Essays on Discrimination and Information Technology

by

Tamara Ren

A dissertation accepted and approved in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Economics

Dissertation Committee
Jonathan MV Davis, Chair
Michael Kuhn, Core Member
Kathleen Mullen, Core Member
Krystale Littlejohn, Institutional Representative

University of Oregon
Spring 2024

DISSERTATION ABSTRACT

Tamara Ren

Doctor of Philosophy in Economics

Title: Essays on Discrimination and Information Technology

This dissertation investigates how decision-makers, such as teachers or law enforcement agencies, may discriminate against individuals from different racial and gender demographic groups in the context of technology usage. It analyzes the impact of seeing identifying information (such as names, photos, and emails) and signals of quality (such as sharing information about one's past in an email) on decisions. This dissertation also explores how leveraging technology may reduce the potential for discrimination by using the features on digital platforms to conceal identifying information from decision-makers or to incentivize objectivity and fairness in assessments. This dissertation includes unpublished coauthored material.

CURRICULUM VITAE

NAME OF AUTHOR: Tamara Ren

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
University of Oklahoma, Norman, OK, USA

DEGREES AWARDED:

Master of Science, Economics, 2020, University of Oregon
Bachelor of Arts, Economics and Mathematics, 2014, University of Oklahoma

AREAS OF SPECIAL INTEREST:

Experimental Economics
Labor Economics
Behavioral Economics

PROFESSIONAL EXPERIENCE:

Graduate Educator, Economics Department, University of Oregon, 2020-2024
Marketing Analyst, Phillips 66, 2013-2018

GRANTS, AWARDS AND HONORS:

Oregon Consumer Justice/Oregon Law Award, 2023-2024
University of Oregon's Dissertation Award, University of Oregon, 2023
Alanson H. Kleinsorge Scholarship, University of Oregon, 2023
Edward G. Daniel Scholarship, University of Oregon, 2020
Graduate Teaching Fellowship, University of Oregon, 2020 - Present
Graduate Student Fellowship, University of Oregon, 2019 - 2020
National Science Foundation Summer Research, Tulane, 2012

ACKNOWLEDGEMENTS

I extend my heartfelt gratitude to my academic committee for their invaluable support and guidance throughout my doctoral journey. I am grateful to Jonathan Davis, whose unwavering encouragement and expert advice since my second year in the Ph.D. program have profoundly influenced my growth as a researcher in experimental economics. I am indebted to Michael Kuhn for his role as the Director of Graduate Students in encouraging me through the program. I also appreciate his collaborative approach in discussing research, theory, and academics. I am thankful of Kathleen Mullen, for her insightful feedback, particularly in framing my research within economic theory, which has been instrumental in shaping my work. I would like to express my appreciation to Krystale Littlejohn for valuing my research and for giving me her time.

Additionally, I extend my thanks to the graduate students and faculty of the economics department whose assistance with pre-grading materials for my online experiment significantly contributed to the success of my research. I am especially thankful to the faculty members who taught me economics, collaborated with me, and provided constructive feedback in my work. Special mention goes to Ed Rubin and Trudy Cameron for their invaluable insights and contributions to refining my research. Finally, I am grateful to the economics department for offering research funding opportunities, and I extend my thanks to the University of Oregon for the Dissertation Award, which funded my online experiment, and to the Law School for the Consumer Protection Grant, which supported my field experiment.

DEDICATION

To the many people who have been instrumental in my journey—my husband, for his unwavering support during the challenges of the program and the pandemic. His patience, understanding, humor, spontaneous trips, and constant encouragement have been my pillar of strength. To my parents, for instilling in me a sense of curiosity and a strong work ethic, especially my father, for always encouraging me to develop my mathematical abilities. I extend my heartfelt gratitude to all my family and friends who have stood by me, forgiven my many moments of absorption in research, and supported me near and far through every phase of this endeavor.

I also pay tribute to both of my grandmothers. To my grandmother who passed away when I was just three, whose determination to pursue higher education late in life before succumbing to ALS serves as a constant inspiration for me to complete my academic journey. To my other grandmother, who passed away during the program, but whose musical talent, humor, strength, and love for people filled my upbringing with music and laughter, inspire me to work hard, remain consistent and diligent in pursuing my own goals, while remembering to always share joy with others.

TABLE OF CONTENTS

Chapter Page

Chapter Page

APPENDICES

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

Table Page

CHAPTER I

OVERVIEW

This dissertation encompasses three chapters. To explore the role of technology and information on discrimination, I conduct two experiments, which are described in chapter two and chapter three. The first experiment is an online laboratory experiment that evaluates fairness in digital grading in an environment similar to learning management systems commonly used in schools. This experiment received an exemption from the Unviersity of Oregon's IRB (STUDY00000848). It was funded by the University of Oregon's Dissertation Award in 2023.

The second experiment is an audit study examining differences in email response rates and sentiment from law enforcement agencies after receiving email inquiries about firearm permits. This chapter is co-authored with Garrett Stanford. This experiment received IRB approval (STUDY00000779) from the University of Oregon. It was also funded by the University of Oregon's Law School's Consumer-Protection Grant in 2023.

In chapter four, I introduce a novel algorithm that uses only Google Trends data to access regional variations in internet search trends over time for more than five regions. The algorithm allows researchers to access more robust data on population-wide search interests.

CHAPTER II

THE ROLE OF INFORMATION AND ATTENTION IN DISCRIMINATION: EVIDENCE FROM A GRADING EXPERIMENT

2.1 Introduction

Online technology has transformed the way students receive their education. In recent years, especially during the COVID-19 pandemic, schools of all grade levels have adopted sophisticated online communication technologies, such as Learning Management Systems (LMSs), for classroom management. With LMSs, teachers and students often rely on remote interactions, even for test-taking or submitting homework. A unique feature of LMSs like Canvas is the ability for the teacher to grade digitally. Unlike grading paper submissions, grading in a digital space allows teachers more access to students' information, such as the student's name, profile photo, and even a web link on the student's homework submission to the student's past performance within the class. Another unique feature of LMSs is that if teachers want to, they can change the grading setting to automatically de-identify all homework submissions into a blind-grading environment. This ability to grade de-identified work raises the question of whether changing the grading environment into a blind one can promote equity in grades. In this paper, I provide causal evidence from an online laboratory experiment that the ability of LMSs to de-identify homework submissions may reduce discrimination in grading.

In the experiment, I recruit five hundred research participants from Prolific to evaluate multiple essays in a Qualtrics survey.[1] The grading environment in the survey simulates the grading experience on an LMS like Canvas. All the

[1] Peer, Rothschild, Gordon, Evernden, and Damer (2021) finds that research subjects from Prolific produce some of the best quality results relative to any other online platform for academic research.

essays answer a homework question that a teacher might ask in an introductory economics course. The essays participants grade are randomly drawn from a larger pool of essays and include randomly assigned student demographic information in the form of a photo and name that denote race/ethnicity/ethnicity and gender. Treatment is determined by the grading environment in which participants are randomly assigned: they are either in a non-blind (treatment group) or a blind grading (control group) environment. Additionally, each participant is randomly assigned a financial incentive between five and forty-five cents (called 'accuracy incentive') they can earn for assigning a grade to each essay that closely matches an unknown pre-determined benchmark, regardless of treatment status. Based on this design, there should be no differences in grades between the grading environments or demographic groups in the non-blind grading environment unless bias or the accuracy incentive influences how participants grade.

The results reveal that graders can adjust how they assign grades based on their knowledge of the students' racial/ethnic/ethnic or gender identities and whether they are compensated for grading accurately. Participants assign points more generously when they can see a name and photo of the students, especially at lower accuracy incentives. As the accuracy incentives increase, the grade gap between the non-blind and blind grading environments decreases, suggesting participants pay more attention to the signals of essay quality at higher accuracy incentives.

Heterogeneity in the results shows that students' putative racial/ethnic/ethnic identities in the non-blind grading environment differ significantly from the unknown identity in the blind grading environment when participants are assigned a low-value accuracy incentive. At 5 cents, the lowest accuracy incentive offered to participants, Asian students receive the highest average grades (0.459 points or 9.1 pp, significant at the 1 % level), followed by white (0.305 points or 6.1 pp, significant at the 1% level) and Hispanic students (0.271 points or 5.4 pp, significant at the 10% level), with Black students receiving the lowest grades (and statistically insignificant) in the non-blind grading environment. This order of results aligns with prior research on biases in teachers' assessments of students' abilities, as shown in Shi and Zhu (2023). The most prominent and statistically significant grade gap between non-white and white students at 5 cents is the Black-white gap at -0.21 points (about 4.2 pp difference, p-value 0.055).

Yet, when the value of the accuracy incentives increase, all grade gaps between the students' racial/ethnic groups decrease. If participants are assigned an accuracy incentive of 45 cents, the largest grade gap is the Black-white gap at 0.094 points and is statistically insignificant (p-value = 0.38). There are no significant gender differences in grades, nor do the accuracy incentives significantly impact the grades given to the unknown identity in the blind grading environment.

Given the design of the experiment, I also evaluate the role of learning how to grade the essays on the assigned scores. Specifically, I evaluate if there are differences in grades on the first graded essay relative to the last (i.e., essay order effect). This part of the experiment is motivated by Hanna and Linden (2012), who found in a grading experiment that grade gaps between demographic groups are largest on the first graded exams, likely due to the teachers relying on their knowledge of the students' demographics to help them assign grades when they start grading. As the teachers grade subsequent submissions, they learn the distribution of the tests and rely less on the students' observable characteristics for allocating partial credit. However, the results from this study suggest that, when pooling all of the data, there are no essay order effects. In comparing average grades given on the first essay relative to the last essay graded, there are no differences in grades relative to the unknown identity in the blind grading environment, nor are there differences in grades between the racial/ethnic/ethnic or gender groups in the non-blind grading environment.

I investigate whether the order effects are stronger for smaller accuracy incentives than those assigned larger ones by conducting a triple difference that estimates the order effects at the different accuracy incentives. I find strong evidence that the Black-white grade gap is driven by discrimination in the first essay when participants are assigned an accuracy incentive of 5 cents. The Black-white gap is -0.525 points or a 10.1 pp difference (0.2 SE, p-value=0.012). However, those assigned an accuracy incentive of 45 cents grade all student racial/ethnic/ethnic groups equally on the first essay. In the last essay, all grade gaps disappear regardless of the assigned accuracy incentive, except for the Asian-white gap for those assigned the 5-cent accuracy incentive (significant at the 10 % level).[2]

What do these results suggest? Given that the design of this experiment is such that the quality of the essays across the students' putative identities, essay order, accuracy incentive, and grading environments are randomly assigned. Thus, uncorrelated differences in the assigned grades at small but not at large incentives in the first essay suggest a behavioral difference when participants are paid more for their effort. In a theoretical model developed to motivate the experiment, I show that inattention and statistical discrimination can explain this behavioral

[2] This part of the study is exploratory, as I did not pre-register a triple difference. Only after seeing the results did I think of doing so.

difference. Following Phelps (1972) and Maćkowiak, Matějka, and Wiederholt (2023), the theoretical model shows how blind grading can reduce discrimination in grading and that paying graders for their effort to learn the quality of the student's work increases graders attention to learning the underlying quality distribution of the student's homework. The experimental results indicate this behavior. Graders exhibit behavior associated with learning the distribution, where discrimination is most significant in the first essay at smaller incentives but is reduced by the last, and discrimination disappears earlier in the grading sequence at larger incentives.

To further explore the relationship between effort, the grading environment, and the accuracy incentives, I evaluate other grading behaviors related to effort. The behaviors the Qualtrics survey measures are the time spent grading, the usage of the grading rubric or class notes on each essay, and the number of clicks per page. I assess whether these behaviors are related to the grading environment, accuracy incentive, and essay order. I generally find no differences in these behaviors relative to the unknown identity or between the racial/ethnic/gender identities. Regardless of treatment status, all participants spend more time grading the first essay than the last. They use the grading material similarly and put in similar levels of effort. Even participants' self-reported responses to questions about their experience with the grading material and their fatigue reveal no significant differences between treatment groups.

In this experiment, I find evidence of racial/ethnic/ethnic discrimination in grading and the underlying behaviors driving that discrimination. However, the results may be understated and conservative due to skin color discrimination and informing participants that the students and the material are fictitious.

If graders assigned points based on other characteristics like skin color, then discrimination in grading may be more pronounced than what is tested in the experiment. Skin color discrimination can show up as differential treatment towards people with lighter or darker skin tones, even within the same racial/ethnic/ethnic group. Depending on the skin color, graders may perceive the student to be of a different race/ethnicity or ethnicity, which may change how they view the student's work and impact their grading.

For example, the most significant grade gap between non-white and white students in this experiment is the Black-white gap. Although the quality of the essays is the same, I find that Black students receive lower grades on average than white students, dependent on the accuracy incentive and the order in which they appear in the essay submissions. However, graders may be more generous in grading if they believe the student is Indian rather than Black, based purely on only looking at the skin color of the student in the profile photo on the essay. If this is the case, then the Black-white gap may be larger. However, I do not ask participants at the end of the experiment to provide their beliefs about the observed students' race/ethnicity/ethnicity or gender, so I cannot test for this explicitly.

Additionally, the results may be more conservative due to participants being told that the students/essays are fictitious in the instructions. I deliberately told participants about the fictitious nature of the material for human subject considerations with the IRB and to avoid deception by following standards typically used in experimental economics. If participants do not know that the students are not real, it may change how they assign the grades and even their other grading behaviors, like time spent grading or using a grading rubric. I do not ask participants at the end of the experiment if they know the students are fictitious, so I cannot measure the extent of that knowledge on their behaviors.

To the best of my knowledge, this paper represents the first to offer causal empirical evidence of how blind grading can reduce the potential for discrimination in grading in a class setting. This paper is also the first to provide evidence that paying graders for their efforts can reduce grading discrimination. It is also the first to quantify differences in grading behaviors related to effort, such as time and usage of supplementary grading material, when grading in non-blind and blind-grading environments. Furthermore, this paper stands among the first to provide causal evidence that implementing a blind grading policy within Learning Management Systems may be an effective strategy for educational institutions to mitigate grading discrimination that might arise when graders observe signals related to students' identities.

This paper is structured as follows. The second section presents the motivation and the theoretical framework of the experiment. The third section details the experimental design and the empirical specification. The fourth section provides the data and summary statistics, and the fifth section discusses the experimental results. The paper concludes with a summary of the experiment and the results.

Related Literature. This experiment offers causal evidence that blind grading can effectively reduce bias in grading when that bias is due to observing students' putative identities. This paper contributes to the extensive literature on discrimination in education, particularly the impact of teachers' biases on grade outcomes. While previous research has shown such biases, this study stands out as one of the few conducted in an experimental setting.

The seminal work by Lavy (2008) reveals that average grades may be biased when comparing grades by students' demographics with those of externally blindly graded exams. Subsequent research has further identified racial/ethnic discrimination (**?**), ethnic discrimination (Alesina, Carlana, Ferrara, & Pinotti, 2018), and gender discrimination (Terrier, 2020). Blind grading has been shown to benefit underrepresented students (Jansson & Tyrefors, 2022), Breda and Ly (2015), or have no impact on students Hinnerich, Höglin, and Johannesson (2011). Field experiments also indicate that assigned grades can be influenced by teachers' in-group bias in favor (Feld, Salamanca, & Hamermesh, 2016) or against Hanna and Linden (2012) students of the same demographics. Additionally, teachers' generational demographics may play a role in grading bias (Nguyen, Nhu, Ost, Ben, & Qureshi, Javaeria, 2023). Shi and Zhu (2023) finds that teachers may over-estimate the abilities of Asian and White students' and under-estimate Hispanic and Black students' math and reading skills compared to results from blindly-graded exams. Implementing a blind grading policy may be an effective way to enhance school equity, given that teachers' biases (Carlana, 2019; Doornkamp et al., 2022) or biases in grades (e.g., Lavy & Megalokonomou, 2019; Lavy & Sand, 2018; Protivínský & Münich, 2018) can predict future academic performance and educational attainment gaps.

This paper contributes to the growing body of literature examining the impact of technology on students' outcomes. It delves into how teachers may approach grading in digital learning environments. While distance learning has been associated with reduced student achievement and attendance (e.g., Bettinger, Fox, Loeb, & Taylor, 2017; Cacault, Hildebrand, Laurent-Lucchetti, & Pellizzari, 2021; Kofoed, Gebhart, Gilmore, & Moschitto, 2021), how teachers conduct grading in online settings is a less explored area. Some initial evidence suggests that teachers' biases may decrease in online classes Mehic (2022).

This paper explores how an incentive strategy to increase grading accuracy can promote equity in grades, contributing to the literature on using pay-for-performance strategies to improve educational outcomes. Financial incentives have been shown to increase teachers' attendance Duflo, Hanna, and Ryan (2012), enhance teachers' quality Biasi (2021), increase teachers' effort Leaver, Ozier, Serneels, and Zeitlin (2021), and increase students' academic performance (Biasi, 2021; Duflo et al., 2012; Leaver et al., 2021).

This paper delves into the underlying mechanisms that drive biases when uncertain decision-makers must assess the quality of an economic agent based on signals of ability. It contributes to research on selective attention (Schwartzstein, 2014), attention discrimination (e.g., Bartoš, Bauer, Chytilová, & Matějka, 2016; Huang, Li, Lin, Tai, & Zhou, 2021), and rational inattention theory, drawn from from macroeconomics (e.g., Maćkowiak et al., 2023; Sims, 2003) and behavioral economics (Gabaix, 2019). My findings reveal that graders may discriminate in a way consistent with statistical discrimination. Average grades reflect the ordering found in Shi and Zhu (2023) at lower incentives. As participants grade each subsequent essay, discrimination decreases in a way suggesting that graders learn and update in a dynamic way (Bohren, Imas, & Rosenberg, 2019).

2.2 Motivation, Theory and Hypotheses

Motivation. In the United States, academic accessibility and performance are often linked to students' demographics, with Black, Hispanic, and Native American students consistently facing educational disparities (de Brey et al., 2019). If the grading environment influences how teachers assess students, LMSs could offer tools to enhance fairness in grades and improve the equity of students' educational outcomes, such as blind grading, where students' information is de-identified.

Why might de-identifying students' information during grading improve fairness in grading? Consider an essay where one student receives an 'A' while another receives a 'B.' It is assumed that the 'A' student produced a superior essay, but this may only sometimes be true. For instance, people often use cognitive shortcuts to reduce decision fatigue, as shown in the seminal work by Tversky and Kahneman (1974). Teachers might inadvertently rely on preconceived beliefs about students' identities to expedite their grading, especially when dealing with subjective materials like essays or when facing many homework submissions to grade in a single night. If teachers rely on these beliefs to aid their grading, observing students' identities may introduce grading bias. This bias could lead to a better or worse grade for an essay of the same quality, depending on the teacher's beliefs about the student's demographic.

Following Phelps (1972)'s statistical discrimination framework, teachers should learn the quality distribution of the student's work as they develop experience and learn more about their students. To examine the relationship between effort (or attention) and discrimination in grading, I introduce rational inattention theory into the theoretical framework. [3]. In this framework, teachers choose which information in the material to prioritize when determining how to assign grades. These choices are driven by their uncertainties about the quality of work of students from different demographic groups and the costs associated with grading, which could include the time that could be spent on other activities.

[3] For review, see (Gabaix, 2019; Maćkowiak et al., 2023)

The theoretical model predicts that when teachers statistically discriminate, they may be slower to update their beliefs, potentially leading to persistent grading discrimination based on beliefs that do not reflect the true distribution of the students' work quality. However, if teachers are compensated for their efforts, they pay attention to more information and update their beliefs more quickly, ultimately reducing grading discrimination.[4]

Theoretical Model. I model how a teacher grades in a blind or non-blind setting under uncertainty. The teacher is Bayesian and anchors how they grade based on their prior beliefs. The score the teacher, or the decision-maker (DM), allocates to a student's essay is denoted by g. The DM is rational and only experiences psychic costs with grading (i.e., the mental exhaustion of grading and the opportunity costs of spending time grading instead of other leisurely activities).

The DM maximizes the expected payoff between assigning a grade that reflects the student's ability and the cost of mentally assessing the homework submission (i.e., essay) to learn the quality of the student's work. The expected payoff considers psychological feelings like guilt that a teacher may experience from inadvertent unfairness in how they graded a student's submission. The costs could include the cognitive effort of deciphering the quality of the submission and the opportunity cost of the time spent grading or seeking information on how to fairly grade, such as reviewing the grading rubric or class lecture notes.

The model is analogous to the standard model of statistical discrimination in labor markets [i.e., (Phelps, 1972)], with two main differences: (i) beliefs are

[4] Although I do not discuss this in detail in the paper, the other type of discrimination that could happen is based on preferences Becker (1971). If graders exhibit preference-based discrimination, grade disparities persist even when teachers learn about their students' abilities. These gaps stem from teachers' biases against different student groups rather than their beliefs about student capabilities. I do not go into this theory as much of the literature on discrimination in the classroom finds evidence of statistical discrimination.

biased (i.e., unequal) and (ii) the level of attention the DM pays to the essay is endogenous. Even with the availability of perfect information, discrimination in grading can persist due to imperfectly learning signals of the quality of a student's work. This model follows a simple rational inattention model from behavioral economics (Gabaix, 2019).

Furthermore, I show in this model that discrimination can arise even if there are no population differences in the quality of work between demographic groups. I assume that the quality of the student's work is normally distributed $N(\theta, \sigma_\theta^2)$ and does not vary across demographic groups (such as race/ethnicity, gender, or other observable demographics). Additionally, I define discrimination in grading as the systematic difference in grades between two demographic groups, given the same quality of work. The model is shown in two parts: (1) the general setup, which outlines the DM's choice structure, and (2) the DM's problem.

General Setup. Let a student produce some quality of work denoted by θ that is unknown to the decision maker (DM). The student submits his or her homework, such as an essay, to the DM who grades it. The DM faces two choices. First, how much of the student's essay to read and consider carefully. This choice affects the precision of the DM's knowledge about the student's quality of work, θ. Second, what grade, g, should be allocated.

In a non-blind grading setting, I assume only two demographic groups exist. Let $d \in \{A, B\}$ where A denotes group A, and B denotes group B. The DM holds prior knowledge about θ. The DM's prior beliefs about the group quality, $\theta|d$, is normally distributed according to $\theta|d \sim N(\theta_d, \sigma_d^2)$. Prior beliefs can reflect bias or unwarranted differential perceptions about the two groups. In this model, bias is denoted by $\theta_A \neq \theta_B$. The precision of those prior beliefs, σ_A^2 or σ_B^2, depends on the DM's certainty about the distribution of the quality within each demographic group. For simplicity, I assume that $\sigma^2_A = \sigma^2_B$, although this assumption can be relaxed.

In a blind-grading setting where signals of the students' demographics have been removed from the grading environment, the DM cannot observe characteristics such as gender. Instead, the identity of the student is *unknown*. Let a denote a student whose work is being graded in a blind-grading setting. The DM's prior beliefs about the quality of a student in a blind-grading setting is normally distributed according to $\theta \sim N(\theta_a, \sigma^2_a)$. The DM holds prior knowledge about θ_u from the expected quality of group, d, $\theta_u \equiv E[\theta|d]$. The uncertainty of the DM's prior beliefs about a student in the blind-grading setting follows the law of total variance. By definition, the DM will hold more uncertainty about the quality of her students' work when grading in a blind setting, $\sigma^2_u > \sigma^2_B$:

$$\sigma^2_u = V[E[\theta|d]] + E[V[\theta|d]] \tag{2.1}$$

Each submission by a student is a noisy signal, $s = \theta + \epsilon$, concerning the student's innate ability, θ, where ϵ is normally distributed, $\epsilon \sim N(0, \sigma^2_\epsilon)$. Upon receiving signals of the student's demographic group, $i \in \{u, A, B\}$, where a denotes unknown gender, A denotes group A, and B denotes group B, the DM's posterior belief about θ is given by the distribution, $N((1-\phi_i)\theta_i + \phi_i s, \phi_i \sigma^2_s)$ and $\phi = \frac{\sigma^2_i}{\sigma^2_i + \sigma^2_\epsilon} \in [0, 1]$. The parameter ϕ denotes the fraction of attention a DM devotes to a given essay.

Regardless of whether the DM is grading in a non-blind or blind-grading setting, the DM chooses how much attention, ϕ, to devote to the content of the student's essay. The DM chooses ϕ by learning the precision, σ^2_ϵ. The more attention the DM pays to the submission, the more information the DM acquires about θ from reducing σ_ϵ^2.

Implication: Discrimination decreases when ϕ increases. This is because grades reflect the quality of the student's work when ϕ is larger, i.e., there is more weight on the work and less weight on the grader's beliefs about the student's demographics.

The DM's problem. The DM sets a grade, g, equal to the expected utility of the student's ability, θ. I assume the DM's utility is derived from grading accurately involves a quadratic-loss function, $-\frac{1}{2}(g - \theta)^2$, which is a common functional form to approximate the loss used for mathematical tractability. This psychic cost implies the DM experiences an increasingly negative utility if g, the grade they assign, is unequal to the true quality of the student's work, θ.

The DM maximizes the expected quadratic-loss function $E[-\frac{1}{2}(g - \theta)^2|s]$. The DM's utility-maximizing grade for submission s is $g(s) = E[\theta|s]$ where $E[\theta|s] = (1 - \phi_i)\theta_i + \phi_i s$ for $i \in \{u, A, B\}$. The DM's expected utility loss from inattentive grading of a given homework submission, s, is:

$$-\frac{1}{2}(1 - \phi)\sigma_i^2, \tag{2.2}$$

where σ_i^2 is the variance of the DM's prior beliefs derived from the student's demographic group membership.

Attention is mentally costly, so the DM experiences a cost for the mental effort required for grading. Denoting this cost of effort as a function, $C(\phi)$, where ϕ is the parameter of the level of attention given to a student's essay. I assume for simplicity that $C(\phi)$ is a convex function in ϕ, so $C'(\phi) > 0$ and $C''(\phi) \geq 0$. In other words, the more attention devoted to submission s, the greater the cost of effort to the DM.

The DM optimizes her level of attention based on choosing a level of attention ϕ, and the actual cost of paying attention to the essay.

$$max_{\phi \in [0,1]} - \frac{1}{2}(1-\phi)\sigma_i^2 - C(\phi) \tag{2.3}$$

The DM's optimal attention level satisfies the following first-order condition:

$$C'(\phi^{\star}) = \frac{1}{2}\sigma_i^2 \tag{2.4}$$

Implication: An increase to σ_i^2 decreases discrimination due to attention, ϕ, increasing.

Compensating the cost of grading effort. The cost of paying attention, $C(\phi)$, includes opportunity costs for the time and effort to grade. If one were to compensate the DM for accurate grading, ϕ increases, and discrimination is reduced.

Let the DM receive some compensation, P, to grade accurately. Then, the grader's problem becomes:

$$max_{\phi \in [0,1]} - \frac{1}{2}(1-\phi)\sigma_i^2 - C(\phi) + \phi \times P \tag{2.5}$$

The DM's optimal attention level is now influenced by the value of P, given the following first-order condition:

$$C'(\phi^{\star}) = \frac{1}{2}\sigma_i^2 + P \tag{2.6}$$

Implication: If P or σ_i^2 increases (decreases), then ϕ increases (decreases). Hence, compensating graders for their efforts should decrease discrimination.

Testable Hypotheses. Given the implications of the theoretical model, I design a grading experiment to test whether graders statistically discriminate and rely on their beliefs to grade when they are exposed to salient signals of the students' demographics while grading. In particular, I test the following hypotheses:

1. Blind grading reduces discrimination.
2. Compensating graders for their effort reduces discrimination in a non-blind grading environment.
3. Grading behaviors that indicate grading effort differ across the blind and non-blind grading environments.
   (a) Graders who are exposed to demographic signals spend less time assessing a student's essay relative to graders who grade in a blind environment and do not see any demographic signals.
   (b) Graders rely less often on a grading rubric and class notes when exposed to demographic signals, relative to graders who grade in a blind-grading environment and do not see any demographic signals.

2.3 Experimental Methodology and Data

The experiment's design consists of three parts: the essays to be graded, the survey, and the "student" characteristics. Each component is described in depth below.

The Grading Material. For the experiment, I employed ChatGPT 3.0 to generate a diverse range of twenty distinct one-paragraph responses to a homework question. This question, which one might encounter in a university-level or secondary-level economics course, is: 'Should we tax companies that use artificial intelligence in place of human workers?'

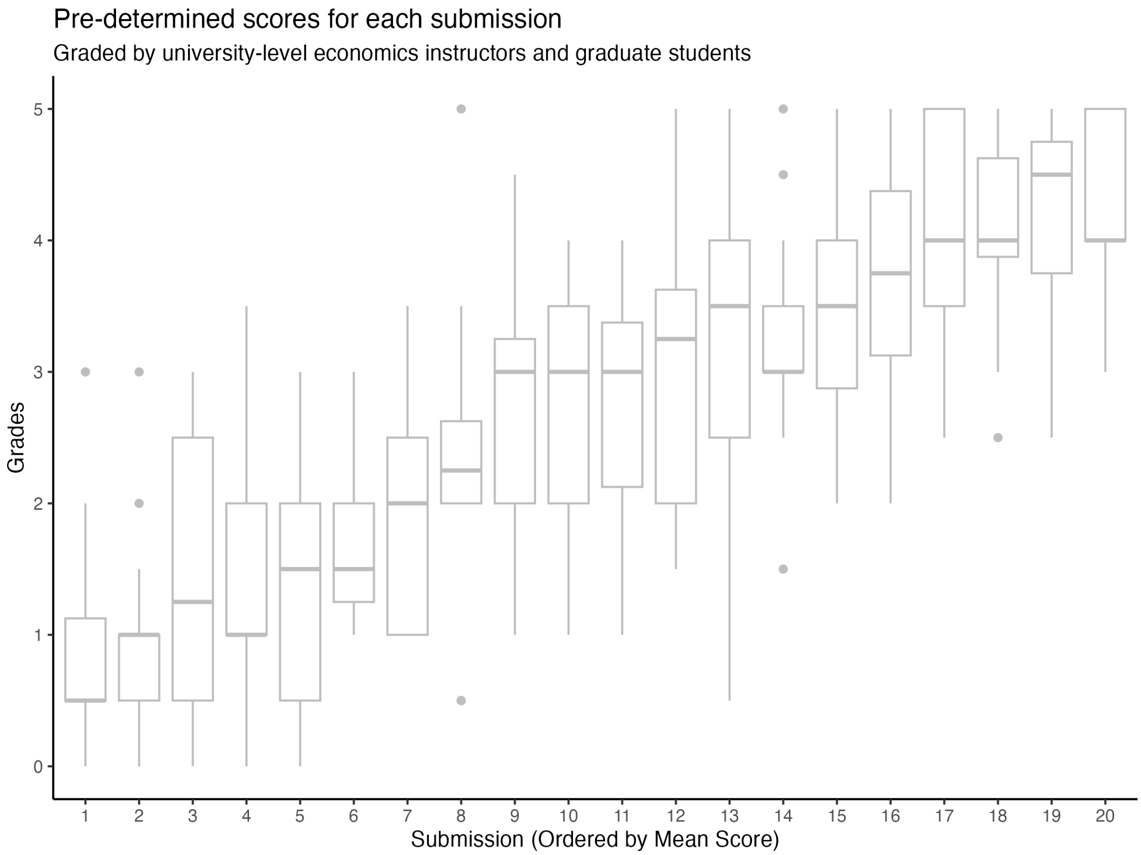
Before launching the experiment, I recruited twenty economics faculty members and graduate students at a large public university to determine the quality of each essay on a scale between zero and five. These volunteers were told to grade the essays as though they were from real students. The volunteers graded the essays using a different Qualtrics survey from the main experiment. The Qualtrics survey randomly ordered the essays to ensure multiple people graded each. [5] Additionally, I provided these volunteers with a grading rubric and a short list of the pros and cons associated with the tax, described as "class notes."

The mean number of essays graded by each person is 16.75. The average score of these preliminary grades, denoted as "pre-determined grades," differs across all twenty essays. Figure 1 illustrates the distribution of these scores by each essay with whisker-box plots. The ordering of the whisker-box plots in Figure 1 is based on the average grades the volunteers gave each essay. The line in the middle of the boxes is the median grade; the hinges (or the lines at the top and bottom of the boxes) refer to the 25th and 75th quartile values. The whiskers, the vertical lines spanning out of the boxes, are based on the distance between the first and third quartiles. The points are outliers.

I used the Grammarly software package to measure the grammar and writing characteristics of the essays prior to launching the experiment. Grammarly

[5] The volunteers were asked to grade all 20 but could opt-out anytime.

Figure 1 Distribution of the Pre-determined Grades



This graph shows the distribution of the pre-determined grades by each essay in whisker-box plots, ordered from smallest to largest mean scores. Twenty university-level instructors and graduate students graded at least five submissions. The line in the middle of the boxes is the median grade; the hinges (or the lines at the top and bottom of the boxes) refer to the 25th and 75th quartile values. The whiskers, the vertical lines spanning out of the boxes, are based on the distance between the first and third quartiles. The points are outliers. The average number of submissions graded by each person was 16.75 essays. All essays were randomly ordered to ensure all essays were graded by multiple people. All graders in this exercise were told the purpose of how these values would be used in the main experiment.

can determine the word count, the estimated length of time to read, the number of sentences, the readability (i.e., Grammarly provides a score of how easy it is to read a text passage—the larger the number, the easier the passage is to read), and the overall grammar/writing score on each submission. The score is lower if there are typos and grammatical mistakes in the passage and higher if the text has fewer grammatical mistakes. Table 1 reports the Grammarly data.

Participants could refer to a grading rubric and class notes that are used in the main experiment. The grading rubric outlines the proportion of the points that should be allocated based on different criteria, and the "class notes" are a short list of the pros and cons of an AI tax. Figure 2 shows the grading rubric. It lists the "homework" question and how to allocate the points based on the "students" understanding of the prompt, reasoning, analysis of the consequences, and the clarity of their writing.

Table 1 Some Descriptive Statistics of the Essay Characteristics

	N	Mean	Std.Dev	Min	Max
Word Count	20	83.65	38.601	23	164
Sentences	20	4.2	1.399	2	7
Reading Time (sec)	20	19.55	9.168	6	39
Readability of Text	20	46.65	26.432	14	100
Overall Grammar Score	20	82.5	13.582	42	100

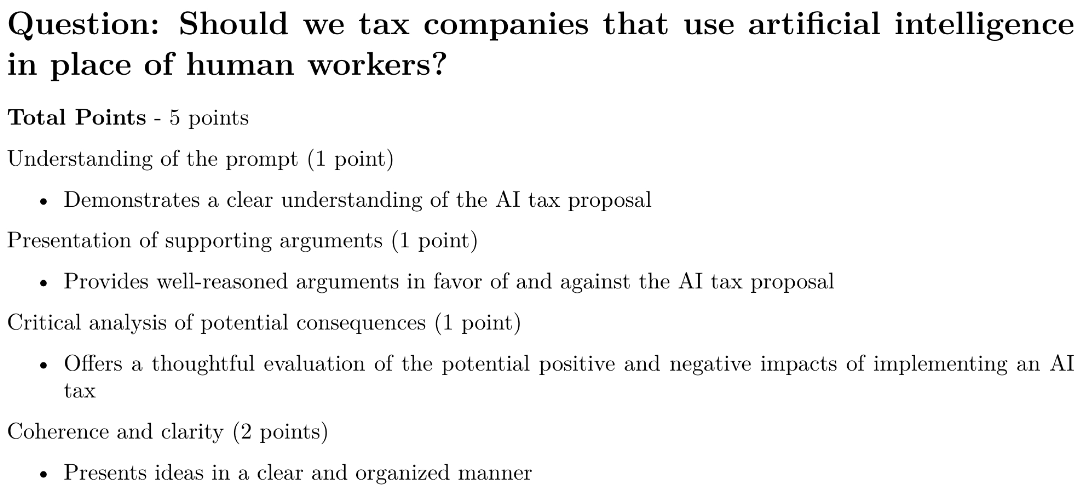
Note: This table reports the distribution of some essay characteristics in the text used in the experiment. The grading rubric I use in the experiment directs the participant to allocate some of the five possible points based on the grammar of the writing. To identify if there was variation in the writing quality of the twenty essays, I use the Grammarly platform to measure the writing characteristics of the essays. *Word Count*: the number of words. *Sentences*: the number of sentences. *Reading Time (sec)*: expected number of seconds required to read each essay. *Readability of Text*: the required comprehension level to read the text, where larger numbers means the text can be read by people with lower reading levels. *Overall Grammar Score*: an index value based on the grammar. A value of 100 means there are no mistakes in the text.

The Grading Session on Qualtrics.

Survey Design. Figure 3 illustrates the Qualtrics survey's experimental design and the randomization points. Reading from top to bottom, the boxes with no background color show the chronological steps of the survey. The dark green boxes denote the type of information randomized.

When a participant opens the Qualtrics survey, the survey randomly assigns them an accuracy incentive amount they could earn as a bonus, a set of essays to grade, and the student identities associated with those essays. The accuracy incentive ranges from five to forty-five cents in ten-cent increments and is assigned only once from a uniform distribution with equal probability on each value. Participants can earn a bonus based on their assigned accuracy incentive if they give a grade within half a point of the specific essay's pre-determined grade. Before the grading portion of the survey, participants answer questions to ensure they understand how they can earn the accuracy incentive.

Figure 2 The Grading Rubric



Question: Should we tax companies that use artificial intelligence in place of human workers?

Total Points - 5 points

Understanding of the prompt (1 point)

- Demonstrates a clear understanding of the AI tax proposal

Presentation of supporting arguments (1 point)

- Provides well-reasoned arguments in favor of and against the AI tax proposal

Critical analysis of potential consequences (1 point)

- Offers a thoughtful evaluation of the potential positive and negative impacts of implementing an AI tax

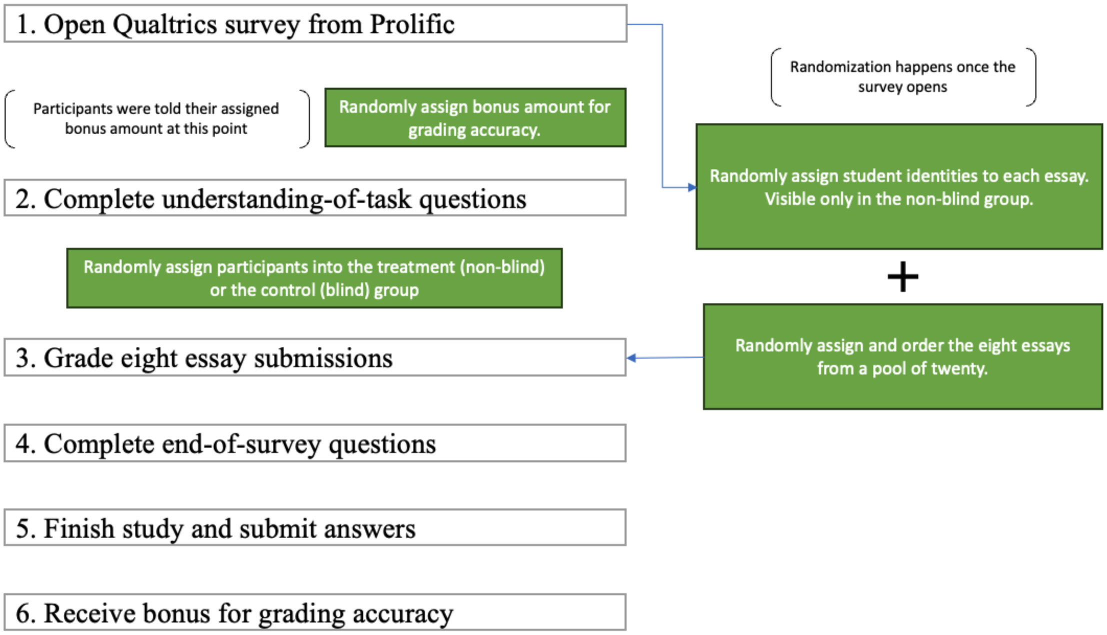
Coherence and clarity (2 points)

- Presents ideas in a clear and organized manner

Note: Participants could look at the rubric and class notes at any time during the experiment. These were available to them within the Qualtrics survey. To access the information, participants had to click a button to reveal the photo.

Once participants finish reading through the instructions and answering the questions on their task, the survey randomly assigns them to a non-blind or blind grading environment. I do not inform the participants of the different

Figure 3 Summary of the Experimental Design


1. Open Qualtrics survey from Prolific
Participants were told their assigned bonus amount at this point
Randomly assign bonus amount for grading accuracy.
2. Complete understanding-of-task questions
Randomly assign participants into the treatment (non-blind) or the control (blind) group
3. Grade eight essay submissions
4. Complete end-of-survey questions
5. Finish study and submit answers
6. Receive bonus for grading accuracy
Randomization happens once the survey opens
Randomly assign student identities to each essay. Visible only in the non-blind group.
+
Randomly assign and order the eight essays from a pool of twenty.

Note: This figure shows a visual summary of the experimental design. It begins with the participant opening the study in Prolific to completing the survey. The green boxes refer to the point at which information is randomized.

grading environments, nor do I publicize them at any stage in the experiment of the different environments.

Regardless of the grading environment assigned to the participants, they grade eight essays, one at a time. This is a common practice on LMS platforms like Canvas, where graders will assess each homework submission chronologically. The essays are randomly drawn from twenty, each with an equal chance of being drawn once. In the non-blind group, each essay includes a randomly assigned student name and photo, indicating a race/ethnicity and ethnicity (Asian, Black, Hispanic, or White) and gender (female or male). In the blind grading group, participants are shown only the essays with the following text replacing the name and photo: "Student [one through eight]."

As the participants grade the essays, they cannot go back and change their answers. Although real-life graders may do this, this restriction is necessary to measure the grading behaviors associated with effort. Moreover, participants do not know of their grading accuracy. When participants assign a grade, they are not shown or told whether that value is within the half-point of the pre-determined grade. They had to wait until after the experiment to receive their bonuses.

After finishing the grading portion of the survey, I ask participants four debriefing questions concerning their grading experience before requesting their demographic information. The debriefing questions are on their level of agreement with statements related to grading fatigue, the usefulness of the grading rubric or class note, and the ease of learning how to allocate points throughout the study. After answering these questions and volunteering their demographic information, participants complete the survey and submit their completion status to Prolific. Prolific then registers their completion, and I pay the participants their bonuses within three days of completing the experiment.

Prolific. Participants are from the Prolific platform. Prolific is an online platform that anonymously and randomly recruits individuals to participate in independent research. Anyone in any country where Prolific is present can join Prolific as a test subject as long as they are at least eighteen. Considered one of the best platforms for ensuring the quality of the recruited research participants, Prolific randomly recruits participants from its general population or some criteria set by the researchers (Peer et al., 2021).

In this experiment, I request that Prolific recruit an equal number of male and female participants. These 500 individuals have completed at least some higher education, speak English fluently, and live in the United States. On the Prolific application, I inform participants that the purpose of the study is to help improve our understanding of grading discretion on online platforms and explore what it is like to grade online. I also explain their task (i.e., "Your task is to grade eight, short homework submissions from fictitious students as if you are a teacher," and "There are no requirements for completing this survey. Simply allocate points you think each homework submission deserves."). I told participants in the Prolific submission that they will receive $1.75 for completing the study and that the study includes bonuses.

Five hundred twenty-eight people clicked the link to participate in the study. Twenty-six dropped out, and two timed out. When an individual drops out or times out from the survey, Prolific replaces their spot with another person until the five hundred cap is met. The correlation between completing the study and the accuracy incentive is 0.01, indicating that the value of the accuracy incentive does not contribute to participants dropping out.

Randomizing student identities. Only those participants in the treatment group (the non-blind grading environment) can see the putative student identities on each essay. The Qualtrics platform separately randomizes the students' demographic characteristics and each essay in the treatment group (non-blind grading). This randomization ensures that the student characteristics are uncorrelated with the quality of the essay—in fact, the distribution of the essay quality for the different demographic groups of the students is, by design, no different. I use both names and small, round profile photos in each essay to signal the students' putative demographics. Although much research that studies discrimination in labor economics typically uses only names, I include photos and names to mimic the features of the online grading environment in LMS platforms.

The students' putative identities in the experiment are two genders—female and male—and four racial/ethnic/ethnic groups—Asian, Black, Hispanic, and white.[6] Although the use of these four racial/ethnic groups increases the number of observed identities and reduces statistical power for any number of participants, these proportions reflect the proportions of these groups in most colleges in the United States. The photos are chosen from generated.photos so that facial features are similar across all racial/ethnic groups and genders(Generated Photos, 2023).[7] Each photo is randomly assigned a name once that reflects the race/ethnicity and gender of the photo.

In total, there are forty possible identities. There are sixteen possible white identities split evenly by gender. For the Asian, Black, and Hispanic identities, there are eight putative identities also split evenly by gender. The survey randomly assigns the identities so each participant in the non-blind grading environment always sees five white and three non-white students. The primary reason for this proportion is that it reflects the proportion of white and non-white students in colleges within the United States. The proportion of the assigned identities on the essay submissions in this study is 62.5% white, with the remaining 37.5% split evenly between Asian, Black, and Hispanic identities.[8]

Data Description. The data collected are the grades for each essay, the treatment status, the value of the accuracy incentives, the students' observed demographics, and the essays' grading order. The participants could allocate values

[6] For simplicity, I refer to the racial/ethnic/ethnic groups as racial/ethnic groups in this paper.

[7] generated.photos is a web service that provides academic researchers free access to their AI-generated photo database.

[8] Since the year 2000, white students have been the majority enrolled in a degree-granting institution, although this proportion has declined from 70% to 56%. The proportion of Black students has been between 12 and 15%, Hispanic students between 10 and 19%, and Asian students at only 6%. NCES reports these proportions.

that range between zero and five points for each essay in half-point increments. Additionally, Qualtrics tracks other data indicating grading effort. The data include (1) the participants' time on each essay in seconds, (2) whether the participant opens and looks at the grading rubric or class notes, and (3) the number of mouse clicks on each page while evaluating an essay. The treatment effect on grades is reported as the difference in points.

In addition, I collect three sets of data provided by participants in the study and from data Prolific maintains. First, I ask participants about their grading experience in the survey. This data are based on 5-level Likert scale responses indicating their level of agreement with statements related to grading fatigue, the usefulness of the grading rubric or class notes (asked in two separate questions), and the ease of learning how to grade the essays.[9] Second, I collect information about the participants' demographic, which includes their age group, gender, ethnicity/race/ethnicity, occupation, and educational level. Third, I acquire demographic information of the participants from Prolific, which I use in conjunction with the demographic information elicited by the survey to fill in any missing demographic information from the survey or the data from Prolific. The Prolific data include the specific age, ethnicity, gender, level of education, employment status, and the number of prior Prolific studies in which the participants have participated.

There are only a few missing data points of the grades and the elicited participant demographic information from the survey. There are a total of twenty-three missing grades from twenty-three different participants. Given that there should be four thousand grades in the data, this rate of missingness is 0.6%.

[9] Of the five hundred participants in the study, two to three participants opted not to answer at least one of these questions but did answer at least one of the questions.

Furthermore, if a participant chose not to grade, they only did it once. Of the participant demographic information elicited, only three people did not provide information on one of the questions. One participant did not provide their ethnicity, another did not include their occupation, and the last did not include their education level. Because there are only three missing observations, I use the data maintained by Prolific to impute the data. Additionally, instead of using the age group elicited from the experiment, I use the specific age of each participant that Prolific maintains and provides. In this data, three individuals had "DATA EXPIRED" for their specific age. I use the median age of their age brackets to impute a specific age.

2.4 Empirical Strategy and Summary Statistics

Empirical Strategy. The main empirical specification is a difference-in-means regression, with dummy variables to indicate treatment status and putative student identity observed in the treatment group when evaluating the heterogeneous effects. For the main differences between non-blind and blind grading, I regress the outcome variables on a dummy variable indicating the participant's treatment status or the students' putative demographics in the non-blind grading environment. To determine the significance of the treatment heterogeneity, I conduct hypothesis tests of the differences in the outcomes between the racial/ethnic and gender groups. Standard errors are robust and clustered at the participant level in all regressions.

Outcome Variables The main outcomes in this study are the participants' assigned grades for each graded essay, the time in seconds grading each essay, indicators that they reference the grading rubric or class notes on each essay, and their clicks per page.

The grade is a discrete variable that ranges from zero to five in half-point increments, denoted as the variable *Grade*. The participant's time spent grading each essay is *Time*, the usage of the grading rubric or class notes is *Information Seeking*, and the clicks per page are *Clicks*. *Time* is the total time in seconds each participant spent on each essay (or page). *Information Seeking* is a discrete variable that is either zero or one. A value of zero implies participants did not open the grading rubric or the class notes; a value of one means participants opened either the grading rubric or the class notes. *Clicks* is the number of clicks on the page, including the number of times participants moved the grading bar when assigning a grade to the essay.

Main specification: differences between non-blind and blind grading environments To identify the causal treatment effect of non-blind grading on the outcomes of interest, I use the following linear regression:

$$y_{pj} = \alpha \text{Treat}_p + X_{pj} + \epsilon_{pj}. \tag{2.7}$$

p indexes participant. j indexes the submission. y_{pj} is the grade (or any of the additional behavioral variables). $Treat_p$ is equal to one if the participant p is in the non-blind grading environment and zero otherwise. It can also be a vector of indicator variables for the racial/ethnic or gender identity observed by participants in the non-blind grading environment. The vector of X_{pj} are pre-randomization controls, including essay fixed effects and participant demographics. ϵ_{pj} is the error term.[10] The coefficient vector α from Equation 2.7 is the average difference in outcomes between non-blind and blind grading environments.

To identify differences in the treatment of race/ethnicity/ethnicity or gender in the non-blind grading environment, I conduct hypothesis tests on the differences in the coefficients in α. If α_4 is the coefficient on the white identity, the hypothesis test is $\alpha_r = \alpha_4$ where $r \in \{Asian, Black, Hispanic\}$

Marginal effects: Accuracy Incentives and Essay order. I also evaluate the effect of the accuracy incentives and the order in which each essay is graded on the outcome variables in the following specification:

$$y_{pj} = \alpha \text{Treat}_p + \zeta \text{Het}_p + \beta \text{Treat} \times \text{Het}_p + X_{pj} + \epsilon_{pj}. \tag{2.8}$$

This regression is similarly defined as Equation 2.7, except Het_p is a variable for the accuracy incentive or essay order. The values of Het_p are re-scaled between zero and one for the min and maximum value of the accuracy incentive and essay order, using the following equation: $\frac{x-min}{max-min}$ where x is the row value of the incentive or essay order. A value of zero means the participant is assigned an accuracy incentive of five cents, or the essay graded is the first; a value of one is a forty-five cent accuracy incentive, or the last essay graded. β is a vector of difference-in-difference coefficients. ζ is the main effect of Het_p on the unknown identity in the blind grading environment. To identify the differences in the treatment heterogeneity of the students' putative racial/ethnic or gender identities, I conduct the following hypothesis tests:

[10] Note that when evaluating the participants' responses to the questions about their grading experiences, I use $y_p = \delta \text{Treat}_p + X_p + \epsilon_p$ for each question and omitted repeating observations related to the essays j. Otherwise, the number of observations in the total data set would artificially increase the power and potentially show significant effects when there may not be any.

The differences at the lowest value of Het_p: If α_1 is the coefficient on the white identity, then the hypothesis test are $\alpha_r = \alpha_1$ where $r \in \{Asian, Black, Hispanic\}$

The differences at the largest value of Het_p: If β_4 is the coefficient on the white identity, then the hypothesis tests are $\alpha_r + \beta_r = \alpha_4 + \beta_4$ for $r \in \{Asian, Black, Hispanic\}$.

2.5 Descriptive Statistics

Balance Table. Table 2 provides averages of differences in the characteristics of the participants' demographics for the control and treatment groups after random assignment. I test the balance of each participant's demographic characteristics in the blind and non-blind groups and conduct a chi-square test of the null hypothesis that the differences in these characteristics are jointly zero. The number of observations in each regression is five hundred for the total number of participants in the study. The omitted variables or reference characteristics not included are whether the participant is a white male who has some or no higher education and whose profession is in government or management.

Panel A of Table 2 shows the average value of each demographic characteristic in the control group, the mean difference between the treatment and control groups, the standard errors in parentheses, and the p-value from regressing each demographic variable on a single indicator for treatment status. This exercise reveals that there is no difference in each of the participant's demographic characteristics between the blind and treatment groups, except for participants who are Hispanic. Panel B of Table 2 shows the results from a joint test of the hypothesis that the coefficient on each of the participants' demographics from Panel A is zero. The chi-square value is 10.7, and the p-value is 0.55, which fails to reject

Table 2 Balance Table

Demographic	Control (Blind)	Treatment (Non-Blind)	P-value
Panel A: T-test of equality across treatment and control groups for each separate variable			
	$\boldsymbol{\hat{\mu_B}}$	$\boldsymbol{x - \hat{\mu_B}}$	
Age	41.075	1.798 (1.158)	0.121
Female	0.484	0.027 (0.045)	0.545
Black	0.092	0.002 (0.026)	0.944
Asian	0.096	-0.007 (0.025)	0.787
Hispanic	0.068	-0.04 (0.019)**	0.038
Other ethnicity	0.032	-0.016 (0.014)	0.249
Master's or Higher	0.239	-0.011 (0.039)	0.784
Bachelor's Degree	0.740	0.022 (0.04)	0.578
Sales/Service/Construction	0.220	-0.001 (0.036)	0.979
Teacher	0.072	0.022 (0.029)	0.438
Unemployed/Retired/Student	0.358	-0.023 (0.043)	0.593
Prolific Experience	2041.361	-53.783 (131.933)	0.684
Panel B: Joint test			
Degrees of Freedom	12		
Chi-square	10.7		
P-value	0.55		

$^{***}p < 0.01$; $^{**}p < 0.05$; $*p < 0.1$

Note: This table shows the slope estimates from regressing each participant's demographic variable on treatment status and the joint test from regressing the indicator of treatment status on the participants' demographic variables. The regressions include essay fixed effects. All standard errors are clustered at the participant level. Panel A shows the control mean μ_B, the difference between the treatment and control means, the standard error in parentheses, and the p-value. Panel B reports the results from a joint test from regressing treatment status on the control variables. In each regression, N is equal to five hundred for the total number of participants in the study.

the hypothesis that the coefficients of the participants' demographics are jointly zero.[11]

Assignment of submissions and observed student identities The Qualtrics' survey is programmed to assign a student identity randomly to each essay. Each participant graded a set of eight essays sampled randomly from a pool

[11] The Appendix includes more balance tables showing the differences in the assigned students' identities with the participant demographics. See Appendix Table A.1

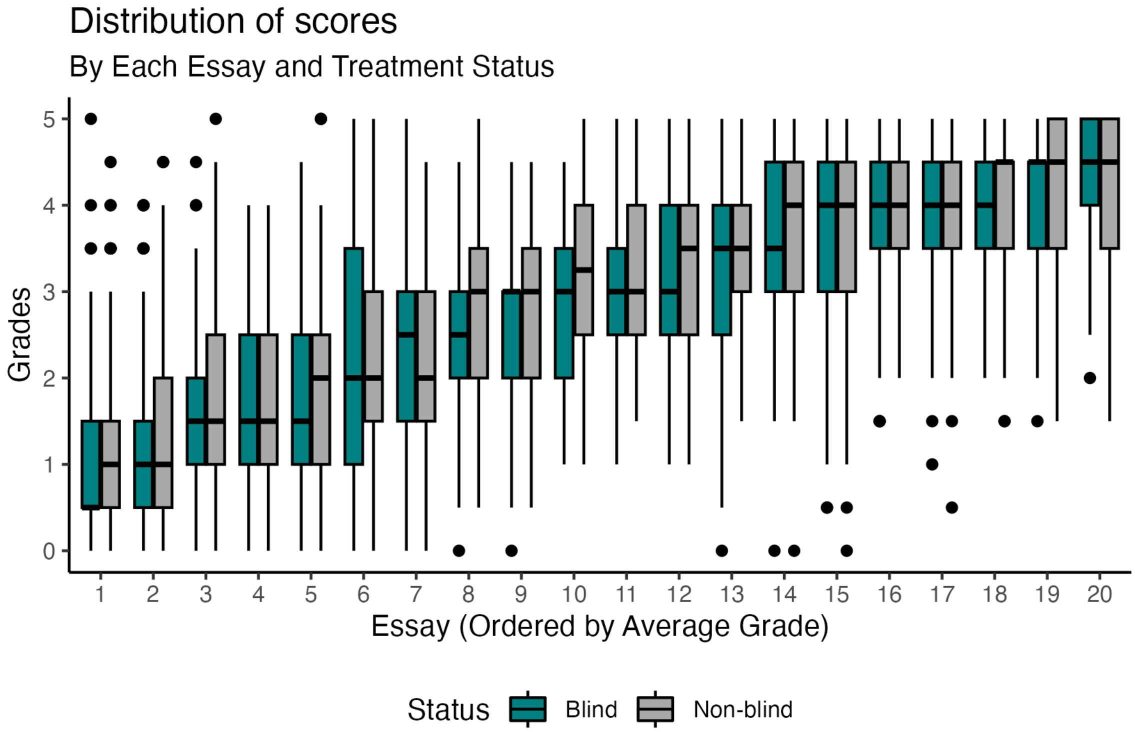
of twenty essays. Although participants in the blind grading environment were not permitted to see the assigned student identities, each essay that participants graded was nevertheless associated with a specific identity, despite these identities being moot. The gender and racial/ethnic proportions of each assigned identity revealed to the treatment group are 50% female, 12% Asian, 12% Black, 13% Hispanic, and 62% White.

Grades The data suggest that participants exercise some discretion in grading while working in both the non-blind and blind environments. Figure 4 illustrates the range of the grades in points between zero and five. The box and lines are whisker-box plots, ordered from smallest to largest average scores given to each essay. The line in the middle of the boxes is the median grade; the hinges (or the lines at the top and bottom of the boxes) refer to the 25th and 75th quartile values. The whiskers, the vertical lines spanning out of the boxes, are based on the distance between the first and third quartiles. The points are outliers. The dark green boxes are for the scores in the treatment group, and the light green boxes are for the scores in the control group. Pooling across all student identities within each group, the distribution appears to be similar between the two groups, although in some cases, the score ranges for a specific essay are larger for one group than the other.

2.6 Results

This section reports the results of the experiment in four sections. The first two sections focus on the impact of the grading environment on the grades given to the putative student identities. Section 1 reports average grade differences in the non-blind and blind grading environments. This section also reports the results from investigating the impact of the accuracy incentives and the essay order

Figure 4 Distribution of the Scores by Essay and Treatment Status



Note: This figure illustrates the distribution of grades assigned to each essay by treatment status (Blind grading vs. non-blind grading). For each essay, the left bar (teal) is the distribution of grades given in the blind grading environment. The right bar (dark gray) is the distribution of grades given in the non-blind grading environment.

effects on grades. I also evaluate how the interaction of essay order and accuