no effect on their optimization at all. On the flip side, if the agent has social preferences but they have perfect beliefs, the treatment has no scope in correcting their beliefs and cannot shift their optimal response in the framework that we laid out.

To summarize the theoretical framework, we use a utility function that can depend on the other players' wages. Under wage transparency all inputs of the utility function are known, but when the wages are secret, the agent does not know the other players' wages ($\mathbf{w}_{-i}$) and hence they will need to maximize their expected utility. If the optimal solutions of the maximization problems differ under wage secrecy and wage transparency, we would see that wage transparency has a treatment effect.

Our framework suggest that it is only possible if participants have inaccurate beliefs about the wage distribution when wages are secret and they have other-regarding preferences.

1.3 Experimental Design

The experiment was designed to measure the treatment effect of wage transparency on two decisions: hostility towards colleagues and sorting to workplaces. We modelled the hostile behavior with a costly punishment game, and the sorting decision with a workplace choice game.

1.3.1 Punishment Game

The punishment game models workers' hostile behavior towards their colleagues. The game is played in teams of four players: one player is in the supervisor's role and three other players are the "workers". First, the supervisor chooses the wage allocation for the team. Then, in the second stage one of the three workers – the agent – decides on whether they want to reduce the payment of anyone else in the team, and if so by how much. The act of reducing an other players' wage is what we refer to as punishment.

We geared the design of this game towards the punishment phase because this stage can answer how hostility would change at a workplace with a given wage allocation if wage transparency is introduced. To control the wage allocations while remaining truthful to the participants, we pick a supervisor to the team who made the desired wage allocation[7]. We also restricted the options of the supervisor to further stir their decisions. For these reasons, the allocation stage should be considered as a

[7]The other option would have been to use a computer or algorithm that creates the desired wage allocations instead of a human supervisor. This, however, would have been a poor solution to test for change in hostility among humans.

setup stage before the punishment decision and not a separate game.

In the punishment stage the information set of the agent contains basic features of the wage distributions. The agents know that the amount of money being allocated among the three workers in the team is 15 experimental currency units (ECU), that the supervisor made the allocation and that the supervisor may earn between 0 and 20 ECU[8]. Besides these guidance, the agent always knows their own wages, but only treated agents know the entire wage allocation in the team. Therefore, the treatment could be thought of as a change in the information set of the agents.

Our punishment stage generally follows the design features of similar games in the literature, such as Fehr and Fischbacher (2004) and Bartling and Fischbacher (2012). The first important design element of punishment games is that punishment is costly. This means that those agents who do not derive marginal utility from reducing other players' wages, will not engage in punishment. Together with the other features of the punishment game detailed below, this will ensure that observing punishment is an evidence for existence of social preferences.

We deviated from the classic design by changing the unit of payment in which participants have to pay for punishment. In a typical punishment game, the cost of punishment is in monetary units (ECUs) and participants pay the cost from their designated fixed punishment budget. The unspent part of the punishment budget becomes part of the participant's payment. Instead of this procedure, participants in our game have to pay for punishing others in time units: for each ECU the agent wishes to subtract from the other players' wages, they will need to solve a given number of time-burning tasks. This change is necessary to mitigate two criticisms about the standard design: house money effects and wealth effect.[9]

[8]The supervisor's wage is separate from the workers' pool of money.

[9]These concerns do not impact estimating the treatment effect of wage transparency for a given wage allocation. However, when we will compare the levels of punishment for different wage alloca-

The first criticism is the house money effect (Thaler and Johnson, 1990). This says that participants are less careful spending money windfall money than they earned money. From the perspective of the participants, the punishment budget from the experimenter is windfall money, suggesting that researchers may overestimate the level of punishment if they use the classic design. Because we do not give a budget for the participants, our method is not susceptible to the house money effect.

The second criticism of the typical punishment game design is the wealth effect. This says that the cost of punishment depends on the starting endowment of the participant, which in our game is the wage of the agent. At low levels of wealth punishment is relatively more costly than at high levels of wealth. This is because when punishment is priced in money and participants get their unspent punishment budget, then the marginal cost of punishment is the marginal utility of money. But because the marginal utility of money is decreasing as the wealth of the participant increases, the marginal cost of punishment does the same. This potentially leads to underestimating the level of punishment when the agent is relatively poor and overestimating the punishment when the agent is relatively rich. With our tweak of pricing the punishment in time, we break the direct relationship between the cost of punishment and the wealth of the participant, which allows us to compare punishment at different wage levels of the agent.

Besides the costly punishment, another important feature of the punishment game is that typically only one player – the agent – can punish the other players. Suppose more than one player could punish. If the preferences of these agents were aligned about whom to punish and by how much, each specific player has an incentive to free-ride on the other players and not pay the cost for the punishment. Therefore punishment would be a public good. If, however, the agent knows that they are

tions that give potentially different wages for the agent, estimating the level of punishment for each allocation is important.

the only one who can punish, the agent has no option to free-ride. It also rules out the motivation for retaliating *anticipated* contemporaneous punishment from the other players. Altogether, this restriction helps eliminating alternative explanations for punishment other than social preferences and so in our game as well only one participant can punish and the participant knows that they are the only one who can punish in the team.

1.3.2 Workplace Choice Game

The workplace choice game helps us to understand how people choose between offers under wage secrecy and wage transparency.

In this game, agents receive two offers from two teams and they have to choose among them. The teams have the same structure as in the punishment game: one player is the supervisor and three players are workers out of which one spot is vacant and can be taken by the agent. The two teams consist of different players and the fact that the supervisors are different means that the wage allocations may differ in the two teams[10]. After seeing the two teams, the agent then needs to decide which offer they would like to accept. The agent is not forced to choose; they may report that they are indifferent between the offers, in which case, they are assigned to a randomly chosen team.

The agent always knows their offered wages in the teams and also the possible ranges of wages (supervisors receive between 0 and 20 ECUs and the total wage bill of the workers is 15 ECU). They also know that the wages were set by the supervisors of the respective teams. In the treatment condition, players additionally know the entire wage distribution of the two teams when they are making the decision. After

[10]We retain control over the wage allocations in this game, too by implanting the supervisors to the team who truthfully selected the allocations we wanted to test.

the agent makes a choice, the game ends.

1.3.3 Procedural Details

Both games are played in a repeated one-shot fashion: we assign the agents to new teams in each round for which they make a new decision. The repeated feature of the games allows us to test the effect of wage transparency for different wage allocations for the entire sample. We add the one-shot feature so that future interactions and reputation building will not be able to explain the agents' decisions – further solidifying our argument that punishment, if any, is due to social preferences.

At the beginning of the experimental session, we randomized half of the participants into the treatment (wage transparency) and another half into the control (wage secrecy) group for the punishment game. All participants made 25 rounds of punishment decisions for the same set of predetermined wage allocations. In each round participants interacted with different team members and faced different wage allocations. Participants in the wage transparency arm first saw the given wage allocation and then were asked to make their punishment decision. We capped the total punishment in a team at 10 ECU, but allowed participants to allocate their punishment in any way they wanted across the other players[11]. If participants selected non-zero punishment, they were reminded of the time cost they needed to pay in order to carry out their decision. Participants in the wage secrecy condition saw only their own wage, the rest of the wage allocation remained hidden for them. After they learned their own wage, we elicited their beliefs about the wage allocation in an incentive compatible way[12]. Next, we asked these participants to make their punishment decision

[11]We did not allow for rewarding/aiding others, self-punishment or fractional punishment (such as 0.5 ECU), but we did allow participants to punish 0 ECUs. The cap was enforced to eliminate the possibility of bankruptcy.

[12]We achieved incentive compatibility in the belief elicitation by rewarding participants if they

in the same way as those in the wage transparency arm: they were allowed not to punish (punish 0 ECU) and they were reminded of the cost of the punishment.

In the workplace choice task, participants decided which of the two teams they wish to join. They could choose either of the teams, or state that they are indifferent, in which case they were randomly assigned to one of the teams. Participants solved 9 of these choice tasks twice: first under the wage secrecy regime and then under wage transparency regime. Similarly to the punishment game, when wages were secret, we elicited participants' beliefs about the wage allocations in the two teams. Because this game has a within-subject comparison for the same choice tasks, we did not randomize the order of the wage secrecy and transparency condition as participants' beliefs would be mechanically more accurate if they had seen the wage distribution of the teams.

The wage allocations that we use in the experiment were predetermined, meaning that the wage allocations (and their characteristics) are orthogonal to the participants' characteristics. This will allow us to estimate the causal effect of wage inequality on punishment and sorting decisions. Two additional features of the design support the causal interpretation. First, the wage distributions were picked such that there is variation in wage inequality conditional on the agent's wage. Therefore, we can separate the effect of own payment from the effect of inequality[13]. Second, we randomize the viewing order of the rounds within each tasks for each participant. Because of this, time varying factors – for example, decreasing attention or learning – will not be able to explain our results. All of these design features – the careful selection of wage

were close to the actual wage allocation. Participants formed their beliefs knowing their own wage and based on the general characteristics of the wage allocations: the wage bill of the workers is always 15 ECUs and the supervisors' wage is between 0-20 ECUs.

[13]In two-player games, such as the dictator or the ultimatum game, this is not feasible: there is a mechanical effect between one's own payment and inequality. In our game, we can hold the agent's payment fixed and redistribute payments among the other players to adjust inequality.

allocations before data collection and the randomized viewing order – imply that the wage allocations can be thought of as secondary treatments.

We achieve incentive compatibility in the games by paying incentive bonus payments for participants. At the end of the survey, one punishment round and one sorting round was picked and actualized. In the randomly chosen punishment round all players in the team received their respective wage minus their received punishment (if any) and the agents were asked to pay for the punishment. As discussed earlier, the agents paid in time for the punishment; for every ECU that the agent reduced, they needed to solve a given number of time burning-task that involved clicking on a button every 5 seconds. In the randomly chosen sorting task, the agent received their own offered wage of the chosen team as well as their team mates.

Our design had an additional built in element that guarded coworkers and supervisors not to have negative final bonus payment. Coworkers and supervisors could lose money in the punishment game if they received more punishment from the agent than what they got from the supervisor. To make sure that losing money is not possible, we capped the total punishment in each round at 10 ECUs and gave all players (including the agent) an additional 10 ECU bonus for participating[14]. This way even if a player had 0 ECU after the allocation phase and received all of the punishment from the agent in the actualized round, the player did not go bankrupt.

1.4 Data Collection

We created three different surveys in Qualtrics that we sequentially deployed to Prolific, a market place for survey takers, in 2020 August and September. First, we

[14]Because every player received this surprise bonus, it did not distorted the relative wages and only served as a way not to ask money back from coworkers and supervisors. When we are analyzing the wages of the players after the punishment, we do not take this cushion into account.

collected the survey responses from the coworkers, then from the supervisors and finally from the agents. This procedure ensured that we do not use deception, we preserve the anonymity of the participants in the games and that all Prolific survey takers can participate at most in one of our three surveys. Note that all players in the games are real participants, even the idle coworkers[15]. However, our analysis is solely based on the agents' survey.

The agents' survey was advertised as an hour long decision making study and was available only for US based adults, fluent in English. About 300 respondents completed this survey, but due to survey errors, we have 294 participants in the final sample. The average payment for participants who completed this survey was \$9.74.

The survey had five sections. First we collected participants starting beliefs about the wage allocations. Second, in an effort to mimic a real workplace, participants "worked". The job was to estimate the positions of several numbers on a number line. Participants did not receive feedback about their accuracy on this estimation task, but knew that similar to real workplaces, their supervisor may have some coarse information about how well they did. Next, participants solved the punishment task either in the treatment (wage transparency) or control (wage secrecy) condition depending on the between subjects randomization. Then in the third main section participants completed the 9 workplace choice task first in the secrecy version and then in the transparency version. Finally, at the end of the survey, we collected demographic information about the participants. We also merged some additional information about the participants to our data set, such as their political views or socio-economic status, collected by Prolific.

Table 3.9 summarizes the characteristics of the agents' population using information from our survey and Prolific's information about the participants. The table

[15]We opted for this design to ensure that we can truthfully inform participants that they are interacting with other real participants.

reports the mean, the standard deviation and in case of non-binary variables, the 10th and 90th percentiles of the variables. We report the probability that the treatment assignment in the punishment task is balanced on participants' characteristics in the last column. The table shows that while our sample is likely a selected sample, the between subjects treatment assignment in the punishment game is balanced.

1.5 Analysis

In this section we present our findings. First, we show the main results from the punishment game and then from the workplace choice game. Next, we will provide evidence for the mechanisms that the theoretical framework suggested (Section 1.2). We pre-registered most of the analysis presented here. We mark deviations and exploratory analysis in footnotes.

1.5.1 Punishment Game

We start with the analysis of the punishment game and show that pooling all punishment rounds to calculate the average treatment effect of wage transparency is misleading because the effect of wage transparency depends on the underlying wage allocation.

Average treatment effect of wage transparency across all rounds

We first estimate the average treatment effect of wage transparency on punishment decisions pooled over all rounds. We estimate the following regression:

$$Y = \alpha + \beta * T + \varepsilon \tag{1.5}$$

, where Y is an outcome variable related to the punishment decision and T is the wage transparency treatment indicator that has value one when the participant is in the wage transparency arm. Finally, ε represents the error. The coefficient of interest is the average treatment effect of wage transparency averaged over all rounds (β). Because wage transparency was randomly assigned, this parameter is identified.

Table 1.1 shows the results of estimating Equation 1.5. We have two types of outcome variables: the level of punishment expressed in experimental currency units (Panel A), and indicator variables that show whether there is non-zero punishment for a given player (Panel B). We calculated q-values (Benjamini et al., 2006; Anderson, 2008) to adjust for multiple hypothesis testing.[16] Based on Panel A, the supervisors receive 0.369 ECU more punishment when wages are transparent than when wages are secret. This effect, however, fails to remain significant when we use the sharpened q-values that adjust for multiple hypothesis testing. In columns 2 and 3, we can see that the treatment significantly decreases the amount of punishment towards poorer coworkers and increases it towards the richer coworkers. Both of these effects are statistically significant at 5% level based on the q-values. The average treatment effect on the total amount of punishment (column 4) is statistically insignificant, meaning that wage transparency does not increase total punishment, only reallocates punishment from the poorer coworker to the richer coworker. As a side-effect, the punishment will be more concentrated in the rounds as shown by the last column: the maximum punishment in a round[17] is about 0.5 ECU higher when wages are transparent. This effect is significant at the 10% level after the false discovery rate

[16]The family that we use for calculating the q-values includes all treatment effects in Table 1.1 and two additional variables: the level of punishment received by the two coworkers and whether the two coworkers receive positive punishment in the round. Adjusting for multiple hypothesis testing across all of these outcomes is a conservative method; one could argue that testing the hypothesis on supervisor punishment is different than testing it on the coworkers. For this reason we also report unadjusted p-values in the table.

[17]We did not pre-register the maximum amount of punishment in a round as an outcome variable.

(FDR) adjustment.

Panel B shows the average treatment effect of wage transparency on the probability of positive punishment towards the specific team members. The coefficients are estimated based on Equation 1.5, which assumes a linear probability model specification. According to both the p and q-values, the only treatment effect that is statistically significant at 5% is the coefficient on the probability of punishing the poorer coworker: the poorer coworkers are less likely to be punished under wage transparency by 8.6 percentage points.

The results from Table 1.1 are not sensitive to adding personal characteristics of the participants as control variables. We also estimated the punishment regressions towards the three players (supervisor, poorer coworker, richer coworker) simultaneously. The results were not sensitive to this change either.

Average treatment effect by wage distribution

In this section we show that the results in the previous section mask important heterogeneity across wage allocations. In the theoretical section (Section 1.2) we argued that when the agents assign non-zero weight on the other-regarding preferences part of the utility function, that is $\alpha \neq 0$, the wage allocations matter for the agents' optimization problem. This means that different wage allocations can also generate different treatment effects depending on the beliefs of the participants and the functional form of their the other-regarding preferences. To test if this is the case, we are going to estimate the average treatment effect for each of the 25 distributions separately that we used in the punishment game[18].

We estimate

$$Y_k = \alpha_k + \beta_k * T + \varepsilon_k \tag{1.6}$$

[18]This heterogeneity analysis in not in the pre-analysis plan.

Table 1.1: Treatment Effects on Punishment Behavior

Panel A: level of punishment					
	supervisor	coworker		total	max
		poorer	richer		
treat	0.369	-0.322	0.294	0.341	0.499
	(0.206)	(0.083)	(0.097)	(0.310)	(0.232)
p-value	0.075	0.000	0.003	0.272	0.032
q-value	0.118	0.002	0.013	0.305	0.069
control group mean	1.505	0.632	0.584	2.722	1.977
# observations	7347	7348	7349	7347	7347
# clusters	294	294	294	294	294
R^2	0.004	0.018	0.007	0.002	0.006

Panel B: indicator for non-zero punishment				
	supervisor	coworker		anyone
		poorer	richer	
treat	0.029	-0.086	0.045	0.034
	(0.037)	(0.029)	(0.032)	(0.041)
p-value	0.440	0.004	0.152	0.406
q-value	0.388	0.013	0.198	0.388
control group mean	0.340	0.228	0.211	0.445
# observations	7347	7348	7349	7347
# clusters	294	294	294	294
R^2	0.001	0.012	0.003	0.001

Notes: Coefficients are from regressing each outcome on wage transparency treatment assignment. Heteroskedasticity-robust standard errors are shown in parentheses, clustering at the participant level. Sharpened q-values control the false discovery rate across the family of punishment outcome variables. The family includes all variables in the table as well as the sum of punishment towards the two coworkers and an indicator if this sum is positive. The p-values are unadjusted and are reported for comparison purposes.

for each wage allocation $k \in [1, 25]$. We are interested in the β_k coefficients that are identified due to the random assignment.

Figure 1.1 plots these β_k coefficients for the supervisor punishment. The coefficients are ordered based on the inequality among the workers' wage, which we measure by the Gini coefficient[19]. Low values of the coefficient indicate low levels of inequal-

[19]The Gini coefficient is defined as $G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}}$ (Sen et al., 1997). This coefficient has

ity, higher values mean higher levels of inequality. Our first observation is that the treatment effect of wage transparency on supervisor punishment varies across wage allocations. Some wage allocations have zero, others significantly positive and yet others significantly negative treatment effects. When we averaged over these coefficients in the pooled estimation (Equation 1.5), we found statistically insignificant effects, highlighting that taking into account wage allocation differences is important to understand the effect of wage transparency. Our second observation is that most of the zero treatment effects are concentrated at low levels of inequality and the treatment effects fan out when the inequality among the workers is higher.

This figure was picked for illustration to build intuition about why controlling the wage allocations in our design was crucial. In what follows, we will regress the outcome variables on wage transparency treatment assignment and the characteristics of the wage allocation.

Average treatment effect as a function of the wage distribution

After we showed that the treatment effect of wage transparency has a lot of variation depending on the underlying wage allocation, we extend our model and incorporate three characteristics of the wage allocation to the treatment effect model as regressors: participant's own wage, the supervisor's wage and the Gini coefficient defined over the workers' wages[20]. We also add the interactions of these variables with the treatment to see if the effect of wage transparency is moderated by the wage distribution. These modifications lead to the following model:

a value of 0 for the perfectly equal society and a value of 1 for a perfectly unequal society. We calculate the Gini coefficient over the wages of three workers and not entire societies, which means that the upper limit of the Gini coefficient in our game is $\frac{2}{3}$.

[20]In the pre-analysis plan we considered the Gini coefficient within the entire team. The qualitative message does not depend on which version we use, but the interpretation is cleaner when we define the Gini coefficients over the workers. This change, however, meant that we also needed to include the supervisors' wage in the analysis.

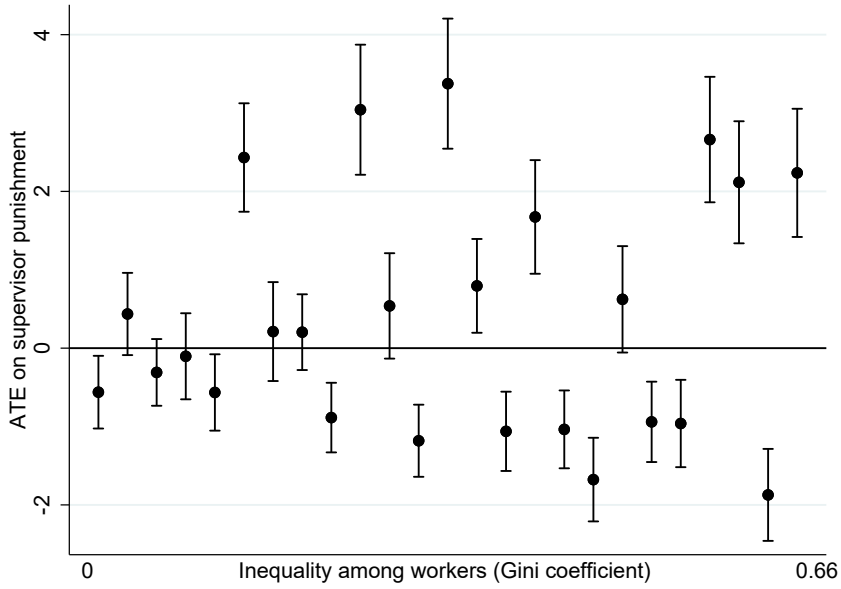


Figure 1.1: Average Treatment Effect of Wage Transparency on the Level of Supervisor Punishment

Notes: This graph shows the average treatment effect of wage transparency on the level of supervisor punishment by the level of inequality among the workers. the x-axis is an ordinal axis where we ordered the wage allocations from the lowest to highest Gini coefficient. The average treatment effects are the beta coefficients from regressing round level supervisor punishment on treatment assignment separately for each round. Heteroskedasticity-robust standard errors clustered at participants level were used to create the wings, which mark the 95% confidence intervals.

$$\begin{aligned} Y = \quad & \alpha \ + \ \beta * T \ + \ \delta_1 * Gini_{workers} \ + \ \delta_2 * (Gini_{workers} \times T) + \\ & + \ \gamma_1 * w_{own} \ + \ \gamma_2 * (w_{own} \times T) \ + \ \theta_1 * w_{sup} \ + \ \theta_2 * (w_{sup} \times T) \ + \ \varepsilon \end{aligned} \tag{1.7}$$

, where T stands for the wage transparency treatment, $Gini_{workers}$ is the level of wage inequality among the three workers as measured by the Gini coefficient, w_{own} is the agents' own wage and w_{sup} is the supervisor's wage. As before, we are clustering the standard errors at the participant level. Our interpretation of the coefficients will remain causal, as the characteristics of the wage allocations are orthogonal to the characteristics of the participants, and we randomized the treatment assignment to

wage transparency.

Table 1.2: Treatment Effects on the Level of Punishment by Allocation Characteristics

Level of punishment outcomes						
	supervisor	coworker			total	max
		poorer	richer	both		
treat	-1.330***	-0.158	0.703***	0.544**	-0.787**	-0.844***
	(0.224)	(0.123)	(0.157)	(0.267)	(0.383)	(0.255)
gini workers	1.164***	0.669***	0.467***	1.136***	2.300***	1.691***
	(0.226)	(0.107)	(0.122)	(0.181)	(0.294)	(0.240)
gini workers × treat	-0.497	-1.220***	3.047***	1.827***	1.330***	2.536***
	(0.317)	(0.142)	(0.283)	(0.314)	(0.448)	(0.387)
own wage	-0.055***	-0.067***	-0.053***	-0.120***	-0.176***	-0.106***
	(0.016)	(0.008)	(0.009)	(0.016)	(0.021)	(0.017)
own wage × treat	0.032	0.052***	-0.146***	-0.094***	-0.062**	-0.083***
	(0.020)	(0.009)	(0.016)	(0.022)	(0.028)	(0.023)
supervisor wage	0.005	0.007**	0.002	0.009**	0.013**	0.010*
	(0.006)	(0.003)	(0.002)	(0.004)	(0.006)	(0.005)
supervisor wage × treat	0.216***	-0.007*	-0.073***	-0.080***	0.136***	0.120***
	(0.016)	(0.004)	(0.006)	(0.007)	(0.013)	(0.013)
control group mean	1.505	0.632	0.584	1.217	2.722	1.977
# observations	7347	7348	7349	7347	7347	7350
# clusters	294	294	294	294	294	294
R^2	0.160	0.064	0.207	0.134	0.129	0.137

Notes: This table estimates the treatment effect of wage transparency using wage allocation characteristics as additional controls and allowing for interaction effects. The Gini coefficient is defined over the three workers and thus excludes the supervisors' wage. *total* refers to the total amount of punishment points in the given round, *max* is the maximum amount of punishment towards another player within a round. Heteroskedasticity-robust standard errors shown in parentheses, clustered at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

The estimated results are summarized in Table 1.2 and in Table 1.3. In Table 1.2 we present our findings for the level of punishment outcome variables. The table uncovers several important results. First, the table shows that relative to the findings

Table 1.3: Treatment Effects on the Propensity of Punishment by Allocation Characteristics

Indicator outcomes for non-zero punishment					
	supervisor	coworker			anybody
		poorer	richer	both	
treat	-0.138***	0.000	0.183***	0.114**	-0.062
	(0.045)	(0.045)	(0.050)	(0.052)	(0.052)
gini workers	0.108***	0.124***	0.092***	0.128***	0.156***
	(0.034)	(0.034)	(0.031)	(0.033)	(0.036)
gini workers × treat	-0.310***	-0.450***	0.166***	0.132***	0.081
	(0.053)	(0.053)	(0.050)	(0.050)	(0.051)
own wage	-0.009***	-0.019***	-0.017***	-0.022***	-0.018***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)
own wage × treat	0.010***	0.013***	-0.017***	-0.012***	-0.005
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
supervisor wage	0.001	0.001	0.000	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
supervisor wage × treat	0.029***	0.001	-0.010***	-0.010***	0.013***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
control group mean	0.340	0.228	0.211	0.269	0.445
# observations	7347	7348	7349	7347	7347
# clusters	294	294	294	294	294
R^2	0.104	0.063	0.103	0.103	0.070

Notes: This table estimates the treatment effect of wage transparency using wage allocation characteristics as additional controls and allowing for interaction effects. The Gini coefficient is defined over the three workers and thus excludes the supervisors' wage. *total* refers to the total amount of punishment points in the given round, *max* is the maximum amount of punishment towards another player within a round. Heteroskedasticity-robust standard errors shown in parentheses, clustered at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

in Table 1.1, adding the extra terms changes the main effects of the wage transparency treatment. The main effect of the treatment is now statistically significant for all punishment variables, except for hostility towards the poorer coworker. These findings are qualitatively and quantitatively different than what we saw earlier.

Second, two characteristics of the wage distribution (Gini coefficient among the workers and own wage) directly affect the punishment decision of the participants. The Gini coefficient has a positive main effect, while own wage has a negative main effect on hostility. The relationship between own wage and hostility confirms the previous findings of Fehr and Fischbacher (2004), who also see that own wage and punishment are negatively related.

Third, we find that the effect of wage transparency is moderated by the characteristics of the wage allocations. The coefficients on the interaction terms are almost always statistically significant at the 5% level, confirming our hypothesis that the effect of wage transparency depends on what is being revealed.

Finally, the results also indicate that participants treated the supervisors and their coworkers differently. There is a level difference between supervisor and coworker punishment in the control group; under wage secrecy, supervisors receive more punishment on average than the coworkers as shown by the control group means (1.5 ECU vs. around 0.6 ECU). In addition, the treatment effect of wage transparency is also different between supervisors and coworkers. The difference arises in terms of magnitude (i.e. Gini coefficient $\times$ treatment) and also in terms of sign (i.e. supervisor wage $\times$ treatment). These suggest that the effect of wage transparency does not only depends on what is being revealed, but also what is the role of the player in the game, whom we study.

Table 1.3 presents the findings for the indicator outcome variables that show whether there is positive punishment for the given players. Because the estimated

equation for this panel is also Equation 1.7, the coefficients are the results of linear probability models and thus should be interpreted as percentage point changes.

The results in this panel support our four main findings in Panel A of the Table. As before we see that adding the wage allocation characteristic change the main effect of wage transparency relative to the case when they are not included (Table 1.1, Panel B). Second, the Gini coefficient for the workers and the own wage of the participants have a significant main effect on the probability of punishment at the 1% level. Next, we also see evidence for statistically significant interaction effects: the characteristics of the wage distribution moderate the treatment effects. Lastly, this panel confirms that participants treated supervisors and coworkers differently. The control group means show that supervisors are punished in the wage secrecy arm more often than the coworkers: 34% versus about 22% for each of the coworkers. Besides the level effect in the control group, the main effect of wage transparency also varies for the different players.

We conclude the punishment analysis by highlighting our most surprising result. After incorporating the characteristics of the wage allocation to the analysis, we still see that the richer coworkers receive more punishment due to the treatment. Punishing the coworkers in our game means punishing the innocent by-standers. We offer our interpretation of this finding when we discuss the mechanisms in Section 1.5.3.

1.5.2 Workplace Choice Game

In this game we test the prediction of the social preference theory on the extensive margin: conditional on their own offered wage, will participants care about the wage allocation when they choose between offers? We find that our participants behaved as if the wage allocations were important for them when they chose among work-

places. We also recovered that a share of participants reversed their choice relative to their choice in wage secrecy. These findings propose a possible avenue how wage transparency policies can have general equilibrium effects: workers may change how they sort to workplaces when the policy is rolled out. This behavior in turn could affect the wage setting practices of the firms, leading to further treatment effects in general equilibrium.

We rely on the within-subject design of this game to test if participants accept the same offer under wage secrecy and wage transparency. All participants completed 9 choice tasks in this game, first in the wage secrecy version, then in the wage transparency version.[21] We find that in 14.11% of the cases, participants chose differently for the same choice task under the two regimes[22], that is, a share of our participants reversed their workplace choice when they learned the true wage allocation.

In order to understand choice reversals, we run a regression analysis. Our outcome variable is whether the participant chose differently in a choice task when the wages were transparent relative to when they were secret. We regress this choice reversal indicator on the characteristics of the previously chosen and not chosen teams' wage allocations[23]. We cluster the standard errors at the participant level.

Table 1.4 shows our results. If participants learn that the level of inequality among the three workers is higher in the team that they chose in the secrecy condition, they are more likely to switch away from that team and reverse their choice. On the flip side, if the previously not chosen team has higher level of inequality among the workers, participants are less likely to reverse their choice. Participants' own wage, on the other hand, is regarded positively: higher own wage in the previously chosen

[21]We fixed the order because we elicited participants' beliefs about the wage allocation when wages were secret. Had we randomized the order of the two regimes, those who started with wage transparency would have had mechanically more accurate beliefs.

[22]This number excludes the cases when choice reversals are mechanical due to indifference.

[23]This analysis is not included in the pre-analysis plan.

Table 1.4: Effect of Wage Allocation Characteristics on Choice Reversals in Workplace Choice Task

	Reversed chosen team
gini worker previously chosen team	1.051***
	(0.138)
gini worker previously NOT chosen team	-0.592***
	(0.063)
own wage previously chosen team	-0.056***
	(0.009)
own wage previously NOT chosen team	0.021***
	(0.005)
supervisor wage previously chosen team	-0.007
	(0.007)
supervisor wage previously NOT chosen team	0.004
	(0.007)
constant	0.402***
	(0.043)
# observations	1301
# clusters	287
R^2	0.151

Notes: The table shows the effects of wage allocation characteristics on the probability to reverse the preferred teams in the workplace choice task when wage allocations became transparent. The analysis excludes cases when participants were indifferent between teams and those rounds where participants own offered wage is the same in the two teams. Previously chosen team refers to the team that the participant chose in the secrecy condition. Previously not chosen team refer to the other team that the participant did not choose in the secrecy condition. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

team decreases the probability to switch and if the other team offers higher wage for the participant, the participant is more likely to switch on average. The sign of these results suggests that participants dislike inequality and prefer higher own wage. We do not find evidence for supervisors' wage affecting participants' choice reversals.

Our interpretation of the results is that wage transparency made some participants to reverse their choices in the workplace choice task, and these choice reversals were informed by the revealed wage allocations in the two teams.

1.5.3 Mechanisms

Having shown that wage transparency affects participants' optimal decisions in our games and that the wage allocation moderates these effects, in this section we test what mechanisms could explain these results. Our theoretical framework points to two possible mechanisms that could explain the treatment effects of wage transparency: beliefs and social preferences. We show empirical evidence for both in this section.

Beliefs

We show three pieces of evidence that beliefs are one of the possible mechanisms. We first demonstrate that participants' beliefs correlate with their punishment and workplace choice decisions when wages are secret, suggesting that participants' beliefs potentially have a role in the decisions they make. Second, we find that participants' beliefs are imperfect under wage secrecy. This finding checks one of the necessary conditions of finding a treatment effect based on our theoretical framework in Section 1.2. Finally, we show that wage transparency corrects participant's inaccurate beliefs and so the treatment is successful in creating a shock to beliefs.

Beliefs correlate with decisions in the wage secrecy arm

First, we analyze if participants' beliefs correlate with their punishment decisions. We regress the punishment outcomes on participants' beliefs about the wage allocation in the wage secrecy arm[24]. The results in Table 1.5 indicate that participants' beliefs about what an other player received correlate positively with the amount of punishment the participant gave to the respective other player (Panel A), and the

[24]We elicited participants' beliefs about the wage allocation only for those who were in the wage secrecy arm. Because the participants always know their own wage, we regress the outcomes only on the uncertain elements of the wage allocation: the beliefs about the supervisor's wage and the two coworkers' wages.

probability that the participant punished them (Panel B). Besides these, the beliefs about the supervisor's and the richer coworker's wage correlate significantly with several other punishment outcomes as well. These suggest that participants' beliefs – especially about the richer coworker's and the supervisor's wage – matter for their punishment decisions.

In addition to the punishment decisions, participants' beliefs also correlate with their choices in the workplace choice task. We regressed whether a team is chosen by the participant on the participant's own wage, their believed level of inequality among the workers[25] and their beliefs about the supervisors' wage. In Table 1.6, we show that conditional on participants' own wage, the believed inequality among workers correlates negatively with the chance that a team is going to be chosen. We also see that participants' beliefs about the wage of the supervisor does not correlate significantly with the likelihood that the team is going to be chosen.

Beliefs are inaccurate in the wage secrecy arm

After showing that participants' beliefs correlate with their punishment and workplace choice decisions, we next illustrate that participants' beliefs are inaccurate in our games and so the wage transparency treatment provides a shock[26]. We have two pieces of evidence for this. At the beginning of the experiment, we elicited participants' expected wage allocation for each player including themselves. Using these expectations in the Gini formula we calculate participants' expected inequality among the three workers. Figure 1.2 shows the histogram of this calculated variable for those participants who were in the wage secrecy arm. The histogram is concentrated on

[25]The believed level of inequality among workers is imputed using the elements of the believed wage distribution and the Gini formula.

[26]Their beliefs may be accurate in a world where the supervisors can freely select an allocation. We do not claim that their beliefs are universally inaccurate, only that they were inaccurate with respect to our wage allocations.

Table 1.5: Correlates of Beliefs and Punishment Behavior

Panel A: level of punishment

	(1) supervisor	(2) poorer cw	(3) richer cw	(4) both cws	(5) total	(6) max
supervisors' $\tilde{w}$	0.092***	-	-	-	0.032	0.064***
		0.022**	0.038***	0.060***		
	(0.021)	(0.009)	(0.010)	(0.018)	(0.029)	(0.022)
poorer cw's $\tilde{w}$	0.011	0.054**	-	0.011	0.021	-
			0.043			0.052
	(0.048)	(0.021)	(0.029)	(0.043)	(0.069)	(0.056)
richer cw's $\tilde{w}$	0.068*	0.005	0.158***	0.163***	0.231***	0.182***
	(0.039)	(0.010)	(0.027)	(0.032)	(0.053)	(0.044)
control group mean	1.505	0.490	0.727	1.217	2.722	1.977
# observations	3475	3475	3475	3475	3475	3475
# clusters	139	139	139	139	139	139
R^2	0.056	0.029	0.113	0.078	0.048	0.054

Panel B: indicator for non-zero punishment

	(1) supervisor	(2) poorer cw	(3) richer cw	(4) both cws	(5) any player
supervisor $\tilde{w}$	0.008**	-0.008***	-0.011***	-0.011***	0.002
	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)
poorer cw $\tilde{w}$	0.007	0.025***	0.008	0.002	0.006
	(0.008)	(0.006)	(0.007)	(0.007)	(0.009)
richer cw $\tilde{w}$	0.010	0.002	0.028***	0.032***	0.023***
	(0.006)	(0.003)	(0.005)	(0.005)	(0.006)
control group mean	0.340	0.192	0.248	0.269	0.445
# observations	3475	3475	3475	3475	3475
# clusters	139	139	139	139	139
R^2	0.020	0.048	0.092	0.086	0.031

Notes: This table shows how participants' beliefs about other's wages ($\tilde{w}$) correlate with their punishment decisions. The estimations use the sample in the wage secrecy arm. *cw* stands for coworker, *others* refer to the supervisor and the two coworkers, *max* is the maximum amount of punishment towards another player within round. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

the left side of the x-axis, suggesting that participants expect low inequality, and about 60% of the times they expect zero inequality among workers at the beginning of the experiment. The dots on the x-axis mark the actual wage inequalities among

Table 1.6: Correlation of Choice and Beliefs in the Sorting Task

	(1) team is chosen
own wage	0.061***
	(0.002)
believed gini worker	-1.296***
	(0.067)
supervisor believed wage	-0.000
	(0.001)
cons	0.440***
	(0.016)
# observations	2601
# clusters	287
R^2	0.143

Notes: The table shows the coefficient of a linear probability model where the outcome is whether a given team is chosen and the explanatory variables are participants beliefs about the wage distribution in the team. The sample is restricted for non-indifferent choices in the secrecy condition and does not include rounds where the participants' own wage is the same in the two teams. Item nonresponse on beliefs or choice explains that this estimation has fewer clusters than participants. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

the workers that participants saw in the punishment game. Several dots are on the right side of the x-axis, where there is little mass in the histogram. This means that participants expect more equal wage allocations relative to most of our actual wage allocations.

The previous graph is based on participants' expected wage allocation, however, because those expectations cannot be incentivized, we also show evidence for inaccurate beliefs from incentivized questions. In the wage secrecy arm we elicited participants' beliefs about the other players' wages in an incentive compatible way. On Figure 1.3 we plot the histogram of their accuracy, measured by the total absolute distance between their believed and the actual wage allocation. The histogram reveals that participants rarely have perfect beliefs, only about in 4% of the cases.

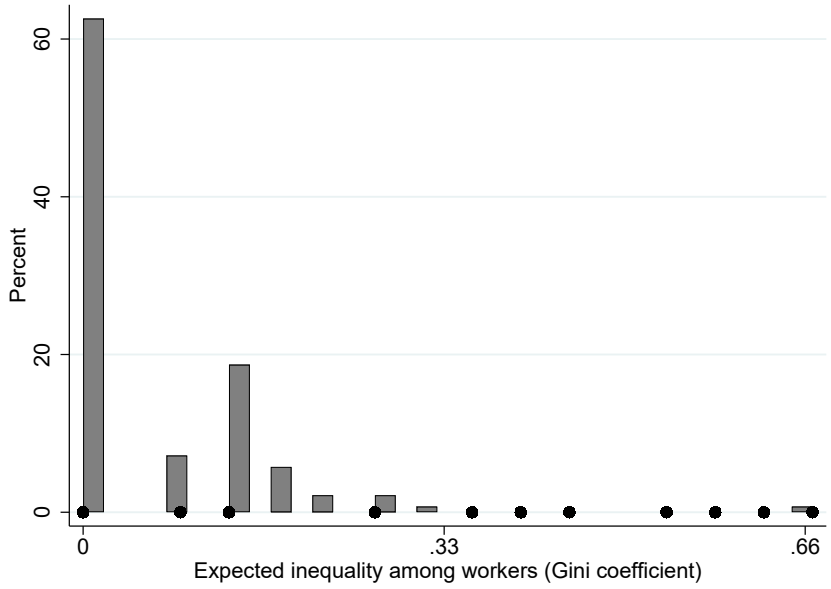


Figure 1.2: Histogram of Participants' Expected Inequality

Note: This graph shows the histogram of expected Gini coefficient among workers in the control group. The dots on the x-axis show the Gini coefficients among workers of the actual wage allocations that the agents faced. The dots overlap, we have 25 allocations in the punishment game.

Wage transparency treatment corrects inaccurate beliefs

After showing that participants' beliefs correlate with their decisions and that their beliefs were generally inaccurate, in the final step, we provide evidence that wage transparency corrects beliefs. Wage transparency mechanically corrects inaccurate beliefs because it reveals the true wage allocation in the team. However, we also have a non-mechanical check for this effect. At the beginning of the experiment all participants were asked to state their beliefs for a wage allocation. Those in the secrecy arm saw only their own payment and thus made a real guess. Those in the wage transparency arm on the other hand saw the full wage allocation and so their task was merely to copy the numbers of the wage distribution to the answer fields. While this elicitation was not incentivized, the results in Table 1.7 reassuringly indicate that the errors in beliefs are largely offset due to the treatment. The gap

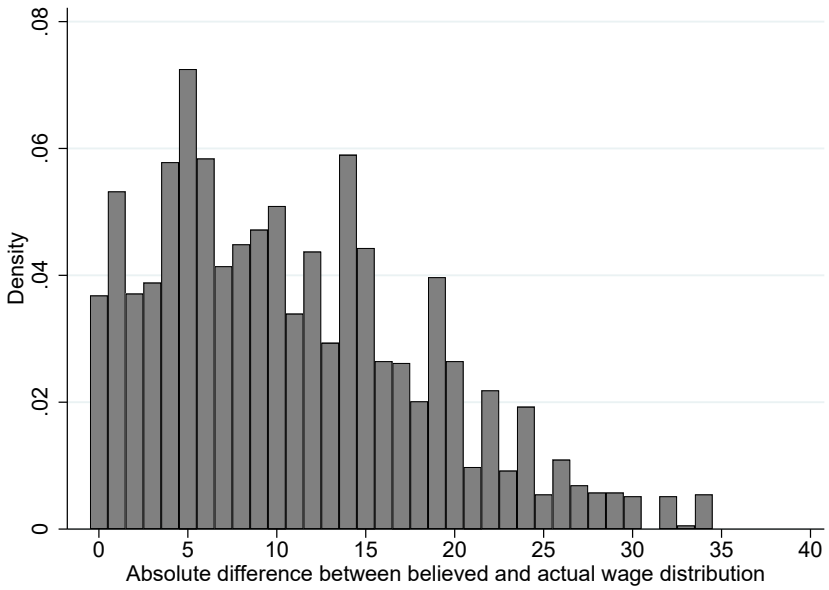


Figure 1.3: Histogram of Belief-accuracy in the Punishment Game

Notes: This graph shows the histogram of how accurate participants' beliefs were about the wage allocation. The sample includes all 25 rounds of the punishment game for those in the wage secrecy arm. In each round, we elicited participants' beliefs about the three other players' wages. We take the absolute distance of these three beliefs and their actual value. 0 absolute difference means that the participant in a given round had perfect beliefs.

is not perfectly offset though, which we interpret as errors in copying numbers or disbelief in the instructions.

We argued in this section that the treatment effect can possibly exist because of beliefs. We showed that participants' beliefs correlate with their decisions, these beliefs are inaccurate and they are corrected in the treatment arm. Our theory suggests that the second, simultaneous mechanism behind the treatment effects is the mechanism of social preferences. We next test if social preferences can be detected in our sample.

Table 1.7: Treatment Effect on Accuracy of Beliefs

	Accuracy on workers' wages		Accuracy on all members' wages	
	abs difference	Δ Gini	abs difference	Δ Gini
treat	-2.781***	-0.071***	-10.299***	-0.057***
	(0.278)	(0.004)	(0.625)	(0.006)
control group mean	3.439	0.081	11.525	0.075
# observations	294	293	294	293
# clusters	294	293	294	293
R^2	0.238	0.499	0.483	0.254

Notes: The table shows the effect of treatment on accuracy of beliefs. We use one round where where we asked the beliefs of all participants about the other players wage. *abs difference* is the sum of absolute differences between participants' beliefs about the other players' wage and their respective actual payment. Δ *gini* is the difference between the imputed believed and actual Gini coefficient, where we calculate the believed Gini coefficient based on the believed wages of the other players and the participant's own payment. The first two columns consider only the three workers, while the last two columns consider all team members, including the supervisor. The Gini coefficient is maximum $\frac{3}{4}$, when we calculate it over the entire team. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

Social preferences

Earlier we showed that wage transparency has a treatment effect on punishment and sorting decisions, and we provided empirical evidence for beliefs as a possible mechanism. In this section we show that the other mechanism identified in on our theoretical framework, the mechanism of social preferences, is also supported by our data. We begin by describing the social preferences that evoke to explain the decisions in the punishment game. Then we categorize participants to different social preference categories. The categorization reveals that social preferences can be detected in our sample, providing empirical evidence for the channel of social preferences.

Description of social preference types

The other-regarding preference segment of the utility function, $g(\cdot)$, describes how a given agent incorporates the wage of another player to their utility function. Nu-

merous papers have focused on what these preference types are and how to formalize them.[27] These functional forms are important, because as seen in Equation 1.3, the marginal utility of the other player's wage depends on not just the weight that the participant places on their other-regarding segment of the utility function, but also on the partial derivative of $g(\cdot)$ with respect to the other players' wage. Thus, different functional forms of the $g(\cdot)$ function can lead to different optimal decisions that can may lead to heterogenous treatment effects by social preference types.

We borrow some of the most commonly used other-regarding preference notions from the literature and consider whether the participants' decisions are consistent with those notions. We limit our attention to inequality based, reciprocal, rank-based and joy of punishment preferences because these preference types have the potential to explain non-zero punishment decisions and consequently the treatment effects of wage transparency on punishment. We list these preferences along with examples in the literature in Table 1.8.

The first notion we describe is the inequality based social preference. The utility function of this type of agents depends on inequality positively (inequality lover types) or negatively (inequality averse types).[28] Therefore, in our punishment game we expect inequality lover type participants to punish to increase inequality. This can typically be done by punishing the poorest individual. On the other hand, we expect that the punishment decisions of inequality averse type participants will result in lower inequality. Inequality can be typically reduced by subtracting payment from the richest individual.[29] As the optimal punishment decisions of both of these types

[27]For a review see Cooper and Kagel (2016) and Fehr and Schmidt (2006). The review by Rotemberg (2006) surveys social preferences specifically at the workplace.

[28]In the literature Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) formalized utility functions for these preference types.

[29]We describe typical decisions here, because these actions are not always possible. Participants cannot punish themselves, so in cases when the participant is the richest individual and has inequality

depend on the (believed) wage allocation, wage transparency may reveal information to the participants that would change their optimal decisions. This means that the presence of inequality based types could explain not just the positive amount of punishment in the punishment game but also the treatment effects of wage transparency.

Next are the reciprocal type agents. Following Rabin (1993) these agents reciprocate kind and spiteful actions. Because only the supervisors make a decision that affects the agent, reciprocal types would reciprocate only towards the supervisor, never the coworkers – a fact that we heavily use when we are categorizing participants to types based on their punishment decision. Whether a reciprocal agent decides to punish or not, depends on the agent's decision rule that determines if the supervisor was kind or not. To the extent that the (believed) wage allocations enter this decision rule, the reciprocal agents' optimal choice may also be shifted by the treatment, and hence reciprocal agents may also be responsible for the observed treatment effects.

Rank-based type agents are in our next group of social preference category. These agents care about their rank to some degree in the wage allocation. For example, Kuziemko et al. (2014) propose that people may be averse to be the last in a group, but we can also think of other, more general preferences that would motivate people to be ahead of others in general. We expect rank-based type agents to punish only if they can improve their ranks[30]. We also predict that when these type of agents punish, they would target the player who is the closest richer player to the agent. The wage allocation thus helps rank-based type agents to determine if increasing their rank is feasible and what is the best way to achieve it. Because the wage allocation drives the decision of this type, wage transparency can modify their behavior leading to

averse preferences, they cannot punish the richest individual. The same is true in the opposite combination: inequality lover type participants cannot punish the poorest individual if they themselves are the poorest.

[30]This condition can be limiting as agents can punish at most 10 ECU in a round, thus they cannot get ahead of any player who is more than 10 ECU richer than them.

treatment effects.

The joy of destruction type agents have utility functions that positively depend on punishment(Abbink and Sadrieh, 2009; Zizzo and Oswald, 2001). These agents would punish prospectively whether they know the wage allocation or not. This means that while joy of destruction type agents may be responsible for non-zero punishment, they are not responsible for the treatment effects of wage transparency.

Many other conventional other-regarding preference types would never punish in the punishment game. For example, the selfish types would not punish because for them punishment only has cost and no benefits. We expect the social welfare maximizer and altruist types not to punish, because punishment is inconsistent with their motivations. Social welfare maximizers would increase the share of the pie, yet this game only allows them to decrease the total welfare through punishment, so their optimal decision is never to punish. Those who have altruistic preferences would like to aid others potentially even by sacrificing themselves. Aid, however, is not a feasible action in our game, leading to an optimal choice of zero punishment. All of these types – the selfish, the welfare maximizers and the altruists – would always choose not to punish irrespectively on which treatment arm they are in. This means that these types are observationally equivalent in the punishment game and they cannot explain the treatment effects of wage transparency on punishment decisions.

Proxying participants' social preference types

In what follows, we proxy participants' social preference types based on their punishment behavior. This categorization is needed, because as we highlighted in the previous section, not all social preference types can explain punishment behavior, and even among those that do, not all of them can be responsible for the treatment

effects we see. In particular, the joy of punishment types may engage in punishment, but the treatment has no effect on their optimal behavior. To support the mechanism of social preferences behind the treatment effects, we need to find evidence that our sample contains types that can respond to the treatment, a combination of inequality averse, inequality lover, reciprocal and rank-based types.

Our categorization uses five pre-specified wage allocations in the punishment game. For each of these wage allocations we made predictions about which social preference types would punish given the wage allocation and if they would punish, who would be their targeted player.[31] If a participant punished anyone, we considered that decision consistent with the joy of punishment preferences. Similarly if the participant did not punish in a round we took it as a sign for belonging to the never-punisher group. We chose the five allocations in a way that the other preference types (inequality averse, inequality lover, reciprocal and rank-based) were predicted to punish in some but not all rounds, so we could separate these types from those who would always punish (joy of punishment) and never punish (never punishes).

We also use our predictions about whom these types would punish in the categorization. We predicted that reciprocal types would only punish the supervisor and never the coworkers, meaning that coworker punishment is never consistent with reciprocal types. On the other hand, coworker punishment is consistent with inequality based and rank based types.

Finally we also make use of the constraints of our game. Suppose that the nearest player ahead of the agent is more than 10 ECU away. For the agent who wishes to increase their rank, punishment is not optimal because the maximum amount of punishment in a team is 10 ECU which would not be enough to increase their rank. On the other hand, an inequality averse or a reciprocal agent may find it optimal

[31]The categorization steps, the used wage allocations and the respective predictions were pre-specified.

to punish the player ahead to decrease inequality or reciprocate spiteful behavior. Accordingly, seeing punishment in this scenario would not be consistent with rank-based preferences, but would be consistent would inequality averse and reciprocal types (in addition to joy of punishment preferences).

In the next step, the predictions we made for the five selected rounds were compared to the actual punishment decisions of the participants in the wage transparency arm[32]. We assigned the most likely type for each participant based on their decisions and excluded participants from the categorization who could be assigned to more types[33].

This categorization has several assumptions. First we need to assume that the agents have only one type and their type is time invariant. In addition, we assume that the cost of punishment did not deter punishment on the extensive margin. This assumption is needed because if the cost of punishment deters participants whose types are consistent with punishment, we would mistakenly label them as never punishers. We consider these assumptions, especially the first two, as strong assumptions. Their benefit though is that they ensure that we can pool the signals for preference types across the five allocations, which further allows us to categorize participants. Nevertheless, we look at the categories only as proxies to participants' true types.

Results of the social preference categorization

Table 1.8 shows the distribution of social preferences types that we find with our categorization method along with some short descriptions. We find that about 36% of the

[32]We used the wage transparency arm to ensure that participants have the true information about the wage allocation. Participants in the secrecy condition may had inaccurate beliefs and thus typically did not have the allocation in mind that we made the predictions for. Random assignment of the treatment ensures that the distribution of types in the treatment arm should be close to that of the entire sample.

[33]This could happen if a participant was equally likely to have two or more types.

participants are inequality averse, 9% are reciprocal, about 24% are joy of punishment types and 30% make up the never-punisher group. We identify very few rank-based types and no participant whose decisions are consistent with inequality lover preferences. This means that among the four types that are potentially responsible for the treatment effect, we could find evidence for only two in our sample: inequality averse and reciprocal types. We conclude that the social preference channel is supported empirically and that our treatment effects in the punishment game are most likely due to inequality averse and reciprocal agents.

We also argued that social preferences have a role in the workplace choice task as well. However, the predictions about which types could possibly respond to the treatment by changing their preferred team are less clear for this task. In theory all types could reverse their choice when they see the full wage allocations, except the selfish types. For this reason we are using this task to provide further evidence for social preferences and bound the share of selfish types.

In some rounds of the workplace choice task, the participant's proposed wage was the same across the two offers. In these cases, participants cannot distinguish between the two offers when the other players' wages are secret. However, when the entire wage allocations were revealed for these rounds, 65-83% of the participants chose one or the other team instead of claiming that they were indifferent. Among those who said they are indifferent, 20 participants (6.8% of the sample) indicated that they are indifferent in all of these rounds and thus could be categorized as selfish. Consequently, we treat the 6.8% as the upper bound of the share of selfish types in our sample.

To summarize our findings, we find that at most 6.8% of our participants are selfish types in our sample and the other participants have some form of social preferences that we detected either by their punishment or sorting decisions. We have sugges-

tive evidence that the treatment effects in the punishment game is mostly driven by reciprocal and inequality averse type participants, while in the sorting game any social preference type could flip their choice due to the treatment. This means that we found support of the second theoretical channel as well: social preferences can explain our treatment effects both in the punishment and the workplace choice game.

1.6 Discussion

Both of our games are abstract. This raises the question about what types of workplaces can speak to and how could we interpret the punishment and sorting decisions in those environments.

First, we argue that while the games are abstract, certain pockets of the labor market operate under circumstances that generally match the conditions in our games. On the free-lancing platforms employers and workers match quickly for short term online tasks. On these platforms workers and employers do not interact face-to-face, only online. It is also possible to achieve anonymity through the use of usernames. Workers even can be assigned to teams. As examples for such tasks, consider programming a website by two programmers as in Lyons (2017) or an on-demand data entry task as in Huet-Vaughn (2015). Our game is in between these two examples: there is no team work element in our game unlike in the programming task, but people are assigned to teams unlike in Huet-Vaughn (2015)'s work.

Temporary in person jobs also provide a useful benchmark for our games. The call center of Flory et al. (2016) required short term work (few hours), where workers worked individually along with several other coworkers. These jobs match the in-between case of team assignment and individual work and the short term nature of our game. The face-to-face interactions, however, move these jobs away from our

setting. Thus we consider online call centers as a closer alternative than the in-person ones.

The abstractions in our games limit the work settings that we can speak to, but they come with important benefits. First, we have tight control over both wage transparency and the individual wage allocations, which allows us to show the effect of wage transparency as a function of the underlying wage allocations. Second, the abstractions also help us to argue more convincingly about the mechanisms in our games. For instance, we can rule out mechanisms that would stem from in-person interactions and would be unknown to the econometrician.

After providing examples of labor markets where several of our abstractions are present and detailing the benefits of the abstraction, we discuss the interpretation of punishment. It is unrealistic that a worker will be able to decrease the wage of their supervisor or their coworkers directly. Instead of interpreting punishment literally, we think of punishment as any behavior that can reduce the (psychological) utility of the others. Social exclusion, shirking[34], sabotage, gossiping about a team member or refusing to cooperate with them are all real world examples for what punishment in our game can stand for[35]. These types of workplace misbehavior are documented in the literature: Flory et al. (2016) present evidence that their temporary call center workers actively sabotaged each other, and Belot and Schröder (2013) provide evidence for theft, misreporting and mistakes that could be costly for the supervisors.

In the case of the sorting task, we interpret it from the perspective of a job seeker.

[34]Shirking has scope to lead to trouble for others especially in team settings, where individual outputs are difficult to quantify but the success of others also depend on the project. If a person free-rides, others will need to complete their task causing them to do overtime, canceled dinner plans and thus lower utility. For this reason shirking itself can be a way of punishment.

[35]Note that the example work types we gave earlier – coding and clerical task on a free-lancing platform and the call centers – are all workplaces where productivity spillovers between workers are possible and hence punishment is possible.

Importantly, we cannot interpret these decisions from the perspective of a worker who is already matched to a team, because our experimental instructions did not include language that the participant is currently matched to a workplace and if so which one. For this reason, the sorting game is a test for the effects of the policy on the extensive margin: what would happen if the wages become transparency when the a workseeker searches for jobs and chooses among offers. On the other hand, the punishment game measures the effect of the policy at the intensive margin, that is for those, who are already matched to a workplace.

1.7 Conclusion

This paper shows that wage transparency can have unintended consequences on workplace hostility and on workers' workplace choice. Both of these effects stem from the theoretical prediction of social preferences under two conditions: people have inaccurate beliefs about the wage allocation when wages are secret and people have social preferences. We find empirical evidence for both of these mechanisms.

Our most surprising finding is the presence of peer punishment. Through the lens of Bartling and Fischbacher (2012) punishment should follow the responsible actor. The peer punishment we document is not consistent with this idea, as the coworkers were idle by-standers in the game. To reconcile the peer punishment pattern, some form of inequality aversion is needed.

Documenting peer punishment has important consequences to policy as it shows that the agents could partially undo the unequal allocations of the supervisor-like figure. This may be normatively desirable in case of favoritism, but may be normatively undesirable in case of affirmative action policies that allocate resources unequally to overcome previous inequalities. Finding ways to mitigate or increase peer punishment

would be an interesting avenue for future research.

The broader message of the paper is that labor economics questions can be studied through online experiments. The tight control of these experiments may provide important insights that would not be possible with observational data or even field experimental settings. The cost of this precision comes at the external validity front; we can speak only to those settings that share similarities with our design: free-lancing and temporary work arrangements. However, some of our findings are corroborated by field and observational evidence, increasing our confidence that our results are plausible and may be relevant in other settings, too. Cullen and Perez-Truglia (2018a) find in a large bank that workers' beliefs are imperfect about their peers wages, and also point out that workers treat horizontal and vertical wage information in hierarchies differently. Breza et al. (2017) show that wage inequality breaks down cooperation, which is in line with our broader interpretation of hostility. Finally, Falk et al. (2018) provide evidence on social preferences, and in particular on retaliatory preferences, from several countries around the world. These pieces of field evidence are compatible with our findings.

The second main point of the paper is that ideas from behavioral economics can be applied to labor economics. These behavioral effects may seem small and relevant only at the margin[36], but in a general equilibrium framework they could still have meaningful effects if other self interested players take the presence of behavioral agents into account.

[36]For example the punishment amounts are small in the game and less than the majority of the workers switch their preferred team

Table 1.8: Distribution of Estimated Social Preference Types

Type	Motivation	Examples in the literature	Predicted punishment behavior	Wage transparency may change behavior?	N	Estimated share
inequality averse	reduce inequality	Fehr and Schmidt (1999) Bolton and Ockenfels (2000)	Punish the player who is richer than the agent	yes	44	36.07%
inequality lover	increase inequality		Punish the player who is poorer than the agent	yes	0	0.00%
reciprocal	reciprocate kindness and hostility towards the decision maker	Rabin (1993) Falk and Fischbacher (2006) Levine (1998)	Punish the supervisor if deemed unkind	yes	11	9.02%
rank based	get ahead of the other players	Kuziemko et al. (2014)	Punish the player who is right ahead of the agent	yes	2	1.64%
joy of punishment	burn money of the others	Abbink and Sadrieh (2009) Zizzo and Oswald (2001)	Punish anyone	no	29	23.77%
never punisher			Never punish	no	36	29.51%
Total					122	100.00%

Notes: This table shows the distribution of the estimated social preference types. The categorization was carried out only for those participants who were in the treatment condition in the punishment task. Among these, the sample excludes participants who could be categorized to multiple types, which explains the smaller sample size.

Chapter 2

The Black-White Recognition Gap in Award Nominations (with Nayoung Rim, Roman Rivera and Bocar Ba)

2.1 Introduction[1]

For decades, a goal of public policy has been to reduce racial disparities in the labor market.[2] The economics literature has largely focused on firms' hiring decisions because of the ability to experimentally examine hiring (Bertrand and Mullainathan, 2004; Kessler et al., 2019; Neumark et al., 2019). Less is known about racial bias in career recognition and progression, which may arguably be more important for the lack of diversity in upper-management positions and, ultimately, the racial wage gap.

An important question for eliminating discrimination and racial gaps in career outcomes is whether supervisors choose to engage with and acquire information about minority colleagues. We examine this question in the context of the second largest police department in the US, where supervisors do not necessarily observe the officer's day-to-day activities but are required to evaluate the officer's performance annually. Because supervisors do not directly monitor officers, they must exert effort to gather

[1]This chapter is co-authored with Nayoung Rim (US Naval Academy), Roman Rivera (Columbia University) and Bocar Ba (University of California - Irvine). Kiss was responsible for the experiment. We thank the Duke Economics Department, the U.S. Naval Academy, and the Quattrone Center at Penn Law School for generous financial support. IRB approval for this experiment was received from the U.S. Naval Academy (#USNA.2019.0040-PADS) and Duke University (#2020-0338). The experiment was pre-registered in the AEA RCT Registry, AEARCTR-0005929. The views expressed herein do not necessarily reflect the position of the Chicago Police Department or the U.S. Naval Academy.

[2]In regards to understanding the source of discrimination in the labor market, the economics literature has coalesced around two main explanations: taste-based discrimination and belief-based or statistical discrimination. See Lang and Kahn-Lang Spitzer (2020) for a survey of the literature.

information on officers when it comes to the annual evaluation. If the cost of acquiring information differs by race, a racial gap in career recognition and progression may result. Although our application focuses on law enforcement, this organizational structure (autonomous workers operating within a hierarchical organization) is common across all industries.

We construct a novel panel dataset of all Chicago Police Department (CPD) officers between 2009 and 2015 containing detailed personnel information on use of force, arrests, and misconduct—crucial information in an empirical study of bias in the workplace. Using supervisor nominations for departmental awards, we examine whether white supervisors are less likely to acquire information about and nominate their minority officers. We focus on award nominations rather than wage and promotion because nominations are subjective evaluations of officers' performance. In contrast, wages in the CPD vary only by experience due to a union contract, and promotions are largely determined by a written test.[3] Further, awards are an important measure of career recognition and are used in important decisions related to career advancement, such as performance evaluations, merit promotions, and overtime pay.

Our identification strategy exploits two institutional features of the CPD that allows us to obtain plausibly causal estimates of the black-white recognition gap. First, officers are assigned a new supervisor every January, which we use to approximate random assignment of an officer's race to a supervisor. We confirm as-good-as-random assignment by analyzing supervisor-officer assignments and confirm that officers do not sort to supervisors based on work performance measures.[4] Second, all officers

[3]Seventy percent of promotions to sergeant are determined by a written exam, while 30 percent may be based on nominations by higher-ranking officers. These nominations take into consideration an officer's qualifications, such as the number of awards received. Further, promotions are rare because the sergeant exam is not offered on a regular basis, thereby limiting the opportunities for police officers to be promoted.

[4]In particular, we may be concerned that more-productive white officers and/or less-productive black officers sort to white supervisors. In this case, we would see a negative black-white nomination

must be evaluated annually by their supervisor, and the quarter of evaluation is randomized across officers.[5] Under the assumption that supervisors are more likely to engage with and gather information about officers in the evaluation period, this institutional feature allows us to exploit the randomly assigned evaluation quarter and estimate the causal impact of an interaction, which would normally be endogenous.

We find that supervisors are more likely to nominate all officers in the quarter of evaluation relative to two quarters prior, suggesting that statistical discrimination may exist (Altonji and Pierret, 2001).[6] However, black officers are nominated 35.4 to 42.7 percent less than white officers, suggesting that statistical discrimination is not the only explanation for this racial disparity. The fact that the nomination likelihood increases for all officers in the evaluation quarter suggests that white supervisors may be learning about their officers, but the persistent, negative black-white disparity suggests that the learning is not manifesting in changed behavior towards black officers. By contrast, nomination patterns for Hispanic officers are similar to those for white officers. These results suggest that the black-white gap is not due to in-group favoritism towards white officers but rather bias against black officers. This is supported by additional analysis that finds that the negative black-white gap among white supervisors widens with the number of arrests. This is the opposite of what we would expect if the disparity were due to white supervisors' inaccurate beliefs about black officers' work measures.

Because the administrative CPD data do not capture detailed interactions between supervisor and officer, we conduct an online experiment to measure the review process in the nomination decision. We ask Amazon Mechanical Turk (MTurk) work-

gap among white supervisors even in the absence of racial bias.

[5]The evaluation must be held in the quarter prior to the quarter that the officer joined the CPD, and the quarter in which officers join CPD is determined by lottery number.

[6]Altonji and Pierret (2001) argue that as firms learn more about their employees, statistical discrimination should decrease.

ers to evaluate officer profiles and nominate one for an award within a time limit. In addition to being able to experimentally examine whether evaluators choose to engage with minority officers, the online experiment allows us to generalize our findings to a broader evaluator group than Chicago police supervisors.

In one task, evaluators choose between a black officer and a non-black officer. Although officer performance levels are randomly chosen, evaluators are 5.6 to 8.8 percentage points less likely to nominate black officers over white officers. In another task, officer profiles display only demographic information and evaluators must mouse over the profile to reveal full information about the given officer. We monitor mouse movements across the screen and find that black profiles are least likely to be moused over. The black-white difference in engagement more than doubles (from -2.8 percentage points to -7.3 percentages points) when evaluators are choosing among three white officers and one black officer. We do not see similar patterns for Hispanic officers when evaluators are choosing among three white officers and one Hispanic officer. When the officer pool becomes more racially diverse, any benefits black officers had received from greater evaluator engagement disappear. Taken together, our findings suggest that racial issues in policing are not just at issue between police and the public, but also within departments, and thus that simply hiring minority officers may be limited in its efficacy.

Our paper relates to the literature on social networks in the workplace. Prior research documents the importance of gender homophily on career outcomes (Sarsons, 2019; Cullen and Perez-Truglia, 2020; Zeltzer, 2020). This paper expands the literature by examining the importance of race homophily in career recognition. For example, we find that CPD supervisors and MTurk evaluators, both groups that are mostly white, are less likely to gather information on minority officers, leading to a racial disparity in award nominations. These findings are consistent with Bartoš

et al. (2016), which finds that employers are less likely to open and read resumes from minority candidates. Our findings also speak to the literature that documents benefits of same-race matching on racially disparate outcomes. There is mounting evidence that race-matching leads to better outcomes, such as in education (Gershenson et al., 2018; Carrell et al., 2010; Kofoed and mcGovney, 2019) and health (Alsan et al., 2019), but less is known about why. Our results suggest that networks formed through race homophily are important for success in the workplace.

Our paper is also similar to other papers that find that discrimination may arise because biased managers interact less with minorities (Glover et al., 2017). Our experimental evidence finds that evaluators are less likely to engage with black officers, particularly when black officers are in a pool with three white officers. Additionally, evaluators spend more time evaluating black profiles but are not more likely to nominate them. These findings are consistent with studies that find that minorities are less likely to be acknowledged for their work (Hengel, 2019; Sarsons, 2020) and a strand of literature that establishes the existence of bias among managers and work colleagues (Bertrand and Mullainathan, 2004; Giuliano et al., 2009; Glover et al., 2017; Egan et al., 2018; Sarsons, 2019; Bohren et al., 2019).[7] By analyzing the black-white recognition gap among police officers, our paper links this literature to studies on racial disparities in law enforcement.

With respect to law enforcement, our study adds to the growing research that is uncovering racial bias in policing.[8] Prior studies largely use data on officer-initiated encounters, which may be biased because they do not include the universe of all

[7]Hengel (2019) and Sarsons (2020) find that female minorities are less likely to be acknowledged for their work.

[8]See, for example, Ajilore and Shirey (2017); Cunningham and Gillezeau (2018); Horrace and Rohlin (2016); Mason (2007); Close and Mason (2006); Knowles et al. (2001); Anwar and Fang (2006); Antonovics and Knight (2009); Nix et al. (2017); Goncalves and Mello (2020); Bacher-Hicks and de la Campa (2020); Hoekstra and Sloan (2020); Weisburst (2018); West (2018); Rim et al. (2020).

possible police interactions (Knox et al., 2020). Recent papers have sought to use officer dispatches to 911 calls or to investigate automobile crashes to address this issue (Weisburst, 2018; West, 2018; Hoekstra and Sloan, 2020). Our paper attempts to bypass the truncated data problem by focusing on supervisor nominations of as-good-as-randomly assigned officers. Importantly, we ask whether racial bias on the part of officers carries over to their colleagues, a question that was previously unanswered due to a lack of detailed personnel data.

We begin the rest of the paper with a short description of CPD's organizational structure and the awards nomination process (Section 2.2). Section 2.3 describes our data collection efforts and presents summary statistics on our CPD analysis sample. Section 2.4 discusses the identifying assumptions. We present results using administrative CPD data in Section 2.5 and the experimental evidence in Section 2.6. We conclude with a discussion of the policy implications for law enforcement agencies in Section 2.7.

2.2 Background

2.2.1 Basic Facts about CPD's Structure

After passing a written exam, all Chicago Police Department candidates are placed on an eligibility list according to a randomly assigned lottery number and called off in lottery order to enroll in police academy. Upon graduation from Police Academy, Police Officers begin their career in one of the 25 geographic districts spanning the city of Chicago.[9] These initial assignments are generally outside the officer's control, with the exception of a small number of officers who received academic and other

[9]Between 2012-2014, three districts were dissolved leaving 22 geographic districts.

distinctions in the Academy (Police Accountability Task Force, 2016).[10]

Police Officers are supervised by Sergeants in their district. Daily responsibilities for sergeants include participating in roll call, supervising criminal investigations (e.g., protecting the scene, establishing the perimeter), and ensuring officers carry out their responsibilities.[11] Every year, sergeants conduct performance evaluations of their assigned supervisees.[12] To assist supervisors with performance evaluations, an electronic database called the Performance Recognition System (PRS) tracks exceptional or adverse behavior related to job performance. Information is entered by Human Resources staff, and supervisors have the ability to monitor and track information in PRS, though it is uncertain whether any actually do in practice (U.S. Department of Justice, 2017, p. 111-112).

CPD patrol officers work on a rotational schedule, where they rotate their off-days each week. Therefore, officers are not necessarily assigned to work on the same days as their supervisors who conduct the annual performance evaluation (U.S. Department of Justice, 2017, p. 108).

2.2.2 CPD Awards Nomination Process

The Chicago Police Department distributes department awards to recognize the accomplishments, performance, and service of its Department members. In addition to highlighting officers' accomplishments, awards are used in important decisions related

[10]When vacancies occur, officers may bid for district transfers. Successful bidders are chosen based on their qualifications and seniority.

[11]Section III.A., Employee Resource E05-05, available at `http://directives.chicagopolice.org` and Appendix A, CPD Sergeant Written Assessment Study Briefing 2013, available at `https://www.chicago.gov/content/dam/city/depts/dhr/general/CPD_Sergeant_Assessment_Study_Briefing_2013.pdf`.

[12]The average supervisor conducts 7.8 evaluations a year. The median number of evaluations (supervisees) is seven.

to career advancement, such as performance evaluations[13], merit promotions[14], and overtime pay[15].

We focus on award nominations in this paper because they are subjective evaluations of officers' performance.[16] The importance of subjective evaluations is underscored in the 2016 report by the Police Accountability Task Force, which states that—despite the test-based promotional process in the CPD—98 percent of CPD officers felt promotions were due to connections instead of merit. This sentiment is consistent with a growing literature that documents the importance of mentoring and networks in the workplace (Beaman et al., 2018; Cullen and Perez-Truglia, 2020).

More practically, we focus on award nominations because of institutional facts that restrict our study of wage and promotion in the CPD. Officer base salaries are set by a pay schedule determined solely by experience. Therefore, there is no racial disparity in pay by construction. Additionally, promotions are rare in the CPD because they depend on the number of vacancies, which occur when a higher-ranking officer retires or dies, and because the sergeant exam is not offered on a regular basis. For example, almost ten years passed between the two most recent sergeant promotion exams (Police Accountability Task Force, 2016, p. 140). As it is difficult to study racial disparities in career progression with these traditional measures, one contribution of this paper is its access to data on award nominations.

[13]Chicago Police Department, Career Development Directive, Employee Resource E05-01, Section IV.H., available at `http://directives.chicagopolice.org`.

[14]Merit promotions are a more subjective selection process that relies on a variety of officer qualifications vis-a-vis test-based promotions. See, Section III.E.2, Employee Resource E05-05, available at `http://directives.chicagopolice.org`.

[15]The number of awards is a statistically significant predictor of overtime pay. An additional award last year is correlated with an additional $206.78 in overtime pay this year, which is more than four times the estimate for an additional arrest last year ($46.99). These estimates control for years of experience.

[16]We choose not to examine award receipt because those are determined by an external Awards Committee, which may add an additional layer of bias in the decision process.

There are 33 departmental awards, which range in their competitiveness. Most awards require a nomination process. Nominations may originate from any higher-ranking officer, including one's supervisor.[17] Our analysis focuses on nominations by officially assigned supervisors to leverage the quasi-random assignment and the annual evaluation requirement.

Officers may be nominated for a single award per incident, and nominations must be submitted within 45 days of the incident.[18] There is no restriction on the number of times an officer may be nominated, as long as the nominations are for different incidents. Supervisors are also not restricted in the number of award nominations they are allowed to submit.

2.3 Data

This section describes administrative police records and district-level crime information that are used for the empirical analysis. We first describe the data sources and the linked analysis dataset. Then, we provide descriptive statistics of Police Officers in the Chicago Police Department between 2009 and 2015.

2.3.1 Police Officer Data

Administrative records and information on sworn Chicago Police Department members were obtained by Freedom of Information Act requests through a collaboration with Invisible Institute. In order to connect different datasets, officers are first

[17]Nearly 90 percent of nominations for police officers are from sergeants. Thirteen percent of all award nominations are from an officer's assigned supervisor.

[18]There are a few exceptions to this. The Carter Harrison/Lambert Tree Medal, 100 Club of Chicago Valor Award, Superintendent's Award of Valor, Police Blue Star Award, and Police Blue Shield Award may be awarded to officers who received other departmental awards for the same incident (Chicago Police Department, Department Organization Directive, Special Order S01-01, available at `http://directives.chicagopolice.org`).

identified within a dataset using the available unique characteristics, such as name, appointed date, birth year, and race, and then matched with identified officers in different datasets.

Demographics Data on officer race, sex, birth year, and appointment date are obtained from aggregated data, using the most common observation across datasets.[19] Officer rank is taken from salary data provided by the Chicago Department of Human Resources (DHR), covering 2002 to 2017.

Supervisors This dataset provides information about the supervisor who conducted each officer's annual evaluation between 2009 and 2017. Our analysis focuses on those at the rank of Police Officer, meaning their supervisors are at the rank of Sergeant. In this paper, the term "supervisor" refers to a Sergeant who is officially assigned to conduct a Police Officer's annual evaluation in a given calendar year.

Awards The awards dataset provides information on all department award nominations between 2004 and 2017. The dataset includes the award name, the individual being nominated, the requester, request date, and the final status of the nomination (approved, deleted, or denied).[20] We consider all performance awards that are open to all sworn Department members and require a supervisor's nomination.[21] After these restrictions, our analysis considers 18 awards. Appendix Table 3.10 provides a description of these awards.

Unit Assignment Historical unit assignment data lists all units to which an

[19]Not all demographic information is complete in each file, so an aggregation of demographic variables across multiple files is necessary. Over 99 percent of officers are matched to a unique gender, race, and appointment date.

[20]An award may be deleted for various reasons, including: the form was not filled out correctly; supporting evidence was not included; or the nomination does not meet the eligibility requirements of the award. This differs from an award denial, which means the officer did not win the award. Very few awards (2.4 percent) are deleted.

[21]Most awards are open to all Department members. One example of an exception is the Thomas Wortham IV Military and Community Service Award, which is awarded to current or former members of the U.S. Armed Services.

officer was assigned since the beginning of his or her career, as well as start- and end-dates in each unit. We focus our analysis on Police Officers assigned to geographic patrol districts.

Arrests The arrests dataset contains information on all arrests made by Department members. The dataset includes detailed information about the subject, crime, and arrest location and time. These data cover 2001 to 2017 but arrest day and month are only provided from 2010 onwards. For arrests made in 2009, we use the date the subject was released from the local police station as a proxy for the arrest date.[22] Crimes are aggregated into three categories: violent crime, property crime, and non-index crime. The Federal Bureau of Investigation classifies violent and property crimes as "index crimes" because they are more serious offenses.[23] Non-index crimes capture crimes that are not related to violence or property, such as municipal code violations, traffic violations, warrants, drugs, prostitution, and gambling.[24]

Complaints The complaints dataset contains all recorded allegations of misconduct filed against an officer from 2000 to 2016. Allegations may originate from the public or from other officers in the department.

Tactical Response Reports Data on officer use of force come from 2004 to 2016 Tactical Response Reports (TRR). Officers are required to file a TRR if they used any force while performing their duties. A TRR filing requirement can be triggered by three things: the subject's actions; the officer's actions; or a subject who is injured or alleges injury resulting from the officer's use of force option. CPD publishes a Use of Force Model, which provides guidelines on the appropriate level of force to be used

[22] In 96.9 percent of cases, the release date is on the same day or the day after the arrest date, and 100 percent of release dates are within four days of the arrest.

[23] Violent crimes are crimes related to violence, such as murder and assault. Property crimes are crimes related to property, such as burglary and motor vehicle theft.

[24] A comprehensive list of crime categories can be found at `http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html`.

in response to a subject's actions and levels of resistance. Using the Use of Force Model as a guide, we classify officer force options into two broad categories of "weak use of force" and "strong use of force." Weak use of force includes force mitigation efforts, such as verbal direction and tactical positioning (which involve no physical touch), and control tactics, such as escort holds and wristlocks. Strong uses of force involve elevated levels of force that are generally intended to enact harm on or injure the subject.[25] The data only report use of force against adult persons. Appendix Table 3.11 outlines force options and our classification.

Sample restrictions To construct a complete dataset on all officers in the Chicago Police Department, we require that officers receive a salary from DHR and appear in the unit assignment dataset. We focus on years 2009 to 2015 to maximize overlap across the different datasets. We further restrict our sample to officers at the rank of Police Officer who are always assigned to a geographic district[26] and officer-supervisor relationships that lasted for 12 months. Our final analysis dataset has 6,518 Police Officers and 1,284 supervisors. In terms of the outcome variable, we consider nominations for 18 awards that require a supervisor's nomination and is open to all Department members.

2.3.2 Crime Data

We use crime data from the Chicago Data Portal (`https://data.cityofchicago.org`), which contains reported incidents of crime that occurred in the City of Chicago since 2001. The dataset contains the primary type of crime, the date, location, and

[25]Strong use of force may or may not use weapons. Examples of strong use of force without weapons are take-downs, kicks, and punches. Weapons are further classified into lethal and non-lethal weapons. Examples of non-lethal weapons are chemical weapons and long-range acoustic devices. Examples of lethal weapons are tasers, batons, and firearms.

[26]We remove the three districts that closed between 2012-2014 (13, 21, and 23) from our analysis sample because we do not have crime statistics for these districts.

whether the crime led to an arrest. We construct monthly crime rates for each district, separately for total crimes, property crimes, and violent crimes.[27]

2.3.3 Summary Statistics

This section provides descriptive statistics of Police Officers in our analysis sample. From Table 2.1, we see that most officers are male (73.7 percent) and white (46.4 percent), but blacks and Hispanics are also well-represented (23 to 27 percent). In fact, these three racial groups make up nearly 97 percent of our sample. The average CPD officer in our sample joined the force in 2000 at age 30. This indicates that at the start of our analysis dataset (year 2009), the average officer had been on the force for nine years.

Relative to Police Officers, the racial makeup of supervisors[28] in our analysis sample is more homogeneous. About 81 percent of supervisors are male, and 70 percent are white. Blacks and Hispanics each make up around 14 percent of supervisors. At the start of our analysis dataset, the average supervisor had worked for 17 years or eight years longer than the average Police Officer. The average supervisor has 7.3 officers to evaluate every year, and the median number is seven. The 25th percentile is three officers, and the 90th percentile is 14 officers.

Table 2.2 presents racial differences in various work measures. The first row is the probability of being nominated for an award in a particular month. For example, the average officer has a 2.5 percent chance of being nominated in a given month, which equates to about a 30 percent chance of being nominated in a given year. White and Hispanic officers have slightly higher than average likelihoods at 3 percent and 3.2

[27]Crime rate is defined as the total number of reported incidents of crime divided by the population and multiplied by 1000.

[28]Recall *supervisors* are Sergeants who are officially assigned to conduct a Police Officer's annual evaluation.

Table 2.1: Summary Statistics

	Police Officers	Supervisors
Male	73.7%	80.8%
Race		
White	46.4%	69.7%
Black	26.8%	14.7%
Hispanic	23.2%	14.0%
Asian	3.1%	1.6%
Native American	0.4%	0.1%
Birthyear	1970.3	1965.3
Start Year	2000.0	1992.2
Observations	6,518	1,284

Source: CPD analysis sample.

percent, respectively, while the likelihood for black officers is half the sample average (1.3 percent). The black-white difference is statistically significant at the 1 percent level.

The second row lists the number of monthly complaints. The average officer receives about 0.04 complaints in a given month, equating to about 1 complaint every two years. This statistic is similar across race. The third row lists the number of TRR filings, which is a proxy measure for use of force. The average officer files about 0.05 reports a month, equating to about 1.2 filings every two years. Black officers, however, files about half as many reports as white and Hispanic officers.

The remaining rows depict the number of monthly arrests by arrest type. For example, the average officer makes 1.8 arrests every month. White and Hispanic officers are slightly over this average at 2 and 2.2 arrests, respectively, while black officers are below this average at 1.2 arrests. The black-white difference equates to 10 fewer arrests a year ($p < 0.01$). When comparing summary statistics for the different types of arrests, we see that the black-white difference in total arrests is driven by arrests for non-index crimes, which make up around 65 percent of all arrests. Here,

Table 2.2: Racial Differences in Work Measures

	All Officers	White Officers	Black Officers	Hispanic Officers	B-W Difference (p-value)	H-W Difference (p-value)
Nominated	2.5%	3.0%	1.3%	3.2%	-1.7 (0.000)	0.2 (0.016)
Complaints	0.04	0.04	0.04	0.04	0.00 (0.937)	0.00 (0.075)
TRR filings	0.05	0.05	0.03	0.06	-0.02 (0.000)	0.00 (0.039)
Total Arrests	1.82	2.04	1.19	2.16	-0.85 (0.000)	0.12 (0.000)
Violent	0.37	0.37	0.31	0.42	-0.06 (0.000)	0.05 (0.000)
Property	0.27	0.29	0.20	0.30	-0.09 (0.000)	0.01 (0.017)
Non-Index	1.19	1.38	0.68	1.44	-0.69 (0.000)	0.07 (0.000)
Drug	0.31	0.37	0.14	0.41	-0.23 (0.000)	0.03 (0.000)
Traffic	0.12	0.15	0.06	0.16	-0.09 (0.000)	0.01 (0.002)
Observations	250,872	111,876	70,572	59,148		

Source: CPD analysis sample.
Notes: This table lists monthly summary statistics for 6,518 police officers. Sample is at the officer-month level. Non-index arrests include arrests for non-property and non-violent crimes. B-W Difference reports the percentage-point difference between black officers and white officers. H-W Difference reports the percentage-point difference between Hispanic officers and white officers. p-values are the p-value from a t-test of a difference in means.

the difference is about -0.7 arrests per month or 8.4 fewer arrests per year ($p < 0.01$).

Although the data reveal a disparity in number of arrests, we caution the reader from jumping to the conclusion that black officers are less productive than white and Hispanic officers. Arrests are not a comprehensive measure of policing quality and may even be a biased measure (Owens et al., 2018). For example, a study by Harvey and Mattia (2020) finds that police departments that increased their share of black officers subsequently reduced black crime victimization. Similarly, female officers have fewer arrests than male officers but Miller and Segal (2018) finds that increasing the number of female police officers decreased the number of intimate partner homicides and increased the number of reports of domestic violence in the U.S. These outcome measures, which are important measures of social welfare, are not captured by arrests nor would they appear on an officer's record.

Another example is to consider proactive arrests like drug and traffic arrests, which

are proactive in that they are more likely to have originated from an officer-initiated incident. This classification of "proactive arrests", which allow for greater officer discretion, can also be seen as a delineation between appropriate and inappropriate uses of police authority.[29] In Table 2.2, we see that white officers are about 2.4 to 2.6 times more likely than black officers to arrest someone for drugs or traffic violations. In contrast, the black-white difference for more serious crimes, like violent crimes, is economically small at -0.06 arrests a month. Relatedly, Ba et al. (2020) examine daily patrol assignments of CPD officers and find that black officers make fewer stops and arrests and use force less often than their white colleagues. This disparity is driven by a decreased focus on discretionary contact, such as stops for "suspicious behavior". These facts suggest that although it is important to control for work measures in our analysis, we should not automatically interpret differences in overall arrests as differences in policing quality.

2.4 Identifying Assumptions

This section outlines the empirical strategy to examine whether supervisors are less likely to acquire information about minority officers. We exploit two institutional features of the Chicago Police Department that allows us to estimate plausibly causal estimates of the black-white recognition gap.

First, we use the assignment to a new supervisor at the start of a calendar year to

[29] We borrow this term and classification from Worden et al. (2013). We do not know whether an arrest stemmed from an incident that the officer initiated on his or her own authority, but we assume that drug and traffic arrests are more likely to have stemmed from officer-initiated traffic stops as compared to arrests for violent crimes. Importantly, proactive arrests should be considered as a very noisy measure of quality policing. For example, Worden et al. (2013) analyzed the impact of a police agency's early intervention system, which aims at monitoring and managing police misconduct among officers who exhibit patterns of problematic behavior, and found that it lowered the number of proactive arrests with little impact on productivity.

approximate random assignment of an officer's race to a supervisor.[30] Although the vast majority of supervisor relationships last one year[31], we may be concerned that some officer-supervisor relationships may have been arranged outside of the random assignment system. Therefore, we restrict our analysis sample to all supervisor-officer relationships that last one year in order to minimize the number of endogenously formed supervisor relationships. In the next section, we empirically test whether officers are as-good-as-randomly assigned to supervisors in the data.

Second, we exploit the randomized timing of an officer's annual evaluation. All supervisors are required to conduct annual evaluations of their assigned officers, and this evaluation must take place during the quarter prior to the quarter in which the officer joined the Department. Because start dates are determined by a lottery number, this means that the evaluation quarter is essentially randomly assigned across officers.[32] Appendix Table 3.12 lists the evaluation quarters and evaluation due dates by start month. For example, if an officer started his career in July (Q3), then his annual evaluation must take place in the second quarter of every calendar year (Q2).

The randomized timing of the annual evaluation in combination with the new supervisor assignment every January allows us to estimate plausibly causal estimates of the black-white nomination gap and how they evolve as supervisors learn more information about their officers due to the annual evaluation.

[30] About 96 percent of officers are assigned to a supervisor in January of each calendar year.

[31] Of all supervisor assignments between 2009 and 2015, 89 percent last exactly one year.

[32] After passing a written exam, all CPD candidates are placed on a eligibility list according to a randomly assigned lottery number and called off in lottery order to enroll in the police academy.

2.4.1 Exogeneity of Supervisor Assignment and Officer Performance

Throughout the paper, we want to interpret any change in nomination likelihood when white supervisors are assigned white officers relative to when they are assigned black officers as a causal effect of officer race. The key assumption is that minority officers were not systematically assigned to white supervisors in years when officer performance would have been particularly low for other reasons. For example, if high-performing white officers and low-performing black officers sort to white supervisors, then we would see a negative black-white nomination gap. This may appear to be bias against black officers by white supervisors, but in reality it would be the result of sorting of police officers based on work performance measures. We argue that this sorting concern is mitigated in our setting due to the as-good-as-random assignment of supervisors every January.[33]

One way to examine the validity of this assumption is to test whether officers of different races are differentially likely to be assigned to a supervisor of a given race. If there is no supervisor-sorting, then we would expect white officers assigned to white supervisors to look similar to white officers assigned to black supervisors, and similarly for black officers.

Table 2.3 reports average lagged annual work performance measures by officer race and supervisor race. Panel A lists mean lagged annual measures for white officers assigned to white supervisors in column 1, to black supervisors in column 2, and the p-value from a t-test of a difference in means after controlling for officer birth year, experience, unit, year fixed effects, and unit-year fixed effects. Panel B lists the same for black officers, and Panel C lists the same for Hispanic officers. Because officers are assigned to supervisors at the Department level, we use all patrol officers assigned

[33]Further, we are able to control for officer work measures.

to a supervisor rather than the analysis sample that is restricted to officers whose supervisor assignment lasted one year.

Table 2.3: Officer Work Measures and Supervisor Race, Comparison of Means

	White Supervisor (1)	Black Supervisor (2)	p-value (3)
Panel A: White Officers			
Total Arrests	29.15	30.36	0.002
Violent-Crime Arrests	4.60	4.12	0.025
Property-Crime Arrests	3.96	3.23	0.000
Non-Index Crime Arrests	20.58	23.01	0.018
TRR Filings	0.61	0.59	0.015
Strong Force Ratio	0.88	0.91	0.755
Complaints	0.54	0.59	0.031
Panel B: Black Officers			
Total Arrests	16.57	14.47	0.001
Violent-Crime Arrests	3.80	3.25	0.000
Property-Crime Arrests	2.67	2.08	0.001
Non-Index Crime Arrests	10.11	9.14	0.046
TRR Filings	0.38	0.30	0.008
Strong Force Ratio	0.88	0.89	0.121
Complaints	0.53	0.51	0.011
Panel C: Hispanic Officers			
Total Arrests	28.94	27.91	0.000
Violent-Crime Arrests	5.20	4.31	0.000
Property-Crime Arrests	4.06	3.02	0.000
Non-Index Crime Arrests	19.67	20.58	0.017
TRR Filings	0.65	0.48	0.000
Strong Force Ratio	0.91	0.97	0.169
Complaints	0.58	0.54	0.011

Source: CPD analysis sample.
Notes: This table reports mean lagged annual work measures for officers assigned to white supervisors and black supervisors. The third column reports the p-value from a t-test of a difference in means after controlling for officer birth year and tenure, and including fixed effects for unit, year, and unit-year. Non-index arrests include arrests for non-property and non-violent crimes.

The table says that the average white officer assigned to a white supervisor this year had 29.2 arrests last year, of which 4.6 were for violent crimes, 3.96 were for property crimes, and 20.6 were for non-index crimes. The average white officer as-

signed to a black supervisor this year had 30.4 arrests last year, of which 4.1 were for violent crimes, 3.2 were for property crimes, and 23 were for non-index crimes. The average black officer assigned to a white supervisor this year had 16.6 arrests last year, of which 3.8 were for violent crimes, 2.7 were for property crimes, and 10.1 were for non-index crimes. The average black officer assigned to a black supervisor this year had 14.5 arrests last year, of which 3.3 were for violent crimes, 2.1 were for property crimes, and 9.1 were for non-index crimes.

Overall, the numbers suggest that officers assigned to white supervisors vs. black supervisors look very similar in terms of work performance. Although most work measures have a statistically significant black-white supervisor difference, they are all economically small. The largest difference is in non-index crime arrests made by white officers assigned to white supervisors vs. black supervisors, but this is equal to 2.4 annual arrests when the annual average is over 20.

Most importantly, the summary statistics suggest that if there is any selection, it would work against our findings. That is, black officers appear to be positively selected to white supervisors and white officers appear to be slightly negatively selected to white supervisors based on arrest numbers. This suggests that, in the absence of racial bias, white supervisors should nominate black officers more often than white officers based on their work performance measures. Of course, we cannot test whether any unobservable traits are similar between white and black officers assigned to the same supervisor. For example, we may be concerned that, due to homophily, white officers may feel more comfortable than their black colleagues in opening up to their white supervisors. If this is the case, white supervisors may be more informed of the achievements of white officers relative to black officers, leading to a black-white nomination gap. We address this potential concern in Section 2.5.1 by leveraging the randomized quarter of evaluation.

2.5 Results

2.5.1 Black-White Gap by Evaluation Quarter

This section explores whether the black-white nomination gap is due to lack of information acquisition by supervisors. The 2016 report by the Police Accountability Task Force found little stability in supervisor-officer relationships. Not only are supervisors reassigned every January, but officers may not be assigned to work with their officially assigned supervisor during the course of their shift. Second, personnel information does not necessarily get transferred to supervisors when officers switch assignments. Therefore, one potential explanation for why white supervisors may be less likely to nominate black officers is because they are less likely to interact with them and, therefore, are less likely to be informed of their accomplishments (Glover et al., 2017).

To test this theory, we exploit an institutional feature that randomizes the quarter in which officers are evaluated by their supervisor. Although there appears to be little interaction between officers and supervisors on a daily basis, we assume that the annual evaluation requires supervisors to acquire information about the officer's work record.

We exploit this institutional feature and compare nomination likelihoods of black vs. white officers assigned to white supervisors across quarters. Because the evaluation quarter is randomly assigned, this simple comparison allows us to isolate the effect of acquiring information. If a lack of information is the reason for a black-white nomination gap, then we would expect this to disappear in the quarter when supervisors are required to evaluate their assigned officers. For this analysis, the sample is at the officer-month level. We estimate the following model, separately for white supervisors and black supervisors:

$$
\begin{aligned}
Nom_{it} = \beta_0 + \sum_{q=-1}^{2} \mathbb{1}\{RQ = q\}\delta^q + \left(B_i \times \sum_{q=-1}^{2} \mathbb{1}\{RQ = q\} \right) \beta_1^q \\
+ \left(H_i \times \sum_{q=-1}^{2} \mathbb{1}\{RQ = q\} \right) \beta_2^q + \left(A_i \times \sum_{q=-1}^{2} \mathbb{1}\{RQ = q\} \right) \beta_3^q \qquad (2.1) \\
+ \left(N_i \times \sum_{q=-1}^{2} \mathbb{1}\{RQ = q\} \right) \beta_4^q + X'\alpha + \tau_t + \epsilon_{it}
\end{aligned}
$$

where i denotes officer and t denotes month. Nom_{it} is equal to 1 if officer i was nominated for an award in month t and 0 if not. B_i is a binary indicator variable if the officer is black, H_i if Hispanic, A_i if Asian, and N_i if Native American. White officers are the reference group.

The second term is a set of binary indicator variables for each quarter relative to the evaluation quarter, which is denoted as $RQ = 0$. The reference quarter is $RQ = -2$, or two quarters prior to the evaluation quarter.[34] The coefficients δ^q tell us how nomination likelihoods for white officers change across quarters. If information acquisition is an important mechanism, then we expect it to be enhanced in the quarter that supervisors evaluate their officers, $RQ = 0$.

The third term in parentheses interacts the black indicator variable and the relative-quarter indicator variables. The coefficients β_1^q depict how the black-white nomination gap evolves relative to $RQ = -2$. If the black-white difference does not change in subsequent quarters, then we expect β_1^q to be zero. Likewise, the coefficients β_2^q tell us how the Hispanic-white nomination gap evolves over time.

X is a vector of officer, supervisor, and district characteristics. Officer controls

[34] We drop observations that are three quarters before the evaluation quarter and three quarters after the evaluation quarter. This is because these observations are all either in January, February, and March (if three quarters prior) and October, November, and December (if three quarters post). Therefore, it is difficult to disentangle the calendar-month effect from the distance-to-evaluation effect.

include officer's birth year, district assignment, tenure, the number of arrests the officer made, and the number of complaints made against the officer. Supervisor controls include supervisor fixed effects and the share of black supervisees. District characteristics include overall crime rate and violent crime rate. All time-varying variables except for tenure, district assignment, and the share of black supervisees are lagged by one month. We also include fixed effects for year and month in τ_t. We estimate robust standard errors to account for heteroskedasticity introduced by the binary dependent variable.

Table 2.4 reports estimates for δ^q (white officers), β_1^q (black-white difference), and β_2^q (Hispanic-white difference) for white supervisors in Panel A and for black supervisors in Panel B. White supervisors are more likely to nominate white officers as they move closer to the evaluation quarter, but this behavior stops afterward. Relative to two quarters before their evaluation, white officers are 0.68 percentage points (23.4 percent) more likely to be nominated in the quarter before their evaluation and 0.94 percentage points (32.4 percent) more likely to be nominated in their evaluation quarter. These estimates are statistically significant at the 1 percent level. After the evaluation, the relative nomination likelihood is 0.22 percentage points higher but not statistically significant. Two quarters post-evaluation, there is essentially no difference in nomination likelihood relative to two quarters prior.

For black officers assigned to white supervisors, the story is a different one. Although black officers are also more likely to be nominated in the quarter leading up to and including their evaluation, relative to two quarters prior, the relative increase is less than half the increase for white officers (specifically, about 35.4 to 42.7 percent less) and becomes negative after the evaluation. In fact, the estimates are stable around -0.4 to -0.5 percentage points across all quarters. Taken together, the relative increase in nomination likelihood suggests that white supervisors may be learning

about their black officers, but the persistent, negative black-white disparity suggests that the learning is not manifesting in changed nominating behavior. This behavior is consistent with taste-based discrimination. By contrast, the nomination patterns for Hispanic officers are similar to those for white officers. In Appendix Table 3.13, we estimate a version with officer fixed effects instead of supervisor fixed effects. The results are similar.

Table 2.4: Racial Difference in Nomination Likelihood by Quarter

Estimates for:	Outcome Variable: Nominated White Officer	Black-White Gap	Hispanic-White Gap
	(1)	(2)	(3)
Quarter relative to two quarters before evaluation			
Panel A: White Supervisors			
One quarter pre-evaluation	0.00683***	-0.00421	0.00133
	(0.00210)	(0.00282)	(0.00363)
Evaluation quarter	0.00937***	-0.00507*	0.00182
	(0.00213)	(0.00275)	(0.00351)
One quarter post-evaluation	0.00219	-0.00531**	-0.00203
	(0.00227)	(0.00270)	(0.00348)
Two quarters post-evaluation	-0.000280	-0.00425	-0.000953
	(0.00257)	(0.00304)	(0.00386)
Observations		154,964	
Panel B: Black Supervisors			
One quarter pre-evaluation	-0.000218	0.00450	0.0180*
	(0.00648)	(0.00697)	(0.00939)
Evaluation quarter	0.00127	0.00276	0.0227**
	(0.00631)	(0.00667)	(0.00905)
One quarter post-evaluation	-0.0129**	0.0139**	0.0245***
	(0.00588)	(0.00620)	(0.00847)
Two quarters post-evaluation	-0.00625	0.00450	0.0210**
	(0.00694)	(0.00701)	(0.0107)
Observations		26,556	
Baseline B-W Nomination Gap		-0.0045	

Source: CPD analysis sample.
Notes: This table depicts how the probability of nomination changes by quarter relative to two quarters before the officer's evaluation. Estimates are reported for white supervisors in Panel A and for black supervisors in Panel B. Each panel is a single OLS regression with estimates for white officers in column 1, the black-white difference in column 2, and the Hispanic-white difference in column 3. All estimates include supervisor, month, and year fixed effects, and control for officer birth year, tenure, district, lagged arrests, lagged complaints, lagged overall crime rate, lagged violent crime rate, and the share of black supervisees. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

We can examine black supervisors' behavior in Panel B as a contrast to provide context for white supervisors' behavior. The analysis suggests that black supervisors largely do not change their nomination behavior for white and black officers before and in the evaluation quarter. But, white officers are less likely to be nominated after

their evaluation relative to two quarters prior. The black-white gap widens by -1.3 percentage points (137 percent) in the quarter immediately following their evaluation, and this is significant at the 5 percent level.

To summarize, white supervisors nominate white and Hispanic officers at similar rates and black officers at a lower rate. These patterns suggest the black-white gap is not due to in-group favoritism towards white officers but rather bias against black officers.

2.5.2 Black-White Gap by Arrest Record

Next, we further examine whether the black-white nomination gap is due to bias against black officers or in-group favoritism towards white officers. Specifically, we ask whether an officer's arrest record affects a supervisor's likelihood of nomination and whether there are any differential effects for minority officers. That is, conditional on the officer's arrest record, is there a black-white disparity in the probability of nomination? The regression sample for this analysis is at the officer-month level. We estimate the following model, separately for white supervisors and black supervisors:

$$\begin{aligned} Nom_{it} = \beta_0 &+ \left(\sum_{c=1}^{5} \mathbb{1}\{Arrests_{i,t-1} = c\} \times \beta_1^c \right) + \left(B_i \times \sum_{c=1}^{5} \mathbb{1}\{Arrests_{i,t-1} = c\} \right) \beta_2^c \\ &+ \left(H_i \times \sum_{c=1}^{5} \mathbb{1}\{Arrests_{i,t-1} = c\} \right) \beta_3^c + \left(A_i \times \sum_{c=1}^{5} \mathbb{1}\{Arrests_{i,t-1} = c\} \right) \beta_4^c \\ &+ \left(N_i \times \sum_{c=1}^{5} \mathbb{1}\{Arrests_{i,t-1} = c\} \right) \beta_5^c + X'\alpha + \tau_t + \varepsilon_{it} \end{aligned} \tag{2.2}$$

where i denotes officer and t denotes month. Nom_{it} is equal to 1 if officer i was nominated for an award in month t and 0 if not. $Arrests_{i,t-1}$ is the number of arrests

officer i made last month. We lag arrests because nominations must be submitted within 45 days of an incident. The reference category is zero arrests last month.

B_i is a binary indicator variable if the officer is black, H_i if Hispanic, A_i if Asian, and N_i if Native American. White officers are the reference group.

X is a vector of officer, supervisor, and district characteristics. Officer controls include officer's birth year, district assignment, tenure, and the number of complaints made against the officer. Supervisor controls include supervisor fixed effects and the share of black supervisees. District characteristics include overall crime rate and violent crime rate. All time-varying variables except for tenure, district assignment, and the share of black supervisees are lagged by one month. We also include fixed effects for year and month in τ_t. We estimate robust standard errors.

The parameters of interest are β_1^c, which tell us how the nomination likelihood changes as the number of arrests last month increases, and β_2^c, which tell us how the black-white difference changes by the number of arrests. We expect β_1^c to be positive and increasing in the number of arrests. If the black-white gap in award nominations does not vary by the number of arrests, then β_2^c will be zero. As the baseline black-white gap is negative, a negative β_2^c indicates that the black-white gap widens with the number of arrests, whereas a positive β_2^c indicates that the black-white gap narrows with the number of arrests.

Table 2.5 reports estimates for white officers in column 1, the black-white difference in column 2, and the Hispanic-white difference in column 3. Panel A reports estimates for white supervisors, and Panel B for black supervisors. There are increasing returns to having more arrests, with a marked increase for those with five or more arrests (column 1). Although we do not assert that arrests are an accurate measure of policing quality, we do the analysis this way because police departments seem to value and reward arrest quantity. It is interesting, therefore, that the return

to having more arrests is less for black officers compared to white officers (Panel A, column 2). The black-white difference in nomination probability for officers with one arrest widens by 0.5 percentage points compared to the black-white difference among officers with no arrests last month. This estimate is significant at the 5 percent level.

Table 2.5: Impact of Arrest Record on Nomination Likelihood by Officer Race

Estimates for:	Outcome Variable: Nominated		
	White Officer	Black-White Gap	Hispanic-White Gap
	(1)	(2)	(3)
Panel A: White Supervisors			
One arrest	0.00912***	-0.00497**	-0.000903
	(0.00135)	(0.00193)	(0.00239)
Two arrests	0.0147***	-0.00761***	0.000299
	(0.00190)	(0.00278)	(0.00329)
Three arrests	0.0216***	-0.0105**	0.000148
	(0.00260)	(0.00408)	(0.00447)
Four arrests	0.0250***	-0.00252	-0.000720
	(0.00326)	(0.00586)	(0.00566)
Five or more arrests	0.0566***	-0.0236***	-0.0156***
	(0.00256)	(0.00481)	(0.00428)
Observations		171,094	
Mean Pr(Nom) for White Officers		0.031	
Panel B: Black Supervisors			
One arrest	-0.000465	0.00354	0.00734
	(0.00309)	(0.00354)	(0.00639)
Two arrests	0.00855*	-0.00635	0.000517
	(0.00509)	(0.00570)	(0.00997)
Three arrests	0.0101	-0.000252	0.00940
	(0.00659)	(0.00800)	(0.0137)
Four arrests	0.00844	-0.0151	0.00810
	(0.00889)	(0.0102)	(0.0171)
Five or more arrests	0.0451***	-0.00294	-0.0224*
	(0.00763)	(0.0113)	(0.0125)
Observations		29,413	
Mean Pr(Nom) for White Officers		0.022	

Source: CPD analysis sample.
Notes: This table reports estimates for the impact of an officer's arrest record on the probability of nomination by white supervisors (Panel A) and by black supervisors (Panel B). Each panel is a single OLS regression with estimates for white officers in column 1, the black-white difference in column 2, and the Hispanic-white difference in column 3. All estimates include supervisor, month, and year fixed effects, and control for officer birth year, tenure, district, lagged complaints, lagged overall crime rate, lagged violent crime rate, and the share of black supervisees. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

The black-white nomination gap widens as the number of arrests increases. Among officers with five or more arrests, the relative black-white difference widens by 2.4 percentage points ($p < 0.01$). It is informative to interpret this disparity in the

context of racial differences in work performance. For example, black officers with 5 or more monthly arrests are at the 94th percentile of their distribution, while white officers are at the 81st percentile of their distribution. Yet, white supervisors are even *less* likely to nominate black officers over white officers compared to if both had zero arrests. In Appendix Table 3.14, we estimate a version with officer fixed effects instead of supervisor fixed effects. The results are very similar.

We also examine whether white supervisors are less likely to nominate Hispanic officers, another racial minority in the Chicago Police Department. The Hispanic-white difference is pretty trivial and not statistically significant until the five or more arrests category (Panel A, column 3). Among officers with at least five arrests, the Hispanic-white gap in nomination probability widens by 1.6 percentage points ($p < 0.01$).

When comparing between the two racial minorities, the black-white difference is statistically significantly different from the Hispanic-white difference among officers with one to three arrests and not for those with four or more arrests. This suggests that white supervisors are less likely to nominate black officers relative to white or Hispanic officers among those with average arrest records (recall the average number of arrests is two), but favor white officers when comparing officers with higher than average arrests.

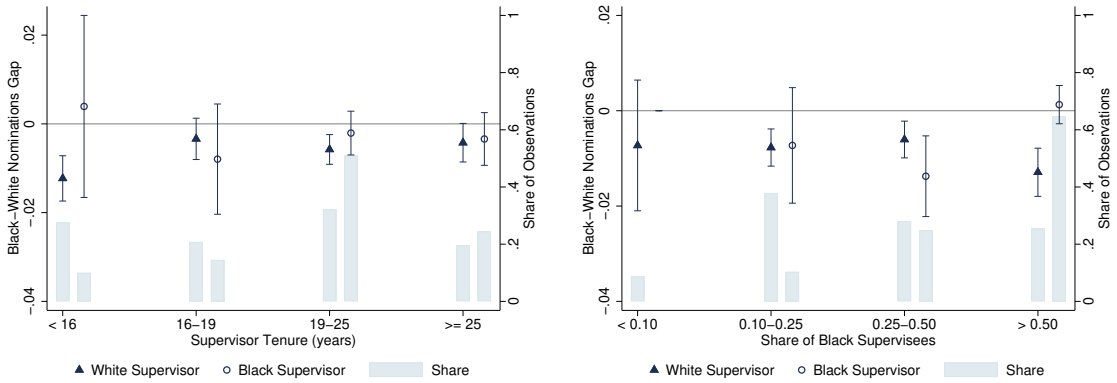
We examine black supervisors' behavior in Panel B. Most of the point estimates are not significant, though this may be due to the fact that there are few black supervisors (190 compared to 893 white supervisors). However, the magnitudes of the point estimates are also very small. One exception is the Hispanic-white gap of -0.022 among officers with five or more arrests ($p < 0.1$).

To summarize, we find that black officers are nominated less frequently than white officers even conditional on arrests. The gap widens as the number of arrests increases,

which is the opposite of what one would expect if the disparity were due to white supervisors' inaccurate beliefs about black officers' work measures. This is supported by the fact that we do not see similar behavior exhibited by black supervisors. Instead, the results suggest that the source of discrimination may be taste-based. These findings are consistent with the analysis from the previous section, which finds that supervisors may not be updating their behavior as they learn more information.

2.5.3 Black-White Gap by Supervisor Characteristics

We also examine how the black-white nomination gap changes by supervisor characteristics in Figure 2.1. Panel A reports the black-white nomination gap by supervisor tenure, with quantiles defined by the 25th, 50th, and 75th percentile values (about 16 years, 19 years, and 25 years, respectively). The patterns indicate that less experienced white supervisors have the largest negative black-white nomination gap at -1.2 percentage points ($p < 0.01$). In contrast, black supervisors with similar levels of experience have a positive black-white gap but this is not statistically significant.



(a) Supervisor Tenure (b) Share of Black Supervisees

Figure 2.1: Black-White Nomination Gap by Supervisor Characteristics

Notes: This figure depicts how the black-white nomination gap changes by supervisor tenure (Panel A) and by the share of black supervisees (Panel B). All estimates include supervisor, month, and year fixed effects, and control for officer birth year, tenure, district, lagged arrests, lagged complaints, lagged overall crime rate, lagged violent crime rate, the number of supervisees, and the share of black supervisees. Wings depict 95% confidence intervals with robust standard errors.

Panel B reports the black-white nomination gap by the share of black supervisees. Here, the patterns indicate that the black-white gap among white supervisors becomes more negative and statistically significant when the share of black supervisees exceeds 50 percent. Among supervisors with a majority of black officers, white supervisors have a black-white nomination gap of -1.3 percentage points ($p < 0.01$), while it is slightly positive and not significant for black supervisors.

2.6 Experimental Evidence

The previous section presented evidence that white CPD supervisors are less likely to acquire information about their black officers, and this may be due to taste-based discrimination. To more concretely test this, we ran an online experiment to measure

the review process in the nomination decision.[35] This section discusses the evidence on whether evaluators choose to engage with minority officers.

Participants were asked to review CPD officer profiles and nominate one for an award. We study how officer race affects two types of choices: attention to an officer profile and the nomination decision. First, we measured which profiles participants hovered over, the order in which participants hovered over the profiles, and how long participants hovered over each profile. Second, we measured which officer was ultimately nominated for an award.

By using the same officers from the CPD analysis sample, we are able to generalize our findings to a broader evaluator group than Chicago police supervisors. At the same time, we do not necessarily expect the two evaluator groups to act very differently; although Chicago police supervisors may be a selected sample, demographically-speaking, Dickinson et al. (2015) finds that police commissioners are no different from non-police civilians when it comes to issuing rewards.[36]

2.6.1 Experimental Design

Survey participants were given two different types of tasks. In the first type of task, participants chose between a black male officer and a non-black male officer, where the black male officer was randomly assigned to be either "high-quality" or "low-quality" and the non-black male officer was assigned the converse.[37] In judging officer profiles, we used the number of civilian complaints and arrests. These classifications are admittedly subjective but they were made independently of officer race and sex.

[35]The experiment was pre-registered in the AEA RCT Registry, AEARCTR-0005929.

[36]In an experiment, Dickinson et al. (2015) finds that police commissioners are slightly more likely than non-police subjects to issue rewards but with less intensity. However, these differences are not statistically significant.

[37]See Appendix Figure 3.9 for a screenshot of the task.

"High-quality" profiles were those with zero civilian complaints and an above-average number of arrests. "Low-quality" profiles were those with one or two civilian complaints and a below-average number of arrests.

In the second type of task, participants were shown four officer profiles and asked to nominate one for an award. In this task, officer profiles displayed only demographic information (race, sex, and age) and participants had to mouse over a profile to reveal full information about the officer.[38] All officers were of "average quality", defined as having zero or one civilian complaints and an average number of arrests. There were two iterations of this task. In the first iteration, three of the four profiles were always white officers and the race of the fourth profile was randomly chosen amongst white, black, and Hispanic. In the second iteration, the officer group was racially heterogeneous. Three of the four profiles always featured a white officer, a black officer, and an Hispanic officer. The race of the fourth profile was randomly chosen amongst these three races. The ordering of these two iterations was randomized.

The ordering of the two tasks was randomized, and the display ordering of officer profiles in each of the tasks was also randomly determined. All tasks were time-constrained to introduce a cost to reviewing profiles. Participants had 20 seconds to complete the first task (pairwise comparison) and 40 seconds to complete the second task (group comparison).[39] For the second type of task, participants were restricted from uncovering any work performance measures and from moving onto the next page for ten seconds. This was to ensure that participants had enough time to view and review the demographic information (e.g., race) of the four officer profiles on the screen. Although participants were asked to nominate an officer, they were

[38]See Appendix Figure 3.10 for a screenshot of the task.

[39]These time limits appear to be within reason; participants took about 9.8 seconds, on average, for the pairwise comparison and 27.9 seconds, on average, for the group comparison. For the group comparison, conditional on mousing over any profile, about 70 percent of participants moused over all four profiles.

not required to do so; participants were able to move onto the next page without nominating an officer. See Appendix 3.10.2 for more information about the online experiment.

2.6.2 Sample Selection and Data

The experiment was conducted on Amazon Mechanical Turk in July 2020.[40] We recruited 411 MTurk workers (hereafter "evaluators") who were 18 years of age or older, based in the United States with English language proficiency, and who had access to a computer with a mouse and Javascript. The technical requirements were necessary in order to capture mouse movements on the screen. The survey had three data quality checks to identify bots and to ensure evaluators paid attention during the survey. For the analysis, we decided to include evaluators who passed at least two of the three data quality checks. This restriction reduces our final analysis sample to 407 evaluators.

2.6.3 Are Black Officers Less Likely to Be Nominated for an Award?

First, we seek to replicate the results from the CPD administrative data and ask whether black officers were less likely to be nominated for an award. Columns 1 through 3 of Table 2.6 report results from the pairwise comparison of a black and non-black (white or Hispanic) officer. Column 1 reports results from all MTurk evaluators, column 2 is restricted to white MTurk evaluators, and column 3 is restricted to black MTurk evaluators.

[40]It is possible that the George Floyd incident on May 25, 2020 and subsequent protests may have altered people's perceptions of the police and black individuals. Specifically, the incident may have increased MTurk workers' interest in and affinity towards black officers because they are black. This would work against our results, which find that black officers are less likely to be moused over and are less likely to be nominated when paired against a non-black officer.

Table 2.6: Impact of being Black on Nomination Likelihood

Outcome Variable: Nominated				
Pairwise Comparison:	Black v. Non-Black			High v. Low
Race of MTurk Worker:	All	White	Black	All
	(1)	(2)	(3)	(4)
High-Quality Profile	0.483***	0.410***	0.771***	0.524***
	(0.0921)	(0.0947)	(0.284)	(0.169)
Black Officer	-0.0883***	-0.103***	-0.0163	-0.0324
	(0.0311)	(0.0361)	(0.0823)	(0.119)
High-Quality x Black Officer	0.0326	0.0487	-0.00390	0.141
	(0.0448)	(0.0518)	(0.121)	(0.171)
Female Officer				-0.0705
				(0.0680)
High-Quality x Female Officer				0.137
				(0.107)
Observations	1,576	1,196	256	794

Source: MTurk survey data.
Notes: This table reports estimates from a pairwise comparison of officer profiles. Columns 1-3 are a pairwise comparison between a black male officer and a non-black male officer. Column 4 is a pairwise comparison between two officers of the same race and sex but differing profile qualities. All estimates control for officer traits and profile location on the screen. Officer traits include officer age, tenure, arrests, and complaints. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Low-quality black officers are 8.8 percentage points ($p < 0.01$) less likely than low-quality white officers to be nominated. This gap largely persists with high-quality officers. Although high-quality officers are almost 50 percentage points more likely to be nominated for an award ($p < 0.01$) than low-quality officers, high-quality black officers are still 5.6 percentage points less likely to be nominated than high-quality white officers. This difference is statistically significant at the 10 percent level. When we focus on the race of the evaluators, we see that the results are driven by white evaluators (column 2). The black-white gap among white evaluators is -10 percentage points ($p < 0.01$) for low-quality officer profiles and -5.5 percentage points for high-quality officer profiles (p-value = 0.111). There is no statistically significant black-white nomination gap among black evaluators. These results are consistent with

the CPD analysis, which found the black-white nomination gap widens for white supervisors.

In column 4, we conduct a robustness check wherein the two officer profiles are of the same race and sex and differ only in terms of quality. As expected, high-quality profiles are more likely to be nominated–about 52 percentages-points–and this is significant at the 1 percent level. This also provides an indirect test that MTurk evaluators were able to discern the quality difference between the two officer profiles. Interestingly, when the officers are both black or both female, evaluators are even more likely to nominate the high-quality officer (about 14 additional percentage points for black and female officers) relative to when the officers are white males. Although these estimates are not statistically significant, the positive point estimates, together with the results from column 1, suggest that either white males are given some slack even if they do not meet a certain standard or that minorities are held to a higher standard.

2.6.4 Do Evaluators Choose to Engage with Black Officers?

Table 2.7 presents summary statistics on MTurk evaluators' engagement with white officers, black officers, and Hispanic officers. We examine three different measures of engagement: ever moused over (Panel A), first mouseover (Panel B), and mouseover duration measured in seconds (Panel C). Table 2.8 reports the findings in a regression framework, where we control for the profile location on the computer screen, and the evaluator's starting mouse position.

The first row in Panel A tells us that evaluators tend to mouse over most of the officer profiles: over 80 percent of officer profiles were moused over. Specifically by race, 84.2 percent of white officer profiles were moused over, 81.5 percent of black officer profiles were moused over, and 81.8 percent of Hispanic officer profiles were

moused over. The black-white difference is borderline significant, with a p-value of 0.107. The Hispanic-white difference is not statistically significant.

When the officer pool is predominantly white—an environment that resembles the Chicago Police Department—the black-white engagement gap widens. Black officers are 7.3 percentage points less likely to be moused over compared to white officers ($p < 0.05$). However, if the minority officer is Hispanic, then there is no statistically significant difference in mouse-over likelihood.

Conditional on being moused over, there does not appear to be a significant black-white difference regarding which officer is moused over first (Panel B). However, there is a racial difference in the amount of time spent reviewing profiles (Panel C). Evaluators spend around half a second more reviewing black and Hispanic profiles, and these are significant at the 1 percent level.

Table 2.7: Evaluator Engagement by Officer Race, Comparison of Means

	White Officer	Black Officer	Hispanic Officer	B-W Difference (p-value)	H-W Difference (p-value)
Panel A: Outcome Variable: Ever Moused Over					
All Officers	84.2%	81.5%	81.8%	-0.028 (0.107)	-0.024 (0.163)
Predom. White Officer Group	83.8%	76.6%	86.7%	-0.073 (0.036)	0.028 (0.420)
Het. Race Officer Group	85.2%	82.7%	80.6%	-0.025 (0.276)	-0.046 (0.054)
Panel B: Outcome Variable: First Mouseover					
All Officers	30.0%	31.9%	28.4%	0.019 (0.419)	-0.016 (0.490)
Predom. White Officer Group	28.9%	35.7%	35.6%	0.069 (0.155)	0.067 (0.152)
Het. Race Officer Group	32.8%	31.0%	26.5%	-0.018 (0.581)	-0.063 (0.048)
Panel C: Outcome Variable: Mouseover Duration (seconds)					
All Officers	2.33	2.78	2.89	0.448 (0.000)	0.559 (0.000)
Predom. White Officer Group	2.36	3.40	3.07	1.041 (0.000)	0.714 (0.001)
Het. Race Officer Group	2.27	2.64	2.85	0.366 (0.006)	0.576 (0.000)

Source: MTurk survey data.
Notes: This table reports mean values for the three measures of information acquisitions: ever moused over in Panel A, first mouseover in Panel B, and mouseover duration in Panel C. B-W Difference reports the percentage-point difference between black officers and white officers. H-W Difference reports the percentage-point difference between Hispanic officers and white officers. p-values are the p-value from a t-test of a difference in means.

When the officer pool is heterogeneous—that is, white, black, and Hispanic officers are represented in equal numbers—the black-white disparity disappears and an Hispanic-white disparity emerges. Hispanic officers are 4.6 percentage points less likely to be moused over and 6.3 percentage points less likely to be the first mouseover, relative to white officers. One potential explanation is that when the two racial minorities (black officers and Hispanic officers) are in the same comparison pool with white officers, black officers crowd out Hispanic officers in regards to evaluator attention. We are uncertain of why this may be the case, but it is possible that the George Floyd protests, which took place about a month prior to the online experiment, may have affected evaluators' decisions on who to mouseover first. This, however, does appear to be a crowd-out effect because when the black officers are the sole minority officer in a group with three white officers, evaluators are less likely to engage with them.

Table 2.8: Impact of Officer Race on Evaluator Engagement

Officer Pool:	All Officers (1)	Predom. White (2)	Het. Race (3)
Panel A: Outcome Var.: Moused Over			
Black Officer	-0.0289	-0.0707*	-0.0271
	(0.0177)	(0.0386)	(0.0230)
Hispanic Officer	-0.0252	0.0332	-0.0475**
	(0.0179)	(0.0329)	(0.0237)
Observations	2,992	1,492	1,500
Mean Outcome for White Officer	0.842	0.838	0.852
Panel B: Outcome Var.: First Mouseover			
Black Officer	0.00355	0.0399	-0.0285
	(0.0210)	(0.0409)	(0.0293)
Hispanic Officer	-0.0136	0.0681	-0.0578**
	(0.0212)	(0.0424)	(0.0293)
Observations	2,488	1,245	1,243
Mean Outcome for White Officer	0.300	0.289	0.328
Panel C: Outcome Var.: Mouseover Duration (sec)			
Black Officer	0.431***	0.984***	0.347**
	(0.107)	(0.230)	(0.135)
Hispanic Officer	0.570***	0.697***	0.575***
	(0.107)	(0.224)	(0.137)
Observations	2,488	1,245	1,243
Mean Outcome for White Officer	2.335	2.361	2.270

Source: MTurk survey data.
Notes: This table reports estimates for racial differences in a group comparison of officer profiles. We examine three different measures of information acquisition: ever moused over in Panel A, first mouseover in Panel B, and mouseover duration in Panel C. All estimates control for profile location on screen and evaluator's starting mouse position. Standard errors are in parentheses. Panels A and B report robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

2.6.5 The Importance of Evaluator Engagement

In this section, we examine how evaluator engagement affects the probability of nomination. Table 2.9 reports the nomination likelihood conditional on officer race and the engagement measure: ever moused over (Panel A), first mouseover (Panel B),

and mouseover duration (Panel C). All estimates control for officer traits, the officer's profile location on the screen, and the evaluator's starting mouse position on the screen.

Profiles that were moused over are on average 8.8 percentage points more likely to be nominated (Panel A). This suggests that mouseover activity is a good measure of the evaluator's interest in the officer. Workers who engage with black or Hispanic officers are even more likely to nominate these minority officers–by an additional 8.7 percentage points and 14 percentage points, respectively. Column 2 reveals that these results are driven by the predominantly white officer pool.

We also analyze first mouseover as a proxy for the intensity of an evaluator's interest in an officer. Conditional on being moused over, being moused over first increases the probability of nomination by 2.5 percentage points though this is not statistically significant (Panel B). Relative to white profiles, however, black profiles who are moused over first are 9.2 percentage points even more likely to be nominated ($p < 0.1$). Column 2 indicates that this finding is driven by a more racially homogeneous officer pool. Since a minority officer will stand out in a predominantly white officer pool, if an evaluator chooses to learn about the minority officer first, then the evaluator is also decidedly more likely to nominate the minority officer (19 percentage points for black officers and 17 percentage points for Hispanic officers). As the mean nomination rate for white officers is 21.6 percent, these estimates imply that being moused over first increases the probability of nomination by 89 percent for black officers and 79.6 percent for Hispanic officers in comparison to white officers.

It is also informative to examine the situation when evaluators do not engage with officers. In this instance, we may expect no racial difference in nomination likelihood since evaluators have no work information about the specific officers. However, we find that black officers are less likely to be nominated relative to white officers whose

work measures were also not revealed. This difference widens to -38 percentage points and becomes statistically significant at the 1 percent level when the officer pool is predominantly white.

It is difficult to ascertain whether this behavior is driven by distaste or inaccurate beliefs about black officers (for example, evaluators may believe black officers are less productive than white officers and therefore may not engage with them in the interest of time). The estimates in Panel B may shed some light. Among officers who were moused over, but were not the first mouseover, black officers were still less likely to be nominated relative to their white counterparts. In a predominantly white officer pool, black officers who were not the first mouseover are 13.8 percentage points less likely to be nominated than their white peers who were not the first mouseover. In contrast, Hispanic officers who were not the first mouseover are 19.4 percentage points more likely to be nominated relative to their white peers. Both of these differences are significant at the 10 percent level. These results suggest that evaluators are not updating their information about black officers, which is consistent with our results using CPD administrative data in Section 2.5.1.

Next, we consider the length of the review process (Panel C). The longer an evaluator spends viewing an officer's profile, the higher the chance of a nomination: an increase of 4.4 percentage points for each additional second ($p < 0.01$). This estimate does not differ for black officers; each additional second on a white officer's profile increases the probability of nomination the same as an additional second spent on a black officer's profile. When the officer pool is predominantly white, the coefficient on the interaction between mouseover duration and black officers is 0.039, which means that an additional second spent reviewing a black profile is associated with an additional 3.9 percentage point increase in nomination likelihood relative to time spent reviewing a white profile. Given the baseline black-white nomination gap of

-0.198, this suggests that an evaluator would need to spend about five additional seconds–more than two standard deviations–on a black officer's profile in order to equalize the nomination probability for white vs. black officers. In contrast, any Hispanic-white differences become insignificant in a predominantly white officer pool.

Taken together, the results appear to suggest that there is a dichotomy in the type of evaluators who are likely to engage with minority officers vs. not. Because all officer profiles are of average quality, the decision to nominate an officer should be independent of officer performance and instead reflect solely the evaluator's preferences. Some evaluators are interested in engaging with black and Hispanic officers, and those evaluators are also more likely to nominate minority officers. This suggests that black officers may benefit from having supervisors who are interested in interacting with them.

On the other hand, some evaluators are less interested in engaging with black officers. Black profiles are less likely to be moused over than white profiles, on average, and this is more salient when evaluators are choosing among three white officers and one black officer. We do not see similar patterns when evaluators are choosing among three white officers and one Hispanic officer. Further, black officers are less likely to be nominated than white officers conditional on neither being the first profile the evaluator moused over.

It is informative to compare estimates when the officer pool is predominantly white vis-a-vis evenly distributed among white, black, and Hispanic officers. Most estimates lose statistical significance, but the point estimates remain similar for the most part; black officers continue to be penalized more than white officers from a lack of evaluator engagement. What disappears when we move to a more diverse officer pool are the *benefits* black officers had received from greater evaluator engagement. For example, the black-white difference among profiles that were the first mouseover

becomes negative (from 0.19 to -0.01). This suggests that having a diverse police force may not actually eliminate the black-white recognition gap. This implication is also consistent with our finding using CPD administrative data, that the black-white nomination gap becomes more negative among white supervisors with a majority of black supervisees. These findings underscore the importance of addressing the black-white *promotion* gap, as simply diversifying incoming police officers may be limited in its efficacy.

Table 2.9: Impact of Engagement on Nomination Likelihood

	Outcome: Nominated Officer		
Officer Pool:	All	Predominantly White	Heterogeneous Race
Panel A: Ever Moused Over			
Ever Moused Over	0.0882***	0.0785***	0.107**
	(0.0244)	(0.0284)	(0.0495)
Black Officer	-0.0629	-0.380***	-0.294
	(0.0388)	(0.0771)	(0.251)
Hispanic Officer	-0.157***	-0.0102	-0.0177
	(0.0378)	(0.0684)	(0.145)
Ever Moused Over x Black Officer	0.0865*	0.364***	-0.00966
	(0.0443)	(0.0797)	(0.0649)
Ever Moused Over x Hisp. Officer	0.140***	0.173**	0.104*
	(0.0412)	(0.0812)	(0.0613)
Observations	2,992	1,492	1,500
Panel B: First Mouseover			
First Mouseover	0.0251	-0.00466	0.109**
	(0.0259)	(0.0313)	(0.0483)
Black Officer	-0.00694	-0.138*	-0.205
	(0.0268)	(0.0834)	(0.302)
Hispanic Officer	-0.0330	0.110*	0.0977
	(0.0326)	(0.0600)	(0.163)
First Mouseover x Black Officer	0.0922*	0.194*	-0.0102
	(0.0493)	(0.106)	(0.0669)
First Mouseover x Hisp. Officer	0.0242	0.172*	-0.113
	(0.0528)	(0.102)	(0.0701)
Observations	2,488	1,245	1,243
Panel C: Mouseover Duration (sec)			
Mouseover Duration	0.0442***	0.0397***	0.0544***
	(0.00638)	(0.00763)	(0.0116)
Black Officer	-0.0191	-0.198***	-0.347
	(0.0258)	(0.0726)	(0.229)
Hispanic Officer	-0.105***	0.0863	0.0550
	(0.0303)	(0.0582)	(0.128)
Mouseover Duration x Black Officer	0.0102	0.0387**	-0.0127
	(0.00966)	(0.0173)	(0.0150)
Mouseover Duration x Hisp. Officer	0.0260***	0.0140	0.0188
	(0.00929)	(0.0150)	(0.0141)
Observations	2,992	1,492	1,500
Mean Outcome for White Officer	0.225	0.216	0.245

Notes: This table reports estimates for racial differences in the impact of information acquisition on nomination likelihood using the MTurk data. We examine three different measures of information acquisition: ever moused over in Panel A, first mouseover in Panel B, and mouseover duration in Panel C. All estimates control for officer traits (officer age, tenure, arrests, and complaints), profile location on screen, and evaluator's starting mouse position. Standard errors are in parentheses. Panel A reports robust standard errors. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

2.7 Conclusion

Racial bias has been extensively documented in a variety of settings, including hiring decisions (Bertrand and Mullainathan, 2004; Agan and Starr, 2017; Doleac and Hansen, 2018; Craigie, 2020), sports umpires (Parsons et al., 2011), judicial and sentencing decisions (Park, 2017; Flanagan, 2018; Mueller-Smith and Schnepel, 2017; Rehavi and Starr, 2014)[41], and bail decisions (Arnold et al., 2018). The increasing availability of police administrative data has allowed researchers to carefully examine and detect bias in law enforcement as well.[42]

A potential solution that has been put forth is to increase racial and gender diversity among officers, who are traditionally homogeneous.[43] A diverse police force may improve policing quality in various ways (Sklansky, 2005). Outwardly, it may improve the police's relationship with the community through unique skills that minority officers may possess (Miller and Segal, 2018; Harvey and Mattia, 2020; Anwar et al., 2012).[44] Inwardly, it may alter the internal dynamics of the department.

This paper asks how racial bias affects career progression, which is of particular relevance to law enforcement, where minorities are less represented at higher ranks.

[41]Mueller-Smith and Schnepel (2017) finds that the practice of diversion, or a halt or termination of one's progression through the justice system, reduces re-offending rates and improves labor market outcomes among young black men charged with misdemeanors.

[42]The release of detailed administrative data has allowed for research on various important questions, such as crime reduction (Blanes i Vidal and Kirchmaier, 2018; Mastrobuoni, 2020).

[43]For example, in their investigative report of the Ferguson Police Department, the U.S. Department of Justice called for a more diverse police force as part of a broader reform effort (U.S. Department of Justice, 2015, p. 58). Several cities, including Chicago, Indianapolis, and Knoxville, have followed this lead and pursued diversity initiatives (Chicago: `https://www.chicagotribune.com/news/breaking/ct-met-chicago-police-hiring-20180503-story.html`; Indianapolis: `https://www.indystar.com/story/opinion/columnists/suzette-hackney/2018/09/27/impd-leads-charge-toward-diversity-columnist-suzette-hackney-writes/1433649002/`; Knoxville: `https://www.knoxnews.com/story/news/local/2017/01/09/knoxville-police-department-recruits-remain-diverse-group/96345092/`)

[44]McCrary (2007) and Garner et al. (2020) do not find that court-ordered affirmative action litigation affects offense and arrest rates, but Garner et al. (2020) acknowledges that there may be racially heterogeneous effects that offset each other.

For example, white males comprised 40 percent of all entry-level police officers in the Chicago Police Department in 2015, and 56 percent of those at the rank of Sergeant or higher.[45] In addition to improving policing quality[46], minority representation—particularly at higher ranks of office—may help to recruit more minorities and close promotion gaps, which may further attract minority applicants. Prior research has found that minorities in management positions can address wage gaps and occupational gaps (Langan, 2018; Kofoed and mcGovney, 2019). At the same time, an extensive literature documents racial and gender bias in the workplace, which may hinder minorities' career progression.[47] In the context of policing, diversity initiatives may be constrained by the extent to which officer bias carries over to their colleagues. Indeed, in the Chicago Police Department, 98 percent of CPD officers believe that promotions are due to connections not merit (Police Accountability Task Force, 2016).

To examine the extent of racial bias in law enforcement, we construct a panel dataset of all CPD officers containing their personnel information. We exploit quasi-random variation in supervisor assignment and randomized timing of the evaluation, and find that white supervisors are more likely to nominate officers in the evaluation quarter, suggesting that information acquisition is important for career recognition, but also that black officers benefit less than their white peers, suggesting that white supervisors are not updating.

To supplement our CPD analysis, we run an online experiment on Amazon Mechanical Turk and, again, find that black officers are less likely to be nominated than

[45]These numbers do not include civilian Department members.

[46]See, for example, Donohue III and Levitt (2001); Miller and Segal (2018); Bulman (2019).

[47]For example, Egan et al. (2018) find that female financial advisors are 20 percent more likely than male financial advisors to lose their jobs following a misconduct. In medicine, Sarsons (2019) finds that physicians are less likely to refer to new female surgeons after a bad patient outcome but not to new male surgeons. Beaman et al. (2018) find that women are harmed in a referral-based hiring process as workplace networks tend to be gender homophilous. Glover et al. (2017) find that manager bias can cause a self-fulfilling prophecy in that biased managers interact less with minority cashiers, leading them to exert less effort.

their non-black peers. In terms of engagement, we find that black profiles are the least likely to be moused over, and that the black-white gap more than doubles when evaluators are choosing among three white officers and one black officer.

Our findings have two important policy implications for law enforcement. First, we find that interactions between supervisors and officers are an important mechanism for career recognition, suggesting that the observed racial gap in award nominations may be due to lack of information. One way to overcome discrimination that is driven by biased beliefs is continuous and sustained positive evaluations (Bohren et al., 2019). But, as we find a persistent negative black-white gap even in the evaluation quarter, our results suggest that the decentralized nature of supervision and oft-changing supervisor assignment in the CPD present a challenge for discrimination to be reversed.

Second, our finding of a persistent black-white recognition gap suggests that simply increasing the diversity of incoming recruits may not be enough to eliminate racial bias in policing. The argument for a diverse police force stems from the "contact hypothesis", or that outsider bias can be reduced if the integrated group has a common goal. Although there is empirical evidence in support of this theory (Lowe, 2019), another study finds that the improved behavior towards out-group members does not extend beyond the intervention setting (Mousa, 2020). Therefore, it is uncertain whether focusing on the diversity of incoming officers will be enough to eliminate racial bias in the department. Further, biased evaluations may lead the discriminated party to exert less effort and have lower performance, affecting pay and promotions (MacLeod, 2003). As such, police departments should also pursue policies that address in-group bias due to its effect on career advancement.

Chapter 3

Sleep: an experiment to improve students' performance

3.1 Introduction[1]

An undergraduate wisdom says: "Sleep, study, socialize. Pick two.". This wisdom manifests the dilemma of a typical college student: how to allocate time to make the best out of college? Making time allocation decisions may be especially difficult for students not just because all of their options may have high opportunity cost, but also because students may face limited information and various behavioral biases that could lead to suboptimal decisions.

This paper is about students' sleep time decisions. The interest in sleep comes from the results of the medical literature: sleep is an important input for memory formation, attention and in general to human capital formation (Pilcher and Huffcutt, 1996; Patrick et al., 2017); yet, some students routinely sleep short (Becker et al., 2018; Avery et al., 2020) and seemingly not capitalize fully on the suggested benefits of sleep.

This paper asks two questions. First, could selected behavioral interventions affect students' sleep, and second, what is the effect of sleep on students' cognitive outcomes.

To answer these questions, I ran a small scale framed-field experiment. In the experiment undergraduate participants randomized into the treatment arms received

[1]This research was facilitated by use of the labs at the Interdisciplinary Behavioral Research Center at Duke University. The experiment received IRB approval from Duke Campus IRB and is registered under the protocol #2018-0393.

light touch interventions. One intervention in particular targeted time inconsistency – a bias that students are particularly prone to (Lavecchia et al., 2016) – with bedtime reminders. Other interventions tackled limited information about sleep. One of them provided tips about how to improve sleep hygiene, while another one was an educational meditation video that taught participants a breathing technique to fall asleep quicker. The experiment lasted for three weeks for each participant, and during this time participants' sleep was tracked with fitness trackers. Besides the sleep data, I also collected survey responses and cognitive outcomes.

The treatment effects on sleep and cognitive performance are large in magnitude, although these estimates are imprecise and statistically insignificant due to the small sample size. Further investigation revealed that this is because the interventions were not successful in increasing the take-up: participants did not adhere to their planned bedtime, they did not tried practicing the meditation and there is also no sign that they followed the recommendations about improving their sleep hygiene. As an alternative strategy I used panel data methods to estimate the relationship between sleep and cognitive performance. The results suggest that higher sleep quality improves performance on the the cognitive tests by reducing the share of wrong answers.

This experiment speaks to the economics literature that studies the impact of sleep on performance (Gibson and Shrader, 2018; Bessone et al., 2020; Giuntella and Mazzonna, 2019), especially in the domain of education (Jagnani, 2020; Avery et al., 2020; Carrell et al., 2011). The findings of this study support the positive relationship between sleep and cognitive performance among college students.

The experiment also contributes to the literature on nudging in education. Previous papers successfully used various behavioral interventions, such as providing rewards for test scores (Levitt et al., 2016) or sending text reminders for students to

file financial applications (Castleman and Page, 2016) .[2] The closest paper to mine in this category is Avery et al. (2020), who offer commitment devices for undergraduate students to meet their sleep duration and bedtime targets. In this paper I show that sleep patterns may not be easy to change with light touch interventions, and stronger interventions, such as the one in Avery et al. (2020), may be needed to see an impact on sleep.

The paper is organized as follows. Section 3.2 provides the background on sleep measurement and the flagship results in the domain of economics. I discuss the theoretical framework in Section 3.3, which builds on the time allocation model of Jagnani (2020). Section 3.4 details the experimental design and procedures. In Section 3.5 I describe the baseline sleep patterns of the participants, while in Section 3.6 I report on the experimental results. These results are then discussed in Section 3.7 and finally Section 3.8 concludes.

3.2 Background

The National Sleep Foundation recommends 7-9 hours of sleep for adults (Hirshkowitz et al., 2015a), yet about a third of the adult population in the US reports sleeping less than this recommendation (Liu et al., 2016). The medical literature identified several negative consequences of sleep deprivation, such increased mortality, impaired cognitive ability and decreased mental health (Hirshkowitz et al., 2015b). In light of these results, recent estimates of the overall economic cost of sleep deprivation are around 1-2% of the GDP in the US alone (Hafner et al., 2016).

Establishing the causal effect of sleep on any outcome is not trivial. The first challenge is measurement. The medical literature considers polysomnography as the most

[2]See Lavecchia et al. (2016) and Koch et al. (2015) for reviews about behavioral economics of education.

reliable measure of sleep. To measure a participant's sleep with polysomnography, the participant has to sleep in a sleep laboratory overnight with several electrodes being attached to their body (Marino et al., 2013). While polysomnography provides precise measurement, it is costly and not practical for large scale studies. For this reason, the economics literature on sleep largely relied on self-reported measures in time use surveys for a long time. These self-reports, however, are extremely noisy, because people lose consciousness when they sleep (Ohayon et al., 2017) and so they cannot provide good estimates of their sleep. Consistent with this, self-reports have been found to systematically overestimate sleep duration (Lauderdale et al., 2008; Avery et al., 2020). As technology improved, sleep can now be measured with wearable devices. These wearables are typically worn on the wrist and measure sleep based on wrist movements (Marino et al., 2013). Because these wearable devices have a lower cost, are less intrusive and could be used at the participants' own home, they provide an opportunity to obtain sleep measurement at a larger scale (Bessone et al., 2020). The cost comes at the loss of precision, wearable devices are less precise than polysomnography (Evenson et al., 2015).

The second challenge is endogeneity. The literature has focused on natural experiments to identify the effect of sleep duration on economic outcomes. These papers are well suited to estimate the effects of the given natural experiment on the economic outcomes, however, they face difficulties in establishing that sleep is the main driver of their effects. For example, one of the often used natural experiments is the daylight savings time (DST) clock change in the spring and the fall. The clock changes are associated with changes in sleep duration, as well as with changes in the amount of light that is available at certain hours (Smith, 2016). This means that activities for which light and sleep are both important are particularly challenging to study with the DST change. Recent experimental work, such as Bessone et al. (2020) and Avery

et al. (2020) have the advantage to overcome the endogeneity issues typically faced by the observational studies.

The economics literature on sleep is small. For a long time economists thought of sleep as an unproductive, biological necessity. Biddle and Hamermesh (1990) were the first to model sleep duration as a choice variable that could respond to economic incentives. They find that if sleep has a higher opportunity cost through higher wages, individuals demand less of it. Since then, the literature adapted a view that sleep may not just be a choice variable that responds to incentives, but could also be a productive input itself that directly affects economic outcomes.

Several papers linked sleep to economic outcomes. On the domain of health, the evidence comes from exploiting natural experiments. Using self-reported time use surveys, sleep deprivation has been found to lower mental health (Giuntella et al., 2017), increase the risk of obesity, diabetes, cardiovascular diseases and breast cancer (Giuntella and Mazzonna, 2019). Sleep also affects mortality directly through accidents: sleeping an additional hour during the fall daylight saving time change (DST) is negatively associated with overall hospitalization (Jin and Ziebarth, 2020), while the DST change in the spring increases fatal automobile accidents due to an hour of sleep loss (Smith, 2016)[3].

Sleep is also linked to labor market outcomes. Gibson and Shrader (2018) modified the original Biddle-Hamermesh time allocation model to incorporate that sleep can also increase productivity. Assuming that wages are the marginal productivity of the worker, sleep could affect wages. Indeed, the authors find that an extra hour of sleep per week induced by different sunset times in the same time zones is associated with higher wages in the United States. Similar results have been reported by Giuntella and

[3]Car accidents plausibly depend on sleep as well as the available light conditions. Smith (2016) uses two exogenous variations to argue that less sleep causes more accidents: the first is the DST clock change in the spring and the second is a policy change in the timing of the DST change.

Mazzonna (2019) exploiting time zone boundaries. The positive association between sleep on productivity, however, does not seem to generalize. Bessone et al. (2020) find in a field experiment that an additional hour of night sleep does not increase productivity in India. The authors hypothesize that this is because their an extra hour of sleep has low quality in their settings. In line with this, enforced daytime naps at a comfortable sleep room increased workers' productivity.

In terms of cognitive performance, natural experiments find a positive link between sleep and cognitive performance. Here the evidence holds whether changes in sunset times (Giuntella et al., 2017; Jagnani, 2020), changes in school start times (Carrell et al., 2011) or experiments (Avery et al., 2020) are used for estimation.

This paper fits into literature on measuring the effect of sleep on cognitive outcomes. Relative to Giuntella et al. (2017); Jagnani (2020) and Carrell et al. (2011), this paper uses measured sleep variables in a field experimental setting. Relative to Avery et al. (2020), this paper uses measured cognitive outcomes that can detect changes at a higher frequency than self-reported grades, the measure that Avery et al. (2020) use. Overall, the results of the current paper support the positive relationship between sleep and cognitive performance.

In light of the positive effects of sleep and the prevalence of sleep deprivation, researchers have attempted to nudge people to sleep more and with a higher quality. Bessone et al. (2020) offered information to their participants about the importance of sleep and loaned sleep aids, such as sleep masks and fans, to them. The authors also used a more intense treatment in which they additionally offered monetary incentives to increase sleep duration. Both of these arms were successful in increasing sleep duration but not sleep efficiency. Handel and Kolstad (2017) ran an experiment where workers of a large firm were randomly invited to sign up for a sleep improvement program. The program produced statistically significant but economically small

improvement in sleep practices. Finally, Avery et al. (2020) provide a commitment device for college students to improve their sleep. Students set bedtime targets and sleep duration targets for themselves and the commitment contract incentivized them to meet these goals. The authors find that there is demand for these commitment devices and so students likely have time-inconsistent preferences.

Relative to these papers, I used milder nudges: in the treatment arms participants receive bedtime reminders without a commitment contract and in the other arms they received some information about sleep but no payment or devices. The results show that these mild nudges may not lead to changes in behavior and one might need to apply stronger interventions to find meaningful changes in behavior.

3.3 Theoretical Framework

This section describes the theoretical framework behind the experiment. This framework is based on the time use model of Jagnani (2020), and is extended for several time periods to allow for time-inconsistent preferences.

The model assumes that students face a time constraint every day: they have to choose between leisure (L_t), study time (H_t) and sleep (S_t). The hours spent on these activities each day have to sum up to 24 hours.

$$24 = L_t + H_t + S_t \tag{3.1}$$

Assume further that students' daily utility positively depends on their current leisure time (L_t), and education level (E_t), that is $u(L_t, E_t)$ and $\frac{\partial u(\cdot)}{\partial L_t} > 0$, $\frac{\partial u(\cdot)}{\partial E_t} > 0$. L_t is a choice variables in the current period, but the level of accumulated knowledge, E_t, depends on past choices. The evolution of the education level is governed by:

$$E_{t+1} = E_t + h(H_t, S_t) \tag{3.2}$$

, where $h(\cdot)$ is the education production function that is a positive function of study hours ($\frac{\partial h(\cdot)}{\partial H_t} > 0$) and sleep ($\frac{\partial h(\cdot)}{\partial S_t} > 0$). This latter relationship covers the productive sleep assumption.

The problem of the student is to allocate their time optimally in every period, subject to the time allocation constraint (Equation 3.1).[4] Leisure (L_t) directly increases students' utility in the current period. Study time (H_t) and sleep (S_t), on the other hand, are investment goods; they take time away from current leisure, while they increase future education level and thus future utility.

Students do not maximize their daily utility in isolation, rather, they maximize their lifetime utility. If students have time-consistent preferences, their lifetime utility is given by Equation 3.3 that assumes hyperbolic discounting. Students in this time-consistent case would be able to find their optimal sequence for each of the choice variables (leisure, study time and sleep) at time 0 and their optimal sequences would remain the same as time passes.

$$U_t = u_t + \delta u_{t+1} + \delta^2 u_{t+2} + \delta^3 u_{t+3} + \ldots = \sum_{j=0}^{T-t} \delta^j u_{t+j} \tag{3.3}$$

If, however, students' have time-inconsistent preferences, their lifetime utility would be given by Equation 3.4. In this case students' optimal sequences would change as time passes. This is because while the discount factor between any two

[4]The time allocation problem changes when students graduate. After graduation we can think of the time allocation problem as a problem to allocate time between sleep, work and leisure. Sleep would be a productive input for work and work would allow the individual to purchase consumption goods at the market that they would value besides leisure. The participants in the experiment are all undergraduate students in the spring semester. The experiment concluded before the graduation of that year, and so the participants' problem during the experiment remains that of the students and not of a working individual.

consecutive time period in the future is δ, just as in the time-consistent case, the time-inconsistent utility function has an additional parameter, $\beta < 1$, that further discounts every future period. Under time-inconsistent preferences, at $t = 0$ the student would solve their time allocation problem and find the optimal solution. But as the next period, $t = 1$, arrives its period utility ($u_{t=1}$) would receive more weight than before relative to the other future periods. This new weighting can lead to the abortion of the previously optimal sequences.

$$U_t = u_t + \beta \cdot \delta u_{t+1} + \beta \cdot \delta^2 u_{t+2} + \beta \cdot \delta^3 u_{t+3} + \ldots = u_t + \beta \sum_{j=1}^{T-t} \delta^j u_{t+j} \tag{3.4}$$

Time-inconsistent preferences can lead to a constant internal battle between the students' current and future desires. The students would be tempted to enjoy leisure today and would prefer to study and sleep in the future. This procrastinating behavior could lead to welfare loss because students may undertake less investment than would be optimal from the perspective of the planner.

The experiment aims to affect the students' optimization problem in two ways. In one of the experimental treatment arms, participants choose a desired bedtime for the next week and receive bedtime reminders half an hour prior to that. The text messages in this context would remind the student to their initial optimal solution and make the choice about disregarding their original plan salient, possibly increasing S_t.[5]

The other two experimental arms have the possibility to reduce the time it takes for students to fall asleep. For this, suppose that S_t is the time spent in bed, and an efficiency parameter governs how much of this time is actual sleep. Let's denote the

[5]Text reminders have been successfully used in the education domain. For example, they increased the chance that students filed their financial aid applications (Castleman and Page, 2016).

efficiency parameter as e_t and let $\tilde{S}_t = g(S_t, e_t)$ be the amount of efficient sleep that enters the education production function ($h(\cdot)$) as an input. Students still choose S_t, which is now the amount of time spent in bed, but its efficiency is affected by the exogenous parameter e_t that the informational treatment arms seek to increase.

If the efficiency parameter is higher ($e'_t > e_t$), students may choose more or less sleep time (S_t) than before, depending on the shape of their educational production function ($h(\cdot)$) and their utility function ($u(\cdot)$). In the education production function an exogenous increase in the efficiency parameter will increase the marginal product of sleep time (S_t) because $\tilde{S}_t$ is a positive function of the efficiency parameter. This means that students will substitute away from study time (H_t) towards sleep (S_t), leading to higher sleep time. Against this effect is an effect akin to a wealth effect. When the efficiency parameter is higher, the same level of utility can be obtained with less amount of sleep time (S_t) and so leisure (L_t) becomes relatively cheaper. Depending on the form of the utility function, the lower opportunity cost of leisure can increase the optimal level of leisure at the expense of sleep and/or study time. The overall effect of increasing the efficiency parameter depends on the sum of these opposing effects.

3.4 Experimental Design

The experiment was set up to answer two research questions: 1) whether various light touch interventions (nudges) can change participants' sleep patterns and 2) what is the relationship between sleep and cognitive outcomes. Three light touch interventions are tested: bedtime reminders, teaching a meditation technique in a video and providing tips for healthy sleep. Participants were randomized into one of these arms or the control group using a between subject randomization. During the

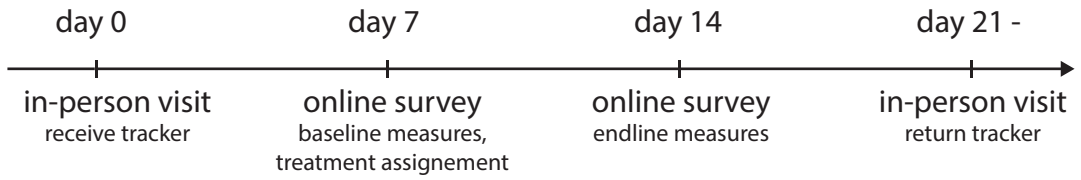


Figure 3.1: Timeline

Note: This figure shows the timeline of the experiment. The last individual in-person visit was scheduled any time after the 21st day of the experiment.

experiment sleep was monitored with fitness trackers and all participants had access to their own data.

The experiment was conducted in 2018 April at Duke University. 41 participants were recruited from the undergraduate student subject pool, who signed up to complete a three-week-long "Wellness Study". During the experiment, participants had two in-person meetings with a member of the research team and completed two online surveys in between the in-person visits. The generic timeline of the experiment is shown on Figure 3.1. Each participant followed their own timeline[6] and may have been on a shifted calendar schedule relative to other participants.[7]

On the first in-person meeting, participants consented, received a fitness tracker for the duration of the experiment, and filled out a demographic survey. Then a member of the research team demonstrated the usage of the fitness tracker and set up the tracker for the participant. Participants were incentivized to wear the fitness tracker every day during the period of the study and were asked to charge the tracker during their sedentary – but not sleep – periods. Wearing the tracker overnight or during sleep sessions was not a requirement of the study.

Approximately a week after the first meeting, participants received a link for an

[6]At the individual level, small deviations from this timeline was possible, for instance if the participant completed the online survey late, or was unable to make an in-person visit on day 21.

[7]The reason behind the staggered timing was twofold: first, the research team had a capacity constraint to conduct the in-person individual meetings with the participants, second, participants could sign up even after the first day of the experiment.

online survey. The survey collected information about participants' well-being, sleep and cognitive performance. The cognitive performance section consisted of two tasks. One of them was a grammatical reasoning test (Baddeley, 1968), in which participants saw sentences about a relationship of two letters (A and B) and then a letter sequence of those two letters. Participants had to decide if the statement is true for the given sequence. For example, they may saw the sentence: "A is followed by B" and the sequence "AB". In this sequence the letter "A" is indeed followed by "B", hence the sentence is true. Participants had 3 minutes to correctly mark as many sentences as they had time for. Because the length of the task is fixed (3 minutes), the outcome variables of interest in this task are the number of correct answers, the number of attempted questions and the fraction of wrong answers relative to the number of attempted questions. This latter fraction is referred to as the error rate in the latter sections.

The second task in the cognitive performance section was a short run memory task, called the recent-probes task (Jonides and Nee, 2006). A sample screenshot of the task is shown on Figure 3.2. Participants first saw 6 letters on the screen for 2 seconds. As the figure sows, the letters were displayed in 2 rows and 3 columns in the center. After the letters disappeared, participants were asked if they saw a given letter in the set. The task was repeated 25 times and had no time limit. The outcomes of interest in this task is the number of correct answers and the total response time.

After the cognitive tests ended in the first online session, participants were randomized into one of the four arms of the experiment. One group viewed a video in which participants learned about a breathing exercise. The breathing exercise is a meditation exercise that is considered to help to fall asleep.[8] Another group received hands-on tips about how to improve their sleep hygiene. The tips were written on the

[8]The shown breathing technique is referred to as the 4-7-8 breathing technique and is demonstrated by Andrew Weil, MD. in the video. The video is available at https://youtu.be/gz4G31LGyog.

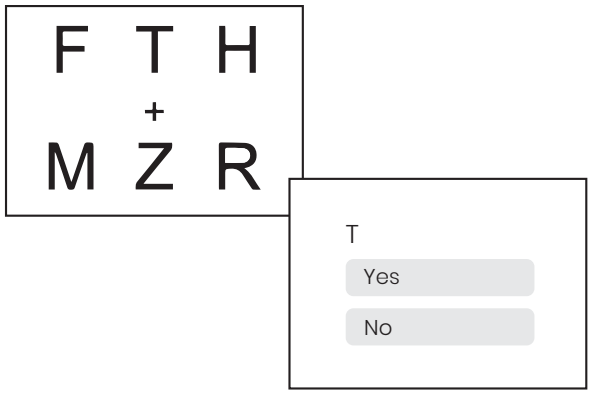


Figure 3.2: Recent Probes Task

Note: The figure shows sample screen shots of one round of the recent probes task. The first screen is shown for 2 seconds, the second screen is displayed until the participant responds.

Table 3.1: Sample Size by Treatment Arms

Treatment arm	N
SMS reminder	9
Meditation video	9
Information	10
Control	13
Total	41

screen and followed by a short quiz to test understanding. See Appendix 3.11.1 for the full list of tips. The third treatment group received bedtime reminders at a time of their choice for the next 7 days. These were pure bedtime reminders, as there were no incentives attached to meeting these targets. Finally, the control group received placebo information about the common cold and answered follow-up questions related to it. The sample size in each of the arms is shown in Table 3.1. The small sample size and the relatively large number of treatments foreshadow the noisy estimation results.

Participants completed the second online survey about a week after the previous one. The structure of the survey was similar to the previous one; the survey collected

information about self-reported well-being over the previous week and concluded with the cognitive test module.

About three weeks after the initial in-person meeting and a week after the second online survey, participants returned to the laboratory. During this meeting, they received their fixed participation fee and their incentive payments, returned the fitness tracker and completed a short leaving survey evaluating their experience with fitness trackers and the research team.

3.4.1 Sleep Measurement

Participants received a Fitbit Charge HR model[9] for the duration of the experiment to measure their sleep. This model is a wrist-type activity tracker that automatically measures sleep. It measures the start and end time of the sleep episode, the time spent in bed, as well as the number of awakenings during the sleep episode. In addition, it also provides estimates for minutes asleep and minutes awake during the sleep episode, but I disregard these two measures in the analysis due to Fitbits' generally poor accuracy on these measures (Montgomery-Downs et al., 2012).

Sleep is characterized by sleep duration and sleep quality. I will use the time in bed information as the upper bound of sleep duration. It is an upper bound because participants may spend some time to fall asleep (sleep latency) and they may wake up for some time during the night. Therefore, longer time in bed duration does not immediately mean healthier sleep if the increase is due to longer sleep latency[10] and longer wake times. Sleep quality answers the question about how well the participant slept and is different than sleep satisfaction. I use the number of awakenings

[9]This specific model has been validated by Lee et al. (2017) among young adults relative to actigraphs, that are the wrist-worn alternatives for polysomnography in medical research.

[10]Sleep latency, the time it takes to fall asleep, is approximately between 10-20 minutes for healthy adult men. Sleep latency is shorter for people suffering from sleep deprivation and narcolepsy, while longer for insomniacs. (Thomas and Anderson, 2013)

measure for sleep quality. The time spent in bed measure is the (S_t) choice variable in the theoretical model (Section 3.3), while the number of awakenings will provide information about the efficiency of the sleep (e_t): I will interpret more awakenings as a sign for less efficient sleep.

The National Sleep Foundation recommends 7-9 hours of sleep for young adults (Hirshkowitz et al., 2015b) and they also created a recommendation system for the objective sleep quality measures, including the number of awakenings. They classify 0-1 awakenings as appropriate for young adults (18-25 years) (Ohayon et al., 2017).

3.5 Summary Statistics

In this section I present the average characteristics of the sample and describe participants' sleep patterns before the treatment.

3.5.1 Sample Characteristics

41 undergraduate students signed up for participating in the three-week-long experiment. They were all undergraduate students enrolled at Duke University in the spring semester of 2018. Their characteristics and the relevant balance tests are shown in the Appendix in Table 3.16. Based on the table, the vast majority of the participants are female (83%) and study at the main liberal arts college of the university[11] (82%). On average participants were born in 1997 and study in their junior year. They did well on both of the two cognitive tests before the treatment: on average they correctly answered 34 statements on the grammatical reasoning test out of the average 39 attempted questions and completed the recent probe memory task almost without

[11]Duke University has three schools for undergraduate education: Trinity College of Arts & Sciences, Pratt School of Engineering and Duke Kunshan University. All participants belonged to either Trinity School or Pratt School.

an error.

Participants' treatment assignments were overall balanced as shown in the last three columns of the table. The columns test balance between one of the treatment arms and the control group. Imbalances in citizenship, year of birth and height will be addressed by controlling for these variables during the analysis.

3.5.2 How Do Participants Sleep Before the Treatment?

In this section I use two pieces of information about students' sleep. First, I will summarize the responses of the participants to the sleep survey module in the first online survey, then I will summarize the sleep data measured by the fitness trackers.

Sleep survey responses

Participants answered several survey questions about their sleep during the online sessions. Based on the first online survey, some participants had sleep issues. Figure 3.3 shows the histogram of the number of nights on which participants had trouble with their sleep in the past 7 days. The median participant reported to have sleep issues on 1 night, and about 20% of the participants reported sleep problems on more than 4 nights on the previous week. On a separate question, approximately half of the participants reported feeling sleepy during the day on at least 3 days out of the past 7 days.

What can be behind these reported sleep difficulties? Some factors are environmental. Based on further survey questions, roughly half of the participants share their sleep environments with others: 54% have at least one roommate and about 40% of the participants sleep in a bed with another person with some frequency. Other factors are related to smoking and caffeine consumption. None of the participants reported smoking, but the majority of them consumed coffee over the past week, and

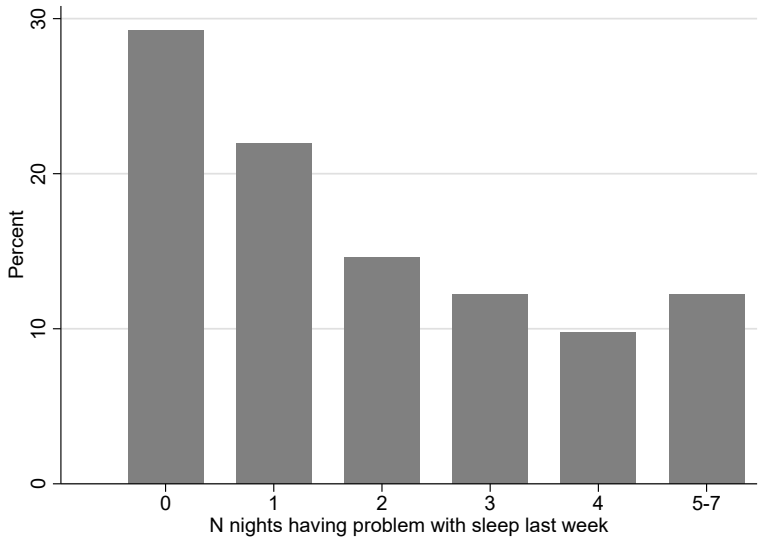


Figure 3.3: Histogram of Number of Nights Participants Had Sleep Problems

Note: This graph is the histogram of the number of nights participants report having problems with their sleep over the past 7 days. The last category pools those who had sleep problems on 5-7 nights.

14% consumed more than 2 cups daily. The third factor that the survey touched upon was mental health (stress and negative emotions). About 30% of the participants reported experiencing all of the five negative emotions in the survey (lonely, overwhelmed, exhausted, stressed, feeling lost) over the past week, and another 30% reported experiencing 4 out of the 5 negative emotions.

Although participants face several challenges to sleep well, especially on the mental health front, they are not without help. About half of the sample reports using some form of sleep aids, such as sleeping pills, earplugs and sleeping masks.

Baseline measured sleep data

Self-reported data about sleep is noisy, because people lose consciousness during sleep (Ohayon et al., 2017). For this reason I measured participants' sleep patterns with fitness trackers. In this part I summarize the baseline sleep data recorded by the trackers. The following observations are described: 1) the vast majority of the sleep

episodes are night sleep episodes that start around midnight and results in about 7-8 hours spent in bed and 2) students' sleep patterns vary over the course of the week leading to inconsistent sleep patterns. These two observed sleep patterns are not unusual, Avery et al. (2020) find similar patterns when studying college students in the US and UK.

To demonstrate the first observation about the sleep episodes, Figure 3.4 shows all the sleep episodes (n = 289) recorded by the trackers during the baseline week. On the x-axis is the hour when the sleep episode started, while the y-axis shows the amount of time spent in bed during the sleep episode in minutes. During the time of the experiment the sun set time was around 19:00 and the sun rose around 7:00 in the morning, so sleep episodes that fell into the 19:00-7:00 window are defined as night sleep episodes (marked with the dots), while sleep episodes that started and ended outside of this interval are defined as daytime sleep episodes (marked with the triangles).

Based on the figure, most of the sleep episodes were night sleep sessions. The majority of these night sleep sessions started between 23:00 and 2:00 lasted for about 7-8 hours.

The graph also hints that students have variable sleep patterns – as noted by the second observation. About half of the night episodes fall between the two dashed horizontal lines that mark the recommended 7-9 hours window for sleep duration, and consequently, the rest of the episodes fall outside of the recommended window. While the time spent in bed measure is only an upper bound for actual sleep time, episodes that were strictly below the 7 hours threshold could be considered as short sleep episodes. This suggests that on some nights students may sleep short and on other nights they may try to compensate leading to longer than 8 hours long sleep episodes.

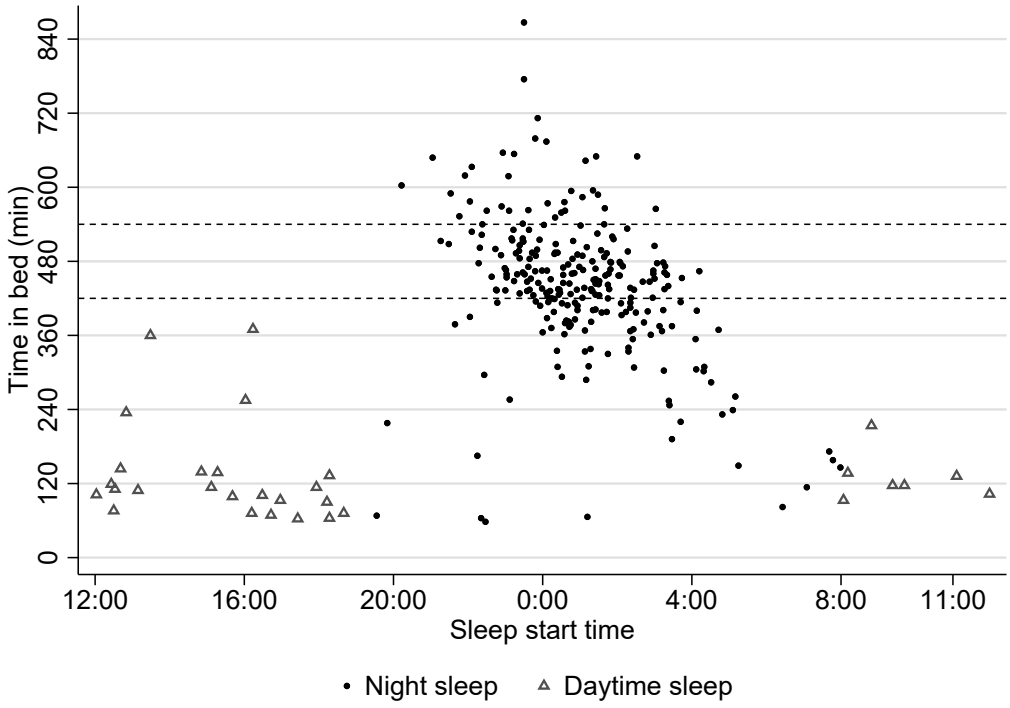


Figure 3.4: Time Spent in Bed by Sleep Start Time During the Baseline Week

Note: This graph shows each recorded sleep episodes during the baseline week. The x-axis represents the start time of the sleep session and the y-axis shows the associated time spent in bed in minutes. The horizontal dashed lines represent the 7 and 9 hour threshold set by the recommendation of the National Sleep Foundation for young adults (Hirshkowitz et al., 2015a). Night sleeps are defined as sleep episodes at least partially overlapping with the 19:00 - 7:00 window.

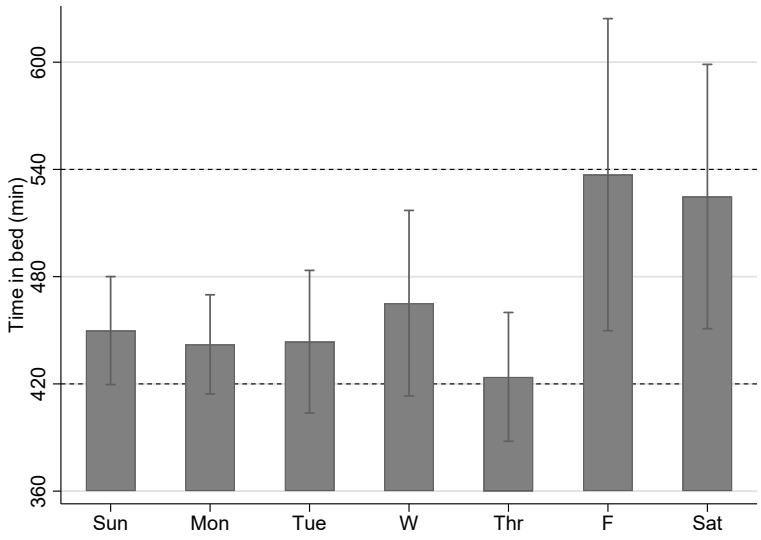


Figure 3.5: Time Spent in Bed by Day of the Week During the Baseline Week

Note: This graph shows the average time spent in bed at night across participants by the days of week during the baseline week. The horizontal dashed lines represent the 7 and 9 hour threshold set by the recommendation of the National Sleep Foundation for young adults (Hirshkowitz et al., 2015a). In case multiple sleep session have been recorded for a night from the same participant, the time spent in bed measures were added up. Wings mark the 95% confidence intervals using heteroskedasticity-robust standard errors clustered at participants level.

The variable sleep pattern over the week is what Figure 3.5 studies. The Figure shows the average time spent in bed by the day of the week during the baseline week.[12] The histogram displays a strong weekend effect: participants on average spend longer times in bed on Friday and Saturday nights than on other nights, especially relative to Thursday nights, when they seem to sleep short. Appendix Figure 3.11 reveals that the variability mostly stems from later wake-times over the weekends and later bedtimes on Thursday evenings. From this respect, Thursday nights are the most unfavorable: students go to bed later but cannot sleep in on the next morning.

It is not just the average sleep pattern that varies within the week by a large extent. In the Appendix Panel E of Table 3.16 shows that during the baseline week,

[12]To prepare this histogram, sleep episodes during the night from the same person and night have been added up, meaning that long awakenings at night that would result in multiple shorter sleep episodes do not pull the average time spent in bed measure down.

participants shifted their bedtime on average by 195 minutes over the week, and their wake-time by 201 minutes.

Consistent sleep pattern schedules are the characteristics of healthy sleep patterns and are associated with better grades (Hershner, 2020), and because this is the margin where the participants fell short, there may be particularly large gains for interventions that could limit the variability.

Missing measured sleep data

Sleep data is obtained with the fitness trackers under three conditions: 1) participants wore the fitness trackers during their sleep episodes 2) the tracker was charged and functioning 3) the tracker was connected to the Fitbit application via Bluetooth connection. The first two conditions were required for measurement, while the last one was required for data transfer. If either of the conditions failed, the sleep episodes were not available for the study. Note, observing missing data does not automatically mean that the participants did not comply with wearing the device; technical issues with the trackers could lead to missing data even if the participant wore the tracker at night.

To judge the severity of the missing data problem, Figure 3.6 shows the estimated probability density function of the percentage of nights with missing sleep data[13]. Based on this figure, the compliance in the study was high, only a few participants had sleep data for less than 80% of the experimental nights.

Having shown that sleep data is available for the vast majority of nights for almost all participants, Table 3.2 test if the rate of missing sleep data is differential across treatment assignment. The table reports the results of a regression that regresses the

[13]Because some participants took part in the study for more than 21 days due to scheduling conflicts, using the number of days without sleep data is not appropriate. Instead I use the percentage of nights with missing sleep data during the entire experiment.

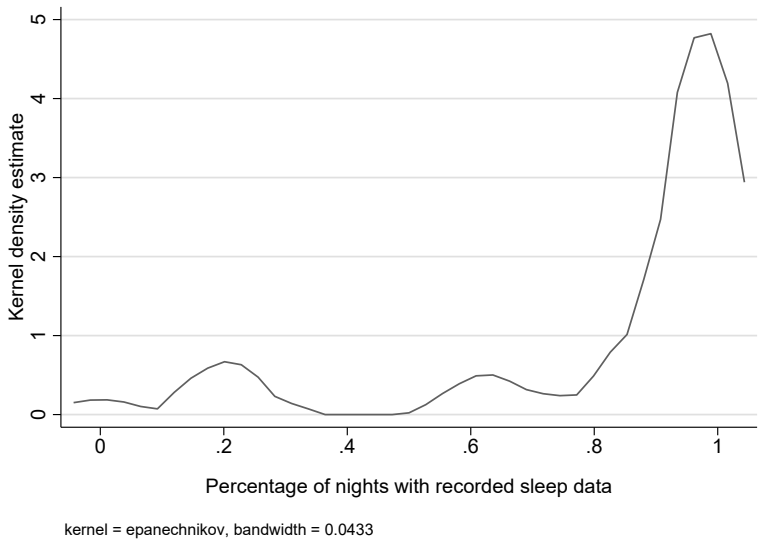


Figure 3.6: Kernel Density Estimate for Percentage of Missing Data

Note: This graph shows the kernel density estimate for missing sleep data at the individual level. The measure for missing data is the percentage of nights during the experiment that had no night sleep. Missing sleep data could results from three causes: if participants did not wear the fitness tracker at night, if the tracker's battery was not charged or if the recorded data was not synchronized with the application of the tracker.

Table 3.2: Missing Sleep Data is Not Predicted by Treatment

	Percentage of nights with missing sleep data
Information	0.072
	(0.131)
Meditation video	0.037
	(0.135)
SMS reminder	0.162
	(0.112)
N clusters	41
N	41
R2	0.048

Note: This table reports the result of regressing the percentage of nights with missing sleep data on treatment assignment. The reference category is the control group. Missing sleep data could results from three causes: if participants did not wear the fitness tracker at night, if the tracker's battery was not charged or if the recorded data was not synchronized with the application of the tracker. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

percentage of nights with missing sleep data on treatment assignment indicator. The estimated coefficients are not statistically significant at the 10% level, albeit large, especially for the SMS reminder group.

While the percentage of nights with missing data is not statistically different for the treatment groups, it is not accurate to hold the missing at random assumption. Appendix Table 3.17 shows that among the demographic characteristics, year of birth and weight significantly predicts the level of missing sleep data. Younger participants and participants with higher weight are less likely to have nights with missing data. Therefore, the rest of the analyses will rely on a conditional missing at random assumption and use year of birth and weight as control variables in the specifications.

3.6 Analysis

3.6.1 Treatment Effect Analysis

This section analyzes the effect of the three behavioral interventions on students' sleep and on students' cognitive outcomes. Because of the design, the interpretation of the first question – what is the effect of the selected behavioral interventions on sleep – is causal. However, the second question – what is the effect of sleep on cognitive performance – is answered through an encouragement design, where the identification comes from shifting sleep patterns with the randomized nudges.

Treatment Effects on Sleep

To analyze the treatment effects on sleep, I first estimate the treatment effects using a night level analysis. Then, I use aggregated sleep outcomes at the week level to understand how participants' sleep patterns have changed during the treatment week relative to the baseline week.

In the night level analysis I regress the sleep outcomes of the baseline and treatment week on treatment assignment indicators (the reference group is the control group), interactions between the treatment week (week 2) and treatment indicators and various control variables. Among the control variables are the day of the week indicators and the "additional control variables" that include indicators for gender, race, US citizenship, birth year, as well as participants' weight and height, and the date of their first in-person visit. The non-standard variables are added to mitigate the concerns about imbalances (US citizenship, height) and to satisfy the conditional missing at random assumption for the sleep measurement (weight, birth year). The exact specification is:

$$Y = \alpha + \sum \beta_{1,j} T_j + \sum \beta_{2,j} T_j \times \mathbb{1}(\text{week 2}) + \gamma \mathbf{X} + \lambda_{dow} + \varepsilon \tag{3.5}$$

, where Y is the night level sleep outcome variable during the baseline and treatment week, T_j represent the treatment indicators, $\mathbb{1}(\text{week 2})$ is an indicator for the treatment week, $\mathbf{X}$ stands for the vector of additional control variables and λ_{dow} are the fixed effects for the days of the week. The reference category is the control group during the baseline week. The coefficients of interests are the β_2 coefficients that are reported in Table 3.3.

Based on the estimates in Table 3.3, the treatments have no statistically significant effects on the outcome variables. The estimates are sometimes large in magnitude, – for instance the SMS reminder and the meditation video decreased the amount of time spent in bed with large magnitude and decreased the number of awakenings too. Table 3.18 in the Appendix reports the estimation results without using the additional control variables. The magnitude and the sign of the estimates are sensitive to the addition of the control variables, an issue that also stems from the small sample size.

As an alternative approach to measure the effect of the interventions on sleep, I use weekly aggregated sleep measures as outcome variables. The rationale is to mitigate missing data issues and to detect changes in the weekly sleep patterns. For the aggregated measures I opted to use the median values of the sleep outcome variables over the course of a week as opposed to the average values, because median values are less sensitive to the missing data problem.

The empirical specifications for the treatment effect estimations are:

$$\Delta Y = \alpha + \sum \beta_j T_j + \gamma \mathbf{X} + \varepsilon \tag{3.6}$$

, where ΔY is the change in the outcome variable between the treatment week and

Table 3.3: Treatment Effect on Night Level Sleep

	Treatment week: night level analysis			
	(1) Time in bed (min)	(2) Awakenings (N)	(3) Bedtime (hr)	(4) Wake-time (hr)
SMS reminder × week 2	-25.536	-0.608	0.233	0.111
	(25.427)	(0.704)	(0.263)	(0.250)
Meditation video × week 2	-43.895	-0.290	0.088	0.119
	(46.044)	(0.645)	(0.470)	(0.227)
Information × week 2	4.613	-0.065	0.273	0.088
	(28.372)	(0.462)	(0.362)	(0.357)
Control × week 2	-9.672	0.589	-0.063	-0.185
	(21.101)	(0.452)	(0.249)	(0.215)
Main treatment indicators	✓	✓	✓	✓
Additional controls	✓	✓	✓	✓
Day of the week indicators	✓	✓	✓	✓
Control mean	427.390	1.909	1.027	8.944
Pre-treatment avg level	460.211	2.796	0.234	9.366
N	485	485	485	485
N cluster	39	39	39	39
R2	0.118	0.378	0.158	0.250

Note: This table reports the treatment effects of the three behavioral interventions on participants' sleep at the night level. The last two columns estimate the effect on measures of consistent sleep schedule. Bedtime and wake-times are the are expressed in fraction of hours, so the control group during the treatment weak has a median bedtime at about 1:01 am (1.027 = 1 am + 0.027 * 60) and their average respective median wake-time is at 8:57 am. The set of additional control variables include indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. All specifications include day of the day indicator variables. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

the baseline week. All Y variables are weekly aggregated outcomes. T_j represents the treatment indicators and $\mathbf{X}$ is a vector of additional controls. The specification does not include day of the week effects because the outcomes are aggregated over the treatment week.

The results of this estimation are shown in Table 3.4. Column 1 reports the effects on the change in the median amount of time spent in bed between the treatment and baseline week. Column 2 shows the estimates for the change in a sleep quality measure, the median number of awakenings during nights. The rest of the columns estimate the effect of the interventions on the consistency of the sleep schedules. The median bedtime and wake-times over the week are again adjusted for nights that have multiple sleep episodes due to long awakenings. For each night, the bedtime is the earliest bedtime and the wake-time is the latest wake-time of the night. The weekly medians are calculated over these nightly measures and then differenced to get the outcome variables of column 3 and 4. In the last column the outcome variable is the change in the longest and shortest time spent in bed recorded over the week between the treatment and the baseline week.

In line with the night level estimation (Table 3.3) this weekly aggregated estimation finds no statistically significant effects, except for the later bedtimes in the SMS reminder group, which is an economically tiny effect. Nevertheless, the estimated coefficient for the SMS reminder intervention still decreases the amount of time spent in bed by a large magnitude. For robustness check, the estimation is also performed without the additional controls. The results are presented in Table 3.19 in the Appendix. Again, the magnitudes are sensitive to the addition of control variables, especially in column 5.

Taken together, the interventions did not affect the sleep outcomes in a statistically significant way. This is true for the daily and weekly aggregated level outcomes as

Table 3.4: Treatment Effects on Sleep Pattern Changes

	$Y_{\text{treatment week}} - Y_{\text{baseline week}}$		
	(1) Median time in bed (min)	(2) Median N awakenings	(3) Median bedtime (min)
SMS reminder	-56.169	0.851	1.319*
	(40.635)	(0.995)	(0.647)
Meditation video	3.332	-0.210	0.740
	(42.793)	(0.828)	(0.633)
Information	-4.841	-1.064	0.482
	(40.593)	(0.923)	(0.610)
Additional controls	✓	✓	✓
Control mean	-4.417	0.167	-0.228
Pre-treatment avg level	440.024	2.280	24.919
N	38	38	37
R2	0.458	0.593	0.646

	(4) Median wake-time (min)	(5) Maximum difference in time in bed (min)
SMS reminder	0.411	66.635
	(0.557)	(281.835)
Meditation video	0.382	-52.453
	(0.573)	(203.893)
Information	0.242	-130.927
	(0.478)	(185.949)
Additional controls	✓	✓
Control mean	-0.317	1.417
Pre-treatment avg level	8.475	307.780
N	37	38
R2	0.588	0.506

Note: This table reports the treatment effects of the three behavioral interventions on the change of participants' night sleep patterns between the treatment week and the baseline week. The outcome variables are aggregated over a week long period to account for the weekly seasonality. The last three columns estimate the effect on measures of consistent sleep schedule. Median bedtime and wake-times are the median time when the first night sleep session started and the last night sleep session ended respectively over the treatment period. The maximum difference in time spent in bed measures the difference between the longest and shortest recorded night sleep over the treatment period. The set of additional control variables include indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

well. Despite of the imprecision, the magnitudes are meaningful and large relative to the control group means.

Treatment Effects on Cognitive Outcomes

Having shown that the behavioral interventions have no statistically significant effects on sleep, there is no ground for treatment effects on cognitive outcomes. This is confirmed with a reduced form estimation in this section.

Participants completed the cognitive modules once during the first online survey before the treatment, and for the second time at the end of the treatment week. The difference between these two measurements will be the outcome variables in the treatment effect estimation that follows the specification of Equation 3.6. Table 3.5 shows the estimated reduced form effects of the treatments on the test scores and on some more subtle outcome variables: the number of attempted questions and the error rate during the grammatical reasoning test, and the total response time during the short term memory test. Overall, while some of the estimated treatment effects are large, – for example the effect of the SMS reminder on the score increase – the treatments have no statistically significant effects on the cognitive outcome measures.[14]

Treatment Effects on Self-reported Sleep Outcomes

Having shown that the treatments did not affect measured sleep outcomes or cognitive outcomes, this part shows that the treatments had no effect on self-reported sleep outcomes, other than lowering the average occurrence of self-reported sleep problems among the information group.

[14]One estimate is statistically significant at the 10% level, which is likely driven by multiple hypothesis testing effects.

Table 3.5: Treatment Effects on Cognitive Outcomes

	Grammatical reasoning test		
	Δ score	Δ attempted	Δ error rate
SMS reminder	7.422	13.015*	0.045
	(4.843)	(7.386)	(0.046)
Meditation video	-1.219	11.389	0.009
	(5.986)	(10.304)	(0.056)
Information	4.121	-3.824	-0.055
	(4.932)	(8.633)	(0.036)
Additional controls	✓	✓	✓
Control mean	9.692	10.615	-0.004
Pre-treatment avg level	33.976	39.049	0.135
N	39	40	39
N clusters	39	40	39
R2	0.474	0.491	0.794

	Memory test		Total
	Δ score	Δ RT	Δ score
SMS reminder	-1.716	10.262	5.705
	(1.097)	(18.997)	(5.469)
Meditation video	-0.679	18.896	-1.898
	(1.486)	(26.196)	(6.462)
Information	0.239	-38.058	4.360
	(0.865)	(23.954)	(4.870)
Additional controls	✓	✓	✓
Control mean	0.846	-0.213	10.538
Pre-treatment avg level	23.683	66.370	57.659
N	39	40	39
N clusters	39	40	39
R2	0.427	0.509	0.435

Note: This table reports the reduced form treatment effects of the three behavioral interventions on the change of participants' cognitive test scores between the treatment week and the baseline week. The grammatical reasoning test follows (Baddeley, 1968) and the related outcome of interests are the test scores, the number of attempted questions and the error rate, which is the number of mistakes over the number of attempted questions. The short-run memory test is an adapted version of the recent probes task (Jonides and Nee, 2006). In this task the test score and the overall response time (RT) are the outcome of interest. The set of additional control variables include indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

Following Equation 3.6 as the main specification, I regress the change in reported sleep variables between the second and first online survey on the treatment dummies and additional controls. Table 3.6 reports that participants' self-reported sleep were not affected by the treatment. The exception is the information treatment. Participants in this treatment arm reported in the second online survey on average 1.5 lower category of their frequency of sleep problems in the past 7 days than what they reported in the first online survey.[15] Since this group did not change their measured sleep behavior, I attribute this effect to desirability bias.

Did the Participants Follow the Suggestions of the Treatments?

The previous sections showed that the treatments left the measured sleep variables, the cognitive outcomes and the reported sleep variables intact, although the estimates are imprecise. The question hence arises if this is because participants did not follow the advice they received in the treatments or they did, but those had no effects.

First, the meditation video and the sleep hygiene information suggested certain behaviors, such as practicing meditation and limiting coffee intake. The online surveys collected information on several of these suggested behaviors, thus it is possible to test if participants' behavior have changed due to the treatments. Table 3.20 in the Appendix shows the treatment effects on the related self-reported variables, again following the specification in Equation 3.6.

The table sporadically shows significant effects at the 10% level, but there is no clear, convincing pattern. Importantly, the meditation video did not increase the usage of meditation as a sleep aid, and the information treatment did not reduce the chance that the participant will drink coffee daily. For these reasons, the information and meditation video interventions do not appear to be successful in leading to

[15]Note that while the categories are in 1-night-long bins until the 4th nights, having problems 5-7 nights past week forms one category.

Table 3.6: Treatment Effects on Self-reported Sleep

	(1) Δ sleep problem reported category past week	(2) Δ above median sleep problems past week	(3) Δ daytime sleepiness min 3 days p.w.
SMS reminder	-1.360	0.100	0.303
	(0.883)	(0.374)	(0.193)
Meditation video	-0.418	-0.011	0.133
	(0.897)	(0.438)	(0.298)
Information	-1.462**	-0.221	0.029
	(0.569)	(0.314)	(0.277)
Additional controls	✓	✓	✓
Control mean	0.077	-0.077	-0.154
Pre-treatment avg level	1.878	0.707	0.244
N	39	39	39
N cluster	39	39	39
R2	0.520	0.458	0.431

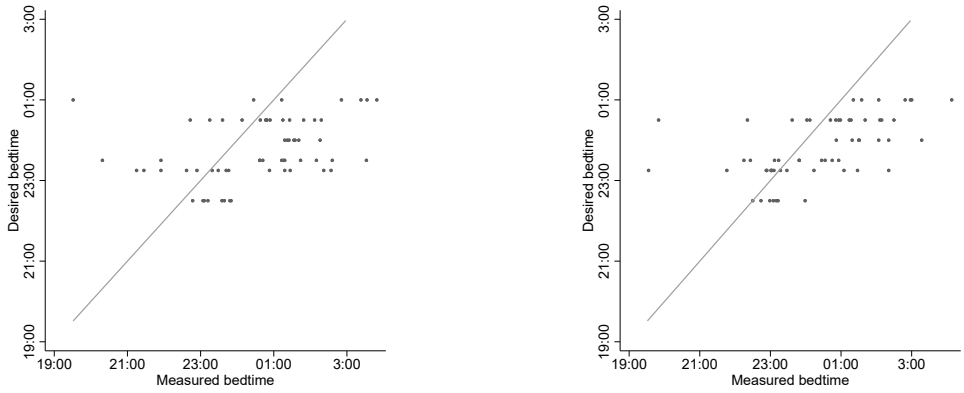
Note: This table reports the treatment effects of the three behavioral interventions on the change in participants' answers on the sleep survey between the treatment week and the baseline week. The outcome variable in column 1) is the change in reported number of nights with sleep problems, measured as categorical variable. The categories are in 1 night long bins until 4 nights, and then having problems on 5-7 nights past week forms one category. In column 2) the outcome variable is the change in whether the participant has above median number of sleep problems. In column 3) it is the change in the indicator variable of being sleepy at least 3 days during the day. The set of additional control variables include indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

behavioral changes.

Second, it is also possible to evaluate the success of the bedtime reminders by comparing participants' desired bedtime with their actual bedtime.[16] Figure 3.7, Panel a) shows a scatter plot of the treatment week nights for the SMS reminder group. The x-axis measures the actual measured bedtime using the data from the fitness trackers, and the y-axis measures participants' desired bedtime that participants chose during the first online survey. Out of the 62 treated nights on the plot, participants went to bed before their chosen bedtime on 13 nights. These successful nights are to the left of the 45 degree line and make up about 21% of the treated nights. Panel b) shows the same comparison but for the baseline week, when no reminders were sent. In other words, Panel b) shows how challenging the bedtime reminders were for the participants. This time, 28% of the nights are to the left of the 45 degree line. The two corollaries are that first, participants indeed targeted a bedtime that was not obvious to meet - a sign that they are time inconsistent to some extent. The second corollary is that the reminders were not effective in pulling participants' bedtime earlier relative to the baseline week. During the baseline week participants would have met their goals 28% of the nights, but during the treatment week this number is 21%.

Taken together, none of the treatments seem to have generated the intended actions. This finding can reconcile why the interventions had no effects on participants' sleep and cognitive outcomes.

[16]The bedtime reminders were sent on 7 nights of the treatment week. The reminders arrived at the same time, half an hour before the desired bedtime. Participants could not adjust the timing of the reminder after their initial timing choice.



(a) Treatment week (b) Baseline week

Figure 3.7: Desired and Actual Bedtime of the Bedtime Reminder Group

Note: This figure shows the desired and the actual bedtime of participants in the bedtime reminder arm during the nights of the treatment week (Panel a) and the baseline week (Panel b) nights. The reminders were only sent during the treatment week. The 45 degree line serves as a comparison; dots on the right of the line represents nights for which a participant missed their target bedtime.

3.6.2 Panel Data Analysis of the Relationship Between Sleep and Cognitive Outcomes

The previous sections revealed that the behavioral interventions had no effect on participants' sleep and because of the missing first stage, the treatments did not affect the cognitive outcomes either. This section relies on panel data methods to estimate the effect of sleep on cognitive measures without using the randomized experiment.

For these methods, I regress cognitive test outcomes on sleep variables measured on the night before the cognitive tests were taken. That is, if the tests were taken on Wednesday, the sleep variables in the regression are the ones that are measured during the night that ends on Wednesday morning, which is Tuesday night. This strategy leads to a smaller sample size because several participants miss recording sleep data on the night before the tests. Therefore these analysis are exploratory and should be interpreted accordingly.

The first estimation pools the information at the time of the two online surveys

and controls for the wave of the data collection. The estimated regression is:

$$Y = \alpha + \beta_1 * S + \beta_2 * Q + \gamma * \mathbf{X} + \delta * w + \varepsilon \tag{3.7}$$

The outcome variables are coming from the cognitive modules of the two online surveys. S is the time in bed measure and Q refers to sleep quality measured as the number of awakenings at night. X is a vector of additional control variables, and just as before it contains indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. w is the wave indicator that equals to one if cognitive measures are from the second online survey.

Table 3.7 reports the corresponding estimates. The coefficients are again overall statistically insignificant. The exception is the coefficient on the number of awakenings in column 5. This result suggests that those who had lower quality sleep (had more awakenings at night) took longer to complete the short-run memory task. This additional effort did not pay off; as column 4 shows, despite of the longer response time, they achieved approximately the same score as those who had better sleep quality.

The estimation results in Table 3.7 are prone to omitted variable bias, especially at the individual level. Therefore, in the next analysis I add individual fixed effects and compare the cognitive outcomes within the same individual.

The estimated regression with individual fixed effects is:

$$Y = \alpha + \beta_1 * S + \beta_2 * Q + \delta * w + \lambda_i + \varepsilon \tag{3.8}$$

, where as before, the outcome variables are the cognitive outcome variables, S is the amount of time spent in bed on the previous night, Q is the number of awakenings

Table 3.7: Pooled Estimation of the Effect of Sleep on Cognitive Outcomes

	Grammatical reasoning test		
	(1) Score	(2) N attempted	(3) Error rate
Time in bed prev. night	0.016	-0.026	-0.000
	(0.013)	(0.026)	(0.000)
N awakenings prev. night	0.257	-0.144	-0.004
	(0.343)	(0.307)	(0.005)
Personal controls	✓	✓	✓
Wave indicator	✓	✓	✓
Control group mean	39.077	45.077	0.120
Control group SD	13.564	13.850	0.178
N	59	60	59
N clusters	33	33	33
R2	0.694	0.534	0.493

	Memory test		Total
	(4) Score	(5) RT	(6) Score
Time in bed prev. night	-0.001	-0.011	0.015
	(0.001)	(0.036)	(0.013)
N awakenings prev. night	-0.051	3.309***	0.206
	(0.038)	(0.461)	(0.358)
Personal controls	✓	✓	✓
Wave indicator	✓	✓	✓
Control group mean	24.115	65.727	63.192
Control group SD	1.336	13.336	13.926
N	59	60	59
N clusters	33	33	33
R2	0.309	0.330	0.687

Note: This table reports the pooled OLS estimates of the effects of sleep on cognitive outcomes. The grammatical reasoning test follows (Baddeley, 1968) and the related outcome of interests are the test scores, the number of attempted questions and the error rate, which is the number of mistakes over the number of attempted questions. The short-run memory test is an adapted version of the recent probes task (Jonides and Nee, 2006). In this task the test score and the overall response time (RT) are the outcome of interest. The set of personal control variables include indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

on the previous night and w is the wave indicator. λ_i represents the individual fixed effects.

Table 3.8 shows that the previous result that lower quality sleep is associated with longer response time during the memory test (Table 3.7, column 5) is not robust to the addition of the individual fixed effects. On the other hand, the table reports other estimates that are statistically significant at the 10% level. Based on those, lower quality sleep increases the number of attempted questions on the grammatical reasoning test, and increases the error rates of the participants.

Students were incentivized to score high on the tests, but error rates or the number of attempted questions were not incentivized in any way. Because on average those with lower sleep quality achieve the same score as those with better sleep quality, it may seem that those who have poorer sleep quality are not worse off in the end. This view, however, would not take into account the relationship between the number of attempted questions and the test scores. Attempting an extra question takes time away from previous questions that may have benefited from an additional few seconds. But attempting more questions could also increase the chance of scoring higher. Either way, attempting more questions is riskier and the seemingly increased risk taking behavior for those who had lower sleep quality fits into the findings of (Venkatraman et al., 2011), who show that sleep deprivation is positively associated with risk seeking behavior.

3.6.3 Summary of the Analysis

The treatment effect analysis showed that the behavioral interventions were not successful in changing students' overall sleep patterns. Because participants' sleep were not affected, the treatment effects on cognitive outcomes are also statistically insignificant. Further analysis revealed that the insignificant treatment effects are due to the

Table 3.8: The Effect of Sleep on Cognitive Outcomes Using Individual Fixed Effects

	Grammatical reasoning test		
	Score	N attempted	Error rate
Time in bed prev. night	-0.007	-0.014	-0.000*
	(0.007)	(0.009)	(0.000)
N awakenings prev. night	0.995	1.616*	0.012*
	(0.844)	(0.849)	(0.006)
Wave indicator	✓	✓	✓
Individual FE	✓	✓	✓
Control group mean	39.077	45.077	0.120
Control group SD	13.564	13.850	0.178
N	54	56	54
N clusters	27	28	27
R2	0.890	0.724	0.911

	Memory test		Total
	Score	RT	Score
Time in bed prev. night	-0.001	-0.001	-0.008
	(0.001)	(0.015)	(0.006)
N awakenings prev. night	-0.161	-1.269	0.835
	(0.113)	(1.052)	(0.829)
Wave indicator	✓	✓	✓
Individual FE	✓	✓	✓
Control group mean	24.115	65.727	63.192
Control group SD	1.336	13.336	13.926
N	54	56	54
N clusters	27	28	27
R2	0.629	0.395	0.895

Note: This table reports the treatment effects of the three behavioral interventions on the change of participants' cognitive test scores between the treatment week and the baseline week. The set of additional control variables include indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively..

lack of behavioral change in response to the interventions: participants did not seem to act upon any of the interventions. These treatment effects are noisily estimated and would benefit from a larger sample size, especially because some of the estimates were large in magnitude. As an alternative strategy, the individual fixed effects analysis shows that lower quality sleep on the night before the tests is associated with higher number of attempted questions and higher error rate on the grammatical reasoning test. The time in bed measure has no effect on cognitive outcomes.

3.7 Discussion

The results show that the behavioral interventions had no treatment effects on sleep and this is because the interventions did not induce changes in participants' behavior. In this section I discuss why could it be the case that the interventions were not effective in this experiment.

One unmistakable candidate is the small sample size and thus low power of the study. Some of the estimates, especially on the time in bed measure outcome, are large in magnitude, yet still statistically insignificant at the 10% level. Further data collection and a larger sample size could mitigate the issue of power. The current study could serve as the basis of sample size calculations and a blueprint for the design.

The second reason is that the interventions may affect sleep outcomes, but not the ones that the current analysis used. Fitness trackers provide a subset of those sleep measures that are available with polysomnography and even the ones that are provided are less reliable. Two of the available measures, time spent asleep and time spent awake, would have been critical to use as outcome variables, yet at the time of the data collection these were considered highly inaccurate. Upon analyzing the data

these variables indeed seemed unrealistic and thus were removed from the analysis. If the interventions affect sleep variables that were not analyzed in this study, for example the time asleep measure, the current paper would not be able to find those effects.

Next, it may be the case that the interventions were too light touch to have any effects. Other experiments that were able to shift participants' sleep patterns applied much stronger interventions: Bessone et al. (2020) required people to use the sleep room at the field experimental site and paid for participants to do so, while Avery et al. (2020) provided incentivized commitment contracts. The interventions of the current study had no such elements. The bedtime reminders were not attached to incentives and participants were not paid to sleep more. The idea that light touch nudges may not be enough to change sleep is supported by similar experiments, such as Blunden et al. (2012) in the domain of sleep information treatments and Tavernier and Adam (2017) in the domain of text reminders.

Finally, the population of the experiment is a particular one. The experimental results could not detect that the nudges affected participants in the sample, but it may be that the interventions would have an effect among other groups, perhaps those who have worse sleep patterns.

Taken together, while probably all of these factors play some role in explaining why the treatments had no effects, the sample size limitation is the one that would be high priority to alleviate. Estimating a precise zero effect would be valuable to understand the possible scope of light-touch sleep interventions.

3.8 Conclusion

This paper studies two research questions: 1) What is the effect of various light touch behavioral interventions on sleep? and 2) What is the effect of sleep on cognitive outcomes?. To answer these questions the paper reports on a small scale experiment. Participants in the treatment arm received gentle nudges (bedtime reminders, meditation video, information about sleep hygiene) that were hypothesized to be able to change sleep patterns.

The results overall show that the interventions did not change participants' sleep or cognitive performance in a way that is detectable with the current sample size. The panel method on the other hand suggests that higher sleep quality reduces the percentage of wrong answers out of the attempted questions. This result supports the idea that sleep is linked to cognitive performance.

Overall, there are several potential factors that could explain why the treatments did not affect participants' sleep and as a consequence participants' cognitive outcomes. In light of these, the first order issue for estimating the effect of the behavioral interventions on sleep is to increase the sample size. Larger studies could provide more precise estimates and inform the practitioners about the effectiveness of these interventions. If, however, the goal is to estimate the effect of sleep on cognitive outcomes in a field setting, it seems reasonable to use stronger incentives or natural experiments (for example daylight saving time changes) to induce large changes in sleep. Relying on larger changes in sleep, in addition to using more advanced trackers than the ones used for this study, may be a fruitful avenue in the future.

Conclusion

This dissertation showed in three chapters that the tools of behavioral economics can provide important insights to labor economics. The first chapter uncovered two likely unintended effects of a proposed policy (wage transparency policy) with an online experiment. The policy can increase workplace hostility, and change how worker sort to workplaces. These patterns can be reconciled with the theory of social preferences.

The second chapter contributes to labor economics by showing that there is racial bias when evaluators choose whom to nominate for an award in the police. We find that part of the racial bias is due to homophily, the tendency that evaluators interact with those who are similar to them. However, part of the gap remains unexplained. Due to the increased attention in policing, this chapter is hoped to contribute to the public debate on policing.

The final chapter speaks to human capital formation. In it, I study how sleep is related to cognitive performance and what interventions can change the sleep patterns of undergraduate students. The estimates of the framed field experiment are noisy due to the small sample size, yet, the results of the panel methods are encouraging: the effect sizes are large in magnitude even if statistically insignificant.

The second main theme of the dissertation is the use of experiments. Each of the three chapters has an experimental element and these experiments include small innovations that are hoped to benefit future experimental designs. In the wage transparency experiment of chapter 1, I revisited the punishment games and changed how participants pay for the punishment to eliminate wealth effects and the house-money effect. The award nomination experiment used mouse-tracking to understand what information evaluators look for when they are nominating officers for an award. Finally, the sleep intervention experiment used fitness trackers to collect sleep data and

provide a prototype for incorporating sleep measurements to economic analyses.

Overall this dissertation shows that approaching problems from another perspective with less trivial tools leads to unforeseen obstacles, but also to exciting new lessons that can change how we understand the problem at hand. My intentions were to demonstrate in three chapters that this path is worthy.

Appendix

3.9 Appendix for Chapter 1

Table 3.9: Sample Characteristics

Variable	# obs	Mean	Std dev.	10^{th} pctile	90^{th} pctile	pr: balance
Panel A: Demographic variables						
female	287	0.491	0.501			0.209
age	292	31.9	10.0	20.0	46.0	0.208
region: northeast	294	0.197	0.399			0.479
region: midwest	294	0.197	0.399			0.102
region: south	294	0.364	0.482			0.110
region: west	294	0.241	0.429			0.351
race: american indian	294	0.031	0.173			0.011
race: asian	294	0.139	0.347			0.255
race: black	294	0.051	0.220			0.268
race: native hawaiian	294	0.010	0.101			0.100
race: white	294	0.786	0.411			0.100
race: other	294	0.037	0.190			0.049
hispanic	286	0.126	0.332			0.626
student	294	0.255	0.437			0.885
hh inc: less than $10,000	293	0.089	0.285			0.584
hh inc: $10,000 - $15,999	293	0.058	0.234			0.974
hh inc: $16,000 - $19,999	293	0.024	0.153			0.807
hh inc: $20,000 - $29,999	293	0.096	0.294			0.496
hh inc: $30,000 - $39,999	293	0.082	0.275			0.870
hh inc: $40,000 - $49,999	293	0.089	0.285			0.495
hh inc: $50,000 - $59,999	293	0.072	0.258			0.169
hh inc: $60,000 - $69,999	293	0.075	0.264			0.847
hh inc: $70,000 - $79,999	293	0.072	0.258			0.357
hh inc: $80,000 - $89,999	293	0.041	0.199			0.319
hh inc: $90,000 - $99,999	293	0.068	0.253			0.106
hh inc: $100,000 - $149,999	293	0.137	0.344			0.740
hh inc: more than $150,000	293	0.099	0.299			0.925
SES on 10 step ladder	293	5.14	1.78	3.00	7.00	0.721
pol: democrat	294	0.469	0.500			0.601
pol: republican	294	0.167	0.373			0.795
pol: independent	294	0.364	0.482			0.733
Panel B: Ability proxies						
prolific score	294	100	1	99	100	0.310
estimation task absolute error	291	23.5	14.6	10.6	36.0	0.389
estimation task high accuracy	291	0.553	0.498			0.431

Note: Table shows summary statistics for selected variables. Percentiles are omitted for binary variables. The final column reports the p-value for testing equality of means of the baseline variables across treatment and control groups in the punishment task, using heteroskedasticity-robust standard errors clustered by participants.

3.10 Appendix for Chapter 2

3.10.1 Additional Tables and Figures

Table 3.10: Department Awards

1	*Superintendent's Award of Valor* for an act of outstanding bravery or heroism by which the member has demonstrated in great degree the characteristics of selflessness, personal courage, and devotion to duty.
2	*Superintendent's Award of Merit* for an outstanding accomplishment that has resulted in improved administration, improved operations, or substantial savings in manpower or operational costs, wherein the member has gone far beyond the requirements of their normal assignment.
3	*Police Blue Star Award* is granted to any sworn member who has been seriously, critically, or fatally injured while in the performance of police duty.
4	*Police Blue Shield Award* is granted to any sworn member who, as a result of accidental causes, has been seriously, critically, or fatally injured while in the performance of police duty.
5	*Superintendent's Award of Tactical Excellence* for exceptional tactical skills or verbal approaches and techniques to mitigate any deadly force situation resulting in the saving or sustaining of a human life.
6	*Arnold Mireles Special Partnership Award* for making a significant impact upon the quality of life within their community by identifying and resolving problems.
7	*Special Commendation* for making a significant impact on public safety or crime prevention.
8	*Lifesaving Award* for a successful effort in saving a human life that involved exceptional courage or performance.
9	*Police Officer of the Month* for performance of duty during a specific month was characterized by such exceptional professional skill that it merits recognition by the entire Department.
10	*Chicago Police Leadership Award* for exemplary service, dedication, and leadership.
11	*Department Commendation* for an outstanding act or achievement that brings great credit to the Department and involves performance above and beyond that required by the member's basic assignment.
12	*Problem Solving Award* for an exemplary effort to identify, analyze, and successfully respond to causes, conditions, and problems that may lead to crime and neighborhood disorder.
13	*Joint Operations Award* for efforts and participation in a broad multi-agency joint operation/event, spanning several days or more, significantly contributing to the overall successes of the operation.
14	*Unit Meritorious Performance Award* for exhibiting exceptional professional skill and conduct during a coordinated action.
15	*Traffic Stop of the Month Award* for excellence in conducting professional traffic stops that result in quality arrests.
16	*Top Gun Arrest Award* for exceptional commitment to the recovery of illegal firearms.
17	*Special Service Award* for contributing to any event that has a significant impact upon the historical direction and operations of the Department.
18	*Honorable Mention Certificate* for demonstrating outstanding performance above and beyond that required by the member's assignment.

Source: Chicago Police Department Special Order S01-01-01 "Description and Eligibility for Department Awards", retrieved from `http://directives.chicagopolice.org/directives/`

Table 3.11: CPD Use of Force Options and Member Response

Use of Force Options	**Our Classification**
Force Mitigation Efforts	
Member Presence	
Zone of Safety	
Verbal Direction/Control Techniques	
Movement to Avoid Attack	
Specialized Units	Mitigation
Tactical Positioning	
Additional Unit Members	
None	
Other	
Control Tactics	
Escort Holds	
Wristlock	
Armbar	
Control Instrument	Control tactics
Pressure Sensitive Areas	
Emergency Handcuffing	
Other	
Response without Weapons	
Open Hand Strike	
Take down	
Elbow strike	
Close hand strike/Punch	No Weapon
Knee strike	
Kicks	
Other	
Response with Weapons	
OC/Chemical Weapon	
OC/Chemical Weapon w/Authorization	Non-Lethal Weapon
LRAD w/Authorization	
Taser	Taser
Canine	Canine
Baton/Expandable baton	Baton
Impact munitions	
Revolver	
Rifle	Firearm
Semi-auto pistol	
Shotgun	
Other	Other Use of Force

Source: Chicago Police Department TRR Form

Table 3.12: Evaluation Quarter and Due Dates by Start Month

Quarter	Anniversary Date Month of the Member	The Quarter the Member Will Be Evaluated	Due Date of the Evaluation
1st	January, February, March	4th	30 January
2nd	April, May, June	1st	30 April
3rd	July, August, September	2nd	30 July
4th	October, November, December	3rd	30 October

Source: Chicago Police Department, Career Development Directive, Employee Resource E05-01, Section IX, B. Retrieved from `http://directives.chicagopolice.org/directives/data/a7a56e3d-12887ea9-ce512-887e-c3dce7cd73e28d57.html?ownapi=1`

Table 3.13: Racial Difference in Nomination Likelihood by Quarter, With Officer Fixed-Effects

Estimates for:	Outcome Variable: Nominated White Officer	Black-White Gap	Hispanic-White Gap
	(1)	(2)	(3)
Quarter relative to two quarters before evaluation			
Panel A: White Supervisors			
One quarter pre-evaluation	0.00661***	-0.00393	0.00372
	(0.00239)	(0.00285)	(0.00372)
Evaluation quarter	0.00908***	-0.00434	0.00525
	(0.00299)	(0.00285)	(0.00370)
One quarter post-evaluation	0.00177	-0.00350	0.00216
	(0.00383)	(0.00297)	(0.00393)
Two quarters post-evaluation	-0.00130	-0.00241	0.00575
	(0.00482)	(0.00367)	(0.00468)
Observations		154,964	
Panel B: Black Supervisors			
One quarter pre-evaluation	-0.00131	0.00696	0.0169*
	(0.00660)	(0.00672)	(0.00961)
Evaluation quarter	0.000675	0.00481	0.0207**
	(0.00759)	(0.00673)	(0.00932)
One quarter post-evaluation	-0.0138	0.0168**	0.0233**
	(0.00900)	(0.00691)	(0.00960)
Two quarters post-evaluation	-0.00577	0.00720	0.0166
	(0.0110)	(0.00810)	(0.0128)
Observations		26,556	
Baseline B-W Nomination Gap		-0.0045	

Source: CPD data.

Notes: This table depicts how the probability of nomination changes by quarter relative to two quarters before the officer's evaluation. Estimates are reported for white supervisors in Panel A and for black supervisors in Panel B. Each panel is a single OLS regression with estimates for white officers in column 1, the black-white difference in column 2, and the Hispanic-white difference in column 3. All estimates include officer, month, and year fixed effects, and control for officer tenure, district, lagged arrests, lagged complaints, lagged overall crime rate, lagged violent crime rate, and the share of black supervisees. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 3.14: Impact of Arrest Record on Nomination Likelihood by Officer Race, With Officer Fixed-Effects

Estimates for:	Outcome Variable: Nominated White Officer	Black-White Gap	Hisp.-White Gap
	(1)	(2)	(3)
Panel A: White Supervisors			
One arrest	0.00626***	-0.00494**	-0.00118
	(0.00143)	(0.00206)	(0.00254)
Two arrests	0.00922***	-0.00540*	0.00138
	(0.00202)	(0.00296)	(0.00350)
Three arrests	0.0127***	-0.00725*	0.00241
	(0.00274)	(0.00418)	(0.00474)
Four arrests	0.0153***	-0.00412	-0.00157
	(0.00342)	(0.00588)	(0.00591)
Five or more arrests	0.0325***	-0.0152***	-0.00828
	(0.00308)	(0.00531)	(0.00509)
Observations		171,094	
Mean Pr(Nom) for White Officers		0.031	
Panel B: Black Supervisors			
One arrest	-0.000327	0.00271	0.00206
	(0.00367)	(0.00414)	(0.00759)
Two arrests	0.00551	-0.00377	-6.91e-05
	(0.00527)	(0.00593)	(0.0108)
Three arrests	0.00576	0.00416	0.00721
	(0.00721)	(0.00848)	(0.0149)
Four arrests	0.00253	-0.00872	0.00859
	(0.0104)	(0.0119)	(0.0189)
Five or more arrests	0.00948	0.0272**	-0.00134
	(0.00897)	(0.0127)	(0.0155)
Observations		29,413	
Mean Pr(Nom) for White Officers		0.022	

Source: CPD data.

Notes: This table reports estimates for the impact of an officer's arrest record on the probability of nomination by white supervisors (Panel A) and by black supervisors (Panel B). Each panel is a single OLS regression with estimates for white officers in column 1, the black-white difference in column 2, and the Hispanic-white difference in column 3. All estimates include officer, month, and year fixed effects, and control for officer tenure, district, lagged complaints, lagged overall crime rate, lagged violent crime rate, and the share of black supervisees. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

3.10.2 Online Experiment

The experiment was conducted on Amazon Mechnical Turk (MTurk) in July 2020. It was pre-registered in the AEA RCT Registry, AEARCTR-0005929. We recruited 411 MTurk workers (hereafter "workers") who were compensated $1.20 for completing a survey experiment. Table 3.15 reports summary statistics on all 411 workers. Figure 3.8 plots the distribution of workers' states of residence.

We included three data quality checks to identify bots and to ensure workers paid attention during the survey. For the analysis, we decided to include workers who passed at least two of the three data quality checks. This restriction reduces our final analysis sample to 407 workers.

To avoid deception in our survey, we used real officer profiles but used officer initials to preserve officers' identities. Workers were informed that the profiles belonged to real officers but were not told which agency they were from. Further, we informed workers that their nominations would be relayed to the police department. This was to achieve incentive compatibility. After the experiment ended, the Chicago Police Department was informed of survey results.

Table 3.15: Summary Statistics

	N	%
Race		
Black	48	11.7%
Hispanic	66	16.1%
White	263	64.0%
Other	21	5.1%
Prefer not to answer	10	2.4%
Missing	3	0.7%
Female	166	40.4%
Age		
18-25	53	12.9%
26-35	189	46.0%
36-45	78	19.0%
46-55	55	13.4%
56+	35	8.5%
Missing	1	0.2%
Is English your first language?		
Yes	401	97.6%
No	5	1.2%
Missing	5	1.2%
Length of Residency in US		
$<$ 1 yr	6	1.5%
More than 1 yr but less than 3 yrs	21	5.1%
More than 3 yrs but less than 6 yrs	16	3.9%
More than 6 yrs	365	88.8%
Missing	3	0.7%
Number of Surveys (MTurk Workers)	411	

Source: MTurk survey data.

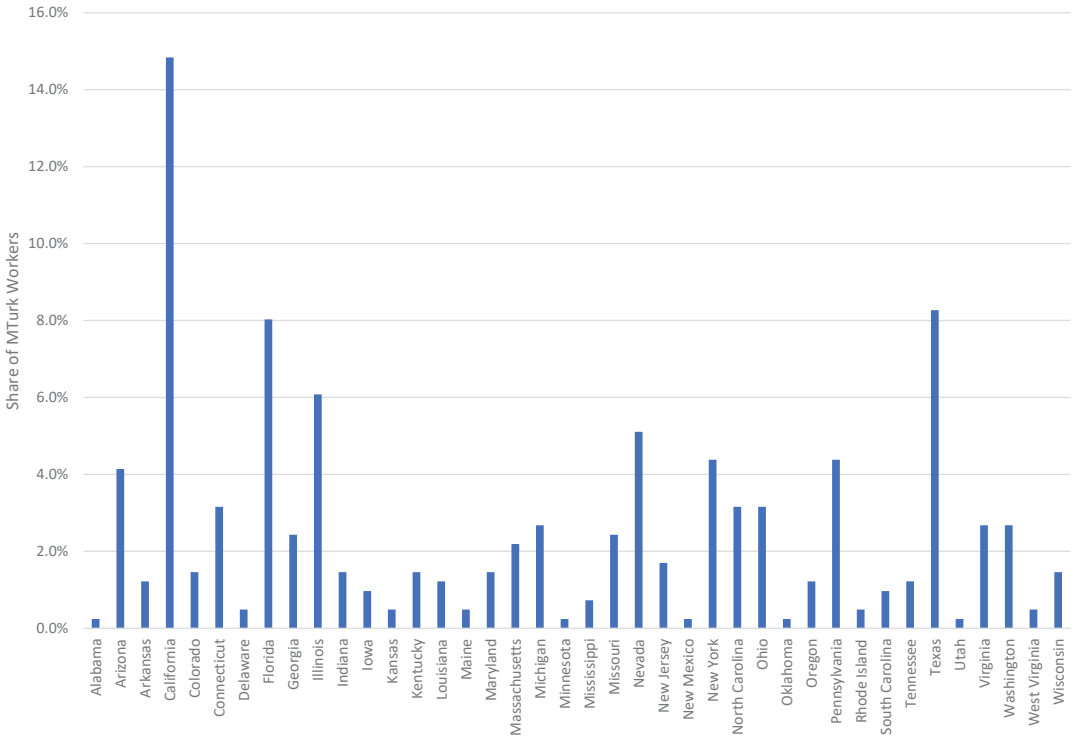


Figure 3.8: Distribution of MTurk Worker State of Residence

Source: MTurk survey data.

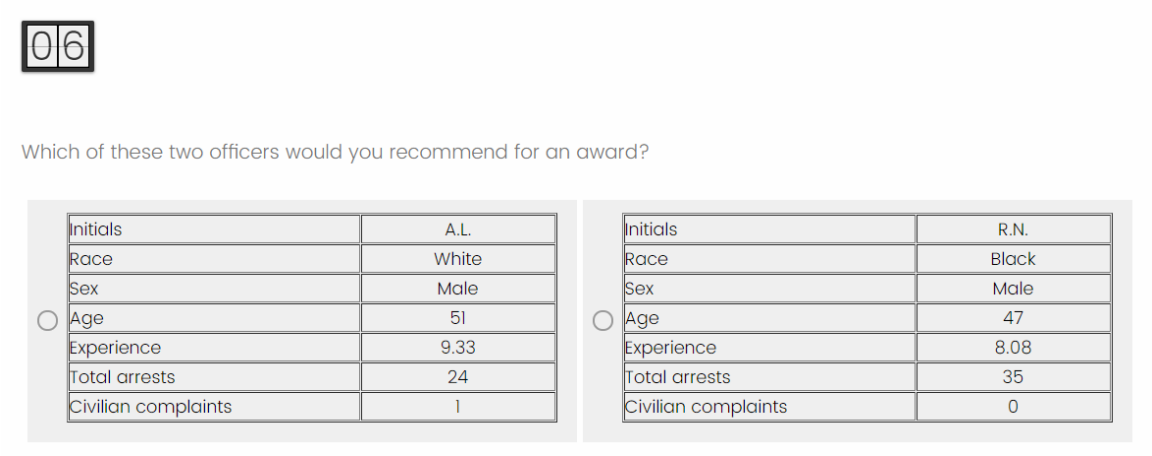


Figure 3.9: Screenshot of Pairwise Comparison Task

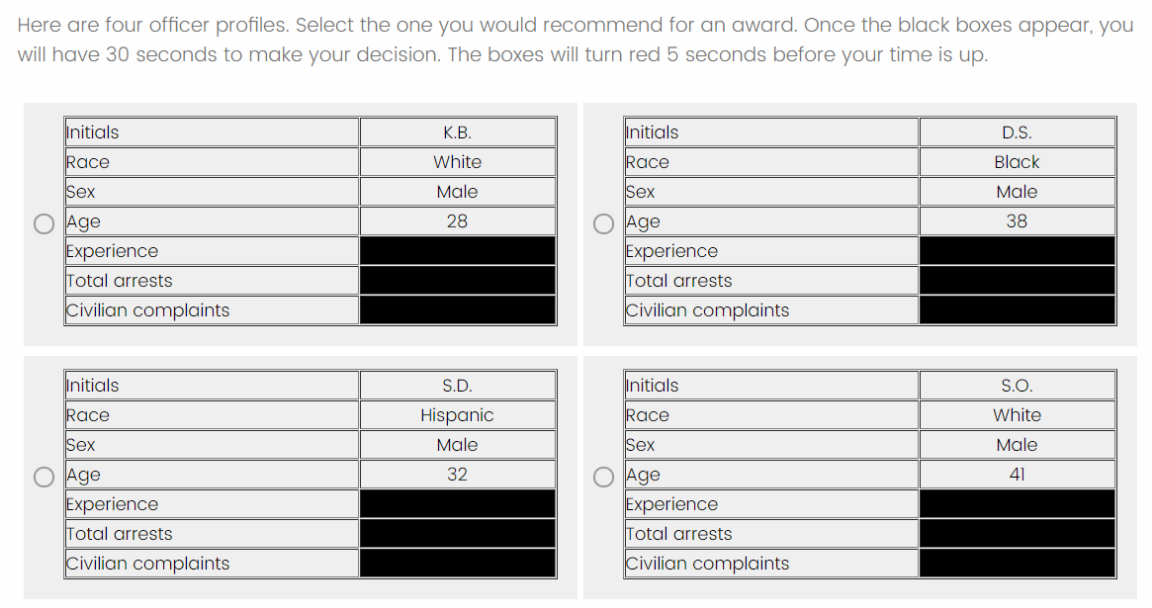


Figure 3.10: Screenshot of Group Comparison Task

3.11 Appendix for Chapter 3

Appendix

3.11.1 Sleep Hygiene Information Treatment

One of the treatment arm, the information arm, received sleep hygiene tips during the first online session. The tips were the Duke Wellness Center's adaptation of Huffington's 2016 advice. The following information was displayed:

Sleep Hygiene has 4 Components

Regular Routine

Daily regular bedtime and wake time for four weeks

Relaxation time, no homework one hour before bedtime

No TV, computer, or other screen use one hour before bedtime.

Bed used only for sleep

Get out of bed after fifteen minutes rather than tossing & turning

Avoid naps six hours before bedtime

Environment

Dark bedroom, cool temperature

Limit noise or use white noise

Minimize allergens

Diet

Caffeine: Under 16 oz daily, none six hours before bedtime

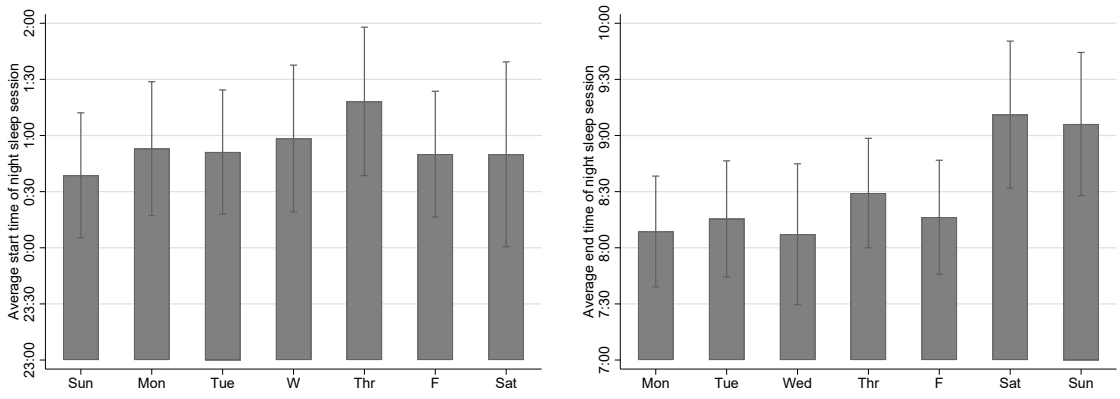
Alcohol: Nicotine: None five hours before sleep

Light bedtime snack: carbohydrate/ protein combination

Exercise

Routine of twenty minutes daily, preferably outdoors

Avoid strenuous exercise five hours before sleep



(a) Average bedtime on nights (b) Average wake-time on mornings

Figure 3.11: Average Start and End Time of Night Sleep During the Baseline Week

Note: This figure shows the average start time and end time of night sleep sessions during the baseline week. To avoid bias from multiple sleep sessions, if more than one sleep session was recorded for a participant for a given night, the first sleep session was taken into account for the sleep start time and the last sleep session was used for the sleep end time measurement. The order of the columns on the two panels correspond. Panel a) shows the average bedtime on nights of the given days, while Panel b) shows the average wake-time on the mornings of the given days. Wings mark the 95% confidence intervals using heteroskedasticity-robust standard errors clustered at participants level.

3.11.2 Additional Tables

Table 3.16: Sample Characteristics

Variable	# obs	Mean	Std dev.	10^{th} pctile	90^{th} pctile	pr: balance with control		
						leaflet	video	SMS
Panel A: Demographic variables								
female	41	0.829	0.381			0.867	0.498	0.498
year of birth	41	1997	1	1995	1999	0.009	0.104	0.442
US citizen	41	0.878	0.331			0.100	.	0.025
Race: Asian	41	0.439	0.502			0.273	0.366	0.941
Race: Black	41	0.049	0.218			.	0.238	0.238
Race: White	41	0.488	0.506			0.273	0.941	0.570
Race: Multi	41	0.024	0.156			.	0.238	.
Panel B: Student life variables								
Trinity School	41	0.829	0.381			0.784	0.787	0.700
Fraternity member	41	0.439	0.502			0.327	0.211	0.211
Althletics member	41	0.317	0.471			0.372	0.676	0.676
Class year	41	2.63	1.26	1.00	4.00	0.027	0.459	0.586
Credits taken	41	4.15	0.87	3.00	5.00	0.393	0.152	0.452
Panel C: Health								
Weight (lb)	40	142	29	109	180	0.236	0.412	0.627
Height (inch)	40	65.8	4.3	62.0	69.0	0.538	0.091	0.224
Panel D: Baseline cognitive score								
Reasoning score	41	34.0	13.3	16.0	50.0	0.670	0.564	0.706
Reasoning task N attempted	41	39.0	12.1	24.0	53.0	0.927	0.826	0.463
Memory score	41	23.7	1.4	22.0	25.0	0.777	0.475	0.301
Memory task duration (sec)	41	66.4	18.8	52.3	75.7	0.478	0.986	0.310
Total score	41	57.7	13.7	39.0	74.0	0.654	0.628	0.794
Panel E: Baseline sleep pattern								
Avg night: time in bed (min)	40	473	88	405	574	0.853	0.489	0.033
Avg night: N awakenings	40	2.67	5.02	0.54	3.79	0.390	0.970	0.263
Largest shift in bedtime (min)	40	195	104	92	320	0.233	0.325	0.319
Largest shift in wake-time (min)	40	201	104	89	338	0.468	0.920	0.395

Note: Table shows summary statistics for selected variables. Percentiles are omitted for binary variables. The final columns report the p-values for testing equality of means of the baseline variables across treatment and control groups.

Table 3.17: Characteristics Predicting the Share of Missing Night Sleep Data

	Percentage of nights without sleep data
Birth year	
1995	-1.178***
	(0.395)
1996	-1.158***
	(0.306)
1997	-0.938***
	(0.288)
1998	-0.983***
	(0.284)
1999	-0.924***
	(0.279)
Female	-0.035
	(0.217)
Race	
Black	-0.012
	(0.213)
White	0.161
	(0.118)
Multiple	0.136
	(0.139)
Weight	-0.008***
	(0.003)
Height	-0.003
	(0.008)
Trinity School	-0.076
	(0.076)
Fraternity member	0.073
	(0.089)
Athletics member	-0.054
	(0.109)
Has a job	-0.066
	(0.096)
Hours studying	-0.009
	(0.008)
N clusters	40
N	40
R2	0.556

Note: This table shows the correlates of personal characteristics with the percentage of nights with missing sleep data. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

Table 3.18: Treatment Effect on Night Level Sleep without Controls

	Treatment week: night level analysis			
	(1) Time in bed (min)	(2) Awakenings (N)	(3) Bedtime (hr)	(4) Wake-time (hr)
SMS reminder × week 2	-26.176	-0.925	0.196	0.032
	(25.629)	(0.963)	(0.252)	(0.207)
Meditation video × week 2	-41.171	-0.422	0.124	0.187
	(43.718)	(0.574)	(0.471)	(0.215)
Information × week 2	23.457	-0.038	-0.123	-0.105
	(30.625)	(0.385)	(0.494)	(0.362)
Control × week 2	-11.206	0.249	-0.055	-0.208
	(22.041)	(0.373)	(0.275)	(0.216)
Main treatment indicators	✓	✓	✓	✓
Additional controls				
Day of the week indicators	✓	✓	✓	✓
Control mean	427.390	1.909	1.027	8.944
Pre-treatment avg level	460.211	2.796	0.234	9.366
N	499	499	499	499
N cluster	40	40	40	40
R2	0.073	0.056	0.043	0.079

Note: This table reports the treatment effects of the three behavioral interventions on participants' sleep at the night level. The last two columns estimate the effect on measures of consistent sleep schedule. Bedtime and wake-times are the are expressed in fraction of hours, so the control group during the treatment weak has a median bedtime at about 1:01 am (1.027 = 1 am + 0.027 * 60) and their average respective median wake-time is at 8:57 am. The set of additional control variables include indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. All specifications include day of the day indicator variables. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

Table 3.19: Treatment Effects on Sleep Pattern Changes

	$Y_{\text{treatment week}} - Y_{\text{baseline week}}$		
	(1) Median time in bed (min)	(2) Median N awakenings	(3) Median bedtime (min)
SMS reminder	-0.500	-0.806	0.967*
	(24.489)	(0.904)	(0.537)
Meditation video	-2.611	-0.583	0.634
	(28.901)	(0.439)	(0.422)
Information	37.833	-0.194	0.196
	(25.007)	(0.610)	(0.540)
Additional controls			
Control mean	-21.667	0.417	-0.300
Pre-treatment avg level	440.024	2.280	24.919
N	39	39	38
N cluster	39	39	38
R2	0.082	0.045	0.111

	(4) Median wake-time (min)	(5) Maximum difference in time in bed (min)
SMS reminder	0.877	-72.250
	(0.754)	(181.061)
Meditation video	0.514	-158.917
	(0.480)	(134.624)
Information	0.515	135.528
	(0.540)	(130.027)
Additional controls		
Control mean	-0.439	36.250
Pre-treatment avg level	8.475	307.780
N	38	39
N cluster	38	39
R2	0.065	0.082

Note: This table reports the treatment effects of the three behavioral interventions on the change of participants' nightly sleep patterns between the treatment week and the baseline week. The outcome variables are aggregated over a week long period to account for the weekly seasonality. The last three columns estimate the effect on measures of consistent sleep schedule. Median bedtime and wake-times are the median time when the first night sleep session started and the last night sleep session ended respectively over the treatment period. The maximum difference in time spent in bed measures the difference between the longest and shortest recorded night sleep over the treatment period. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

Table 3.20: Treatment Effects on the Usage of Sleep Aids

	Δ Mask	Δ Scents	Δ Sleeping pills	Δ Melatonin	Δ Earplug
SMS reminder	0.039	0.079	0.036	0.004	-0.118
	(0.292)	(0.382)	(0.303)	(0.086)	(0.288)
Meditation video	0.185	-0.012	-0.047	-0.321*	-0.140
	(0.193)	(0.456)	(0.226)	(0.165)	(0.213)
Information	-0.037	0.453*	0.055	-0.156	0.038
	(0.167)	(0.243)	(0.147)	(0.164)	(0.239)
Additional controls	✓	✓	✓	✓	✓
Control mean	0.077	-0.154	0.077	0.000	-0.154
Pre-treatment avg level	0.049	0.171	0.049	0.049	0.171
N	39	39	39	39	39
N cluster	39	39	39	39	39
R2	0.514	0.548	0.347	0.547	0.392

	Δ Music	Δ White noise machine	Δ Meditate	Δ Other	Δ Daily coffee
SMS reminder	0.109	-0.248	-0.212	0.000	0.061
	(0.076)	(0.339)	(0.283)	(.)	(0.181)
Meditation video	0.132	0.239	0.137	-1.000	0.088
	(0.119)	(0.266)	(0.292)	(.)	(0.215)
Information	-0.152*	-0.011	0.112	0.000	-0.176
	(0.074)	(0.256)	(0.113)	(.)	(0.273)
Additional controls	✓	✓	✓	✓	✓
Control mean	0.000	-0.077	-0.077	0.000	0.333
Pre-treatment avg level	0.024	0.122	0.122	0.100	0.000
N	39	39	39	12	38
N cluster	39	39	39	12	38
R2	0.740	0.544	0.646	1.000	0.728

Note: This table reports the treatment effects of the three behavioral interventions on the change in participants' use of various sleep aids. The set of additional control variables include indicators for gender, race, US citizenship, birth year, as well as participants' weight, height and finally, the date of the participants' first in-person visit. Heteroskedasticity-robust standard errors shown in parentheses, clustering at the participant level. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

Bibliography

Abbink, K. and Sadrieh, A. (2009). The Pleasure of Being Nasty. *Economics Letters*, 105(3):306–308.

Agan, A. and Starr, S. (2017). Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment. *The Quarterly Journal of Economics*, 133(1):191–235.

Ajilore, O. and Shirey, S. (2017). Do #AllLivesMatter? An Evaluation of Race and Excessive Use of Force by Police. *Atlantic Economic Journal*, 45(2):201–212.

Alsan, M., Garrick, O., and Graziani, G. (2019). Does Diversity Matter for Health? Experimental Evidence From Oakland. *American Economic Review*, 109(12):4071–4111.

Altonji, J. G. and Pierret, C. R. (2001). Employer Learning and Statistical Discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Antonovics, K. and Knight, B. G. (2009). A New Look at Racial Profiling: Evidence From the Boston Police Department. *The Review of Economics and Statistics*, 91(1):163–177.

Anwar, S., Bayer, P., and Hjalmarsson, R. (2012). The Impact of Jury Race in Criminal Trials. *The Quarterly Journal of Economics*, 127(2):1017–1055.

Anwar, S. and Fang, H. (2006). An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence. *American Economic Review*, 96(1):127–151.

Arnold, D., Dobbie, W., and Yang, C. S. (2018). Racial Bias in Bail Decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932.

Avery, M., Giuntella, O., and Jiao, P. (2020). Why Don't We Sleep Enough? A Field Experiment Among College Students. Technical report.

Ba, B., Knox, D., Mummolo, J., and Rivera, R. (2020). Diversity in Policing: The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago. Working paper.

Bacher-Hicks, A. and de la Campa, E. (2020). Social Costs of Proactive Policing: The Impact of NYC's Stop and Frisk Program on Educational Attainment. Working paper.

Baddeley, A. D. (1968). A 3 Min Reasoning Test Based on Grammatical Transformation. *Psychonomic Science*, 10(10):341–342.

Baker, M., Halberstam, Y., Kroft, K., Mas, A., and Messacar, D. (2019). Pay Transparency and the Gender Gap. Technical report, National Bureau of Economic Research.

Bartling, B. and Fischbacher, U. (2012). Shifting the Blame: On Delegation and Responsibility. *The Review of Economic Studies*, 79(1):67–87.

Bartoš, V., Bauer, M., Chytilová, J., and Matějka, F. (2016). Attention Discrimination: Theory and Field Experiments With Monitoring Information Acquisition. *American Economic Review*, 106(6):1437–75.

Beaman, L., Keleher, N., and Magruder, J. (2018). Do Job Networks Disadvantage Women? Evidence From a Recruitment Experiment in Malawi. *Journal of Labor Economics*, 36(1):121–157.

Becker, S. P., Jarrett, M. A., Luebbe, A. M., Garner, A. A., Burns, G. L., and Kofler, M. J. (2018). Sleep in a Large, Multi-University Sample of College Students: Sleep Problem Prevalence, Sex Differences, and Mental Health Correlates. *Sleep health*, 4(2):174–181.

Belot, M. and Schröder, M. (2013). Sloppy Work, Lies and Theft: A Novel Experimental Design to Study Counterproductive Behaviour. *Journal of Economic Behavior & Organization*, 93:233–238.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive Linear Step-up Procedures that Control the False Discovery Rate. *Biometrika*, 93(3):491–507.

Bennedsen, M., Simintzi, E., Tsoutsoura, M., and Wolfenzon, D. (2019). Do Firms Respond to Gender Pay Gap Transparency? Technical report, National Bureau of Economic Research.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013.

Bessone, P., Rao, G., Schilbach, F., Schofield, H., and Toma, M. (2020). The Economic Consequences of Increasing Sleep Among the Urban Poor. Technical report, National Bureau of Economic Research.

Biddle, J. E. and Hamermesh, D. S. (1990). Sleep and the Allocation of Time. *Journal of Political Economy*, 98(5, Part 1):922–943.

Blanes i Vidal, J. and Kirchmaier, T. (2018). The Effect of Police Response Time on Crime Clearance Rates. *The Review of Economic Studies*, 85(2):855–891.

Blunden, S. L., Chapman, J., and Rigney, G. A. (2012). Are Sleep Education Programs Successful? The Case for Improved and Consistent Research Efforts. *Sleep medicine reviews*, 16(4):355–370.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019). The Dynamics of Discrimination: Theory and Evidence. *American Economic Review*, 109(10):3395–3436.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90(1):166–193.

Breza, E., Kaur, S., and Shamdasani, Y. (2017). The Morale Effects of Pay Inequality. *The Quarterly Journal of Economics*, 133(2):611–663.

Bulman, G. (2019). Law Enforcement Leaders and the Racial Composition of Arrests. *Economic Inquiry*.

Card, D., Mas, A., Moretti, E., and Saez, E. (2012). Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *American Economic Review*, 102(6):2981–3003.

Carrell, S. E., Maghakian, T., and West, J. E. (2011). A's From Zzzz's? The Causal Effect of School Start Time on the Academic Achievement of Adolescents. *American Economic Journal: Economic Policy*, 3(3):62–81.

Carrell, S. E., Page, M. E., and West, J. E. (2010). Sex and Science: How Professor Gender Perpetuates the Gender Gap. *The Quarterly Journal of Economics*, 125(3):1101–1144.

Castleman, B. L. and Page, L. C. (2016). Freshman Year Financial Aid Nudges: An Experiment to Increase FAFSA Renewal and College Persistence. *Journal of Human Resources*, 51(2):389–415.

Charness, G. and Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? *Handbook of labor economics*, 4:229–330.

Charness, G. and Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Civil Rights Act of 1964 § 7, 42 U.S.C. § 2000e *et seq* (1964). Retrieved from Equal Employment Opportunity Commission website: `http://www.eeoc.gov/laws/statutes/titlevii.cfm` on 2021-01-31.

Close, B. R. and Mason, P. L. (2006). After the Traffic Stops: Officer Characteristics and Enforcement Actions. *The BE Journal of Economic Analysis & Policy*, 6(1).

Cohn, A., Fehr, E., Herrmann, B., Schneider, F., et al. (2011). Social Comparison in the Workplace: Evidence from a Field Experiment. Technical report, University of Zurich.

Cooper, D. J. and Kagel, J. H. (2016). Other-regarding Preferences. *The Handbook of Experimental Economics*, 2:217.

Craigie, T.-A. (2020). Ban the Box, Convictions, and Public Employment. *Economic Inquiry*, 58(1):425–445.

Cullen, Z. and Perez-Truglia, R. (2018a). How Much Does Your Boss Make? The Effects of Salary Comparisons. *National Bureau of Economic Research*.

Cullen, Z. and Perez-Truglia, R. (2018b). The Salary Taboo: Privacy Norms and the Diffusion of Information. Technical report, National Bureau of Economic Research.

Cullen, Z. B. and Pakzad-Hurson, B. (2019). Equilibrium Effects of Pay Transparency in a Simple Labor Market. Technical report.

Cullen, Z. B. and Perez-Truglia, R. (2020). The Old Boys' Club: Schmoozing and the Gender Gap. Working paper, National Bureau of Economic Research.

Cunningham, J. P. and Gillezeau, R. (2018). Racial Differences in Police Use of Force: Evidence From the 1960s Civil Disturbances. *AEA Papers and Proceedings*, 108:217–21.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic literature*, 47(2):315–72.

Dickinson, D. L., Masclet, D., and Villeval, M. C. (2015). Norm Enforcement in Social Dilemmas: An Experiment With Police Commissioners. *Journal of Public Economics*, 126:74–85.

Dohmen, T. (2014). Behavioral labor economics: Advances and future directions. *Labour Economics*, 30:71–85.

Doleac, J. L. and Hansen, B. (2018). Does "Ban the Box" Help or Hurt Low-Skilled Workers? Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden. *Journal of Labor Economics*, Forthcoming.

Donohue III, J. J. and Levitt, S. D. (2001). The Impact of Race on Policing and Arrests. *The Journal of Law and Economics*, 44(2):367–394.

Egan, M. L., Matvos, G., and Seru, A. (2018). When Harry Fired Sally: The Double Standard in Punishing Misconduct. Working paper, National Bureau of Economic Research.

Equal Pay Act of 1963, Pub. L. No. 88-38, 88th Congress, H.R. 6060 and S. 1409 (1963). Retrieved from Equal Employment Opportunity Commission website: `https://www.eeoc.gov/statutes/equal-pay-act-1963` on 2021-01-31.

Evenson, K. R., Goto, M. M., and Furberg, R. D. (2015). Systematic Review of the Validity and Reliability of Consumer-Wearable Activity Trackers. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1):1–22.

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

Falk, A. and Fischbacher, U. (2006). A Theory of Reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Fehr, E. and Fischbacher, U. (2004). Third-party Punishment and Social Norms. *Evolution and Human Behavior*, 25(2):63–87.

Fehr, E. and Schmidt, K. M. (1999). A theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fehr, E. and Schmidt, K. M. (2006). The Economics of Fairness, Reciprocity and Altruism–Experimental Evidence and New Theories. *Handbook of the Economics of Giving, Altruism and Reciprocity*, 1:615–691.

Fisman, R., Kuziemko, I., and Vannutelli, S. (2018). Distributional Preferences in Larger Groups: Keeping up With the Joneses and Keeping Track of the Tails. *Boston University-Department of Economics-The Institute for Economic Development Working Papers Series DP301.*

Flanagan, F. X. (2018). Race, Gender, and Juries: Evidence From North Carolina. *The Journal of Law and Economics*, 61(2):189–214.

Flory, J. A., Leibbrandt, A., and List, J. A. (2016). The Effects of Wage Contracts on Workplace Misbehaviors: Evidence from a Call Center Natural Field Experiment. Technical report, National Bureau of Economic Research.

Garner, M., Harvey, A., and Johnson, H. (2020). Estimating Effects of Affirmative Action in Policing: A Replication and Extension. *International Review of Law and Economics*, 62:105881.

Gershenson, S., Hart, C., Hyman, J., Lindsay, C., and Papageorge, N. W. (2018). The Long-Run Impacts of Same-Race Teachers. Working paper, National Bureau of Economic Research.

Gibson, M. and Shrader, J. (2018). Time Use and Labor Productivity: The Returns to Sleep. *Review of Economics and Statistics*, 100(5):783–798.

Giuliano, L., Levine, D. I., and Leonard, J. (2009). Manager Race and the Race of New Hires. *Journal of Labor Economics*, 27(4):589–631.

Giuntella, O., Han, W., and Mazzonna, F. (2017). Circadian Rhythms, Sleep, and Cognitive Skills: Evidence From an Unsleeping Giant. *Demography*, 54(5):1715–1742.

Giuntella, O. and Mazzonna, F. (2019). Sunset Time and the Economic Effects of Social Jetlag: Evidence From Us Time Zone Borders. *Journal of health economics*, 65:210–226.

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a Self-Fulfilling Prophecy: Evidence From French Grocery Stores. *The Quarterly Journal of Economics*, 132(3):1219–1260.

Goncalves, F. and Mello, S. (2020). A Few Bad Apples?: Racial Bias in Policing. Working paper.

Gulyas, A., Seitz, S., Sinha, S., et al. (2020). Does Pay Transparency Affect the Gender Wage Gap? Evidence from Austria. Technical report, University of Bonn and University of Mannheim, Germany.

Hafner, M., Stepanek, M., Taylor, J., Troxel, W. M., and van Stolk, C. (2016). Why Sleep Matters–the Economic Costs of Insufficient Sleep. *Europe: RAND Corporation*.

Handel, B. and Kolstad, J. (2017). Wearable Technologies and Health Behaviors: New Data and New Methods to Understand Population Health. *American Economic Review*, 107(5):481–85.

Harvey, A. and Mattia, T. (2020). Reducing Racial Disparities in Crime Victimization. Technical report.

Hauser, O. P., Kraft-Todd, G. T., Rand, D. G., Nowak, M. A., and Norton, M. I. (2019). Invisible Inequality Leads to Punishing the Poor and Rewarding the Rich. *Behavioural Public Policy*, pages 1–21.

Hengel, E. (2019). Publishing While Female. Are Women Held to Higher Standards? Evidence From Peer Review.

Hershner, S. (2020). Sleep and Academic Performance: Measuring the Impact of Sleep. *Current Opinion in Behavioral Sciences*, 33:51–56.

Hirshkowitz, M., Whiton, K., Albert, S. M., Alessi, C., Bruni, O., DonCarlos, L., Hazen, N., Herman, J., Katz, E. S., Kheirandish-Gozal, L., et al. (2015a). National Sleep Foundation's Sleep Time Duration Recommendations: Methodology and Results Summary. *Sleep health*, 1(1):40–43.

Hirshkowitz, M., Whiton, K., Albert, S. M., Alessi, C., Bruni, O., DonCarlos, L., Hazen, N., Herman, J., Katz, E. S., Kheirandish-Gozal, L., et al. (2015b). National Sleep Foundation's Sleep Time Duration Recommendations: Methodology and Results Summary. *Sleep health*, 1(1):40–43.

Hoekstra, M. and Sloan, C. (2020). Does Race Matter for Police Use of Force? Evidence from 911 Calls. Working paper.

Horrace, W. C. and Rohlin, S. M. (2016). How Dark Is Dark? Bright Lights, Big City, Racial Profiling. *Review of Economics and Statistics*, 98(2):226–232.

Huet-Vaughn, E. (2015). Do Social Comparisons Motivate Workers? A Field Experiment on Relative Earnings, Labor Supply and the Inhibitory Effect of Pay Inequality. Technical report, Middlebury College.

Huffington, A. (2016). *The sleep revolution: Transforming your life, one night at a time*. Harmony.

Jagnani, M. (2020). Poor sleep: Sunset time and human capital production.

Jin, L. and Ziebarth, N. R. (2020). Sleep, Health, and Human Capital: Evidence From Daylight Saving Time. *Journal of Economic Behavior & Organization*, 170:174–192.

Jonides, J. and Nee, D. E. (2006). Brain Mechanisms of Proactive Interference in Working Memory. *Neuroscience*, 139(1):181–193.

Kessler, J. B., Low, C., and Sullivan, C. D. (2019). Incentivized Resume Rating: Eliciting Employer Preferences Without Deception. *American Economic Review*, 109(11):3713–44.

Kim, M. (2015). Pay Secrecy and the Gender Wage Gap in the United States. *Industrial Relations: A Journal of Economy and Society*, 54(4):648–667.

Knowles, J., Persico, N., and Todd, P. (2001). Racial Bias in Motor Vehicle Searches: Theory and Evidence. *Journal of Political Economy*, 109(1):203–229.

Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative Records Mask Racially Biased Policing. *American Political Science Review*, 114(3):619–637.

Koch, A., Nafziger, J., and Nielsen, H. S. (2015). Behavioral Economics of Education. *Journal of Economic Behavior & Organization*, 115:3–17.

Kofoed, M. and mcGovney, E. (2019). The Effect of Same-Gender or Same-Race Role Models on Occupation Choice Evidence From Randomly Assigned Mentors at West Point. *Journal of Human Resources*, 54(2):430–467.

Kuziemko, I., Buell, R. W., Reich, T., and Norton, M. I. (2014). "Last-place Aversion": Evidence and Redistributive Implications. *The Quarterly Journal of Economics*, 129(1):105–149.

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

Langan, A. (2018). Female Managers and Gender Disparities: The Case of Academic Department Chairs. Working paper.

Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K., and Rathouz, P. J. (2008). Sleep Duration: How Well Do Self-Reports Reflect Objective Measures? The Cardia Sleep Study. *Epidemiology (Cambridge, Mass.)*, 19(6):838.

Lavecchia, A. M., Liu, H., and Oreopoulos, P. (2016). Behavioral Economics of Education: Progress and Possibilities. In *Handbook of the Economics of Education*, volume 5, pages 1–74. Elsevier.

Lee, H.-A., Lee, H.-J., Moon, J.-H., Lee, T., Kim, M.-G., In, H., Cho, C.-H., and Kim, L. (2017). Comparison of Wearable Activity Tracker with Actigraphy for Sleep Evaluation and Circadian Rest-Activity Rhythm Measurement in Healthy Young Adults. *Psychiatry investigation*, 14(2):179.

Levine, D. K. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, 1(3):593–622.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2016). The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

List, J. A. and Rasul, I. (2011). Field experiments in labor economics. In *Handbook of labor economics*, volume 4, pages 103–228. Elsevier.

Liu, Y., Wheaton, A. G., Chapman, D. P., Cunningham, T. J., Lu, H., and Croft, J. B. (2016). Prevalence of Healthy Sleep Duration Among Adults—United States, 2014. *Morbidity and Mortality Weekly Report*, 65(6):137–141.

Loudenback, T. (2017). More Tech Companies Have Stopped Keeping Employee Salaries Secret — and They're Seeing Results. `https://www.businessinsider.com/why-companies-have-open-salaries-and-pay-transparency-2017-4`. Accessed: 2020-10-22.

Lowe, M. (2019). Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration. Working paper.

Lyons, E. (2017). Team Production in International Labor Markets: Experimental Evidence From the Field. *American Economic Journal: Applied Economics*, 9(3):70–104.

MacLeod, W. B. (2003). Optimal Contracting With Subjective Evaluation. *American Economic Review*, 93(1):216–240.

Marino, M., Li, Y., Rueschman, M. N., Winkelman, J. W., Ellenbogen, J., Solet, J. M., Dulin, H., Berkman, L. F., and Buxton, O. M. (2013). Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography. *Sleep*, 36(11):1747–1755.

Mas, A. (2017). Does Transparency Lead to Pay Compression? *Journal of Political Economy*, 125(5):1683–1721.

Mason, P. L. (2007). Driving While Black: Do Police Pass the Test? *Swedish Economic Policy Review*, 14:79–113.

Mastrobuoni, G. (2020). Crime Is Terribly Revealing: Information Technology and Police Productivity. *The Review of Economic Studies*, 87(6):2727–2753.

McCrary, J. (2007). The Effect of Court-ordered Hiring Quotas on the Composition and Quality of Police. *American Economic Review*, 97(1):318–353.

Miller, A. and Segal, C. (2018). Do Female Officers Improve Law Enforcement Quality? Effects on Crime Reporting and Domestic Violence. *Review of Economic Studies*, Accepted.

Montgomery-Downs, H. E., Insana, S. P., and Bond, J. A. (2012). Movement Toward a Novel Activity Monitoring Device. *Sleep and Breathing*, 16(3):913–917.

Mousa, S. (2020). Building Social Cohesion Between Christians and Muslims Through Soccer in Post-ISIS Iraq. *Science*, 369(6505):866–870.

Mueller-Smith, M. and Schnepel, K. (2017). Diversion in the Criminal Justice System: Regression Discontinuity Evidence on Court Deferrals. Technical report, Working paper.

Neumark, D., Burn, I., and Button, P. (2019). Is It Harder for Older Workers to Find Jobs? New and Improved Evidence From a Field Experiment. *Journal of Political Economy*, 127(2):922–970.

Nix, J., Campbell, B. A., Byers, E. H., and Alpert, G. P. (2017). A Bird's Eye View of Civilians Killed by Police in 2015. *Criminology and Public Policy*, 16(1):309–340.

Nosenzo, D. (2013). Pay Secrecy and Effort Provision. *Economic Inquiry*, 51(3):1779–1794.

Obloj, T. and Zenger, T. (2020). The Influence of Pay Transparency on Inequity, Inequality, and the Performance-Basis of Pay in Organizations. *HEC Paris Research Paper No. SPE-2020-1359*.

Ohayon, M., Wickwire, E. M., Hirshkowitz, M., Albert, S. M., Avidan, A., Daly, F. J., Dauvilliers, Y., Ferri, R., Fung, C., Gozal, D., et al. (2017). National Sleep Foundation's Sleep Quality Recommendations: First Report. *Sleep health*, 3(1):6–19.

Owens, E., Weisburd, D., Amendola, K. L., and Alpert, G. P. (2018). Can You Build a Better Cop? Experimental Evidence on Supervision, Training, and Policing in the Community. *Criminology & Public Policy*, 17(1):41–87.

Park, K. H. (2017). The Impact of Judicial Elections in the Sentencing of Black Crime. *Journal of Human Resources*, 52(4):998–1031.

Parsons, C. A., Sulaeman, J., Yates, M. C., and Hamermesh, D. S. (2011). Strike Three: Discrimination, Incentives, and Evaluation. *American Economic Review*, 101(4):1410–35.

Patrick, Y., Lee, A., Raha, O., Pillai, K., Gupta, S., Sethi, S., Mukeshimana, F., Gerard, L., Moghal, M. U., Saleh, S. N., et al. (2017). Effects of Sleep Deprivation on Cognitive and Physical Performance in University Students. *Sleep and biological rhythms*, 15(3):217–225.

Paycheck Fairness Act (2019). H.R.7 116th Cong. (2019-2020). `https://www.congress.gov/bill/116th-congress/house-bill/7/all-info`. Accessed: 2020-12-30.

Pilcher, J. J. and Huffcutt, A. I. (1996). Effects of Sleep Deprivation on Performance: A Meta-Analysis. *Sleep*, 19(4):318–326.

Police Accountability Task Force (April 2016). Recommendations for Reform: Restoring Trust between the Chicago Police and the Communities they Serve. Technical report.

Rabin, M. (1993). Incorporating Fairness Into Game Theory and Economics. *The American Economic Review*, pages 1281–1302.

Rehavi, M. M. and Starr, S. B. (2014). Racial Disparity in Federal Criminal Sentences. *Journal of Political Economy*, 122(6):1320–1354.

Rim, N., Ba, B., and Rivera, R. (2020). Disparities in Police Award Nominations: Evidence from Chicago. *AEA Paper and Proceedings*, 110:447–451.

Rotemberg, J. J. (2006). Altruism, Reciprocity and Cooperation in the Workplace. *Handbook of the Economics of Giving, Altruism and Reciprocity*, 2:1371–1407.

Sarsons, H. (2019). Interpreting Signals in the Labor Market: Evidence from Medical Referrals. Working paper.

Sarsons, H. (2020). Gender Differences in Recognition for Group Work. *Journal of Political Economy*, Accepted.

Sen, A., Sen, M. A., Amartya, S., Foster, J. E., Foster, J. E., et al. (1997). *On Economic Inequality*. Oxford University Press.

Sklansky, D. A. (2005). Not Your Father's Police Department: Making Sense of the New Demographics of Law Enforcement. *J. Crim. L. & Criminology*, 96(3):1209–1243.

Smith, A. C. (2016). Spring Forward at Your Own Risk: Daylight Saving Time and Fatal Vehicle Crashes. *American Economic Journal: Applied Economics*, 8(2):65–91.

Tavernier, R. and Adam, E. K. (2017). Text Message Intervention Improves Objective Sleep Hours Among Adolescents: The Moderating Role of Race-Ethnicity. *Sleep health*, 3(1):62–67.

Thaler, R. H. and Johnson, E. J. (1990). Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. *Management Science*, 36(6):643–660.

Thomas, D. and Anderson, W. (2013). Multiple sleep latency test (mslt). In Kushida, C. A., editor, *Encyclopedia of Sleep*, pages 96–99. Academic Press, Waltham.

U.S. Department of Justice (January 13, 2017). Investigation of the Chicago Police Department. Technical report.

U.S. Department of Justice (March 4, 2015). Investigation of the Ferguson Police Department. Technical report, Civil Rights Office.

Venkatraman, V., Huettel, S. A., Chuah, L. Y., Payne, J. W., and Chee, M. W. (2011). Sleep Deprivation Biases the Neural Mechanisms Underlying Economic Preferences. *Journal of Neuroscience*, 31(10):3712–3718.

Weisburst, E. (2018). Whose Help Is on the Way? The Importance of Individual Police Officers in Law Enforcement. Working paper.

West, J. (2018). Racial Bias in Police Investigations. Working paper.

Worden, R. E., Kim, M., Harris, C. J., Pratte, M. A., Dorn, S. E., and Hyland, S. S. (2013). Intervention With Problem Officers: An Outcome Evaluation of an EIS Intervention. *Criminal Justice and Behavior*, 40(4):409–437.

Zeltzer, D. (2020). Gender Homophily in Referral Networks: Consequences for the Medicare Physician Earnings Gap. *American Economic Journal: Applied Economics*, 12(2):169–97.

Zizzo, D. J. and Oswald, A. J. (2001). Are People Willing to Pay to Reduce Others' Incomes? *Annales d'Economie et de Statistique*, pages 39–65.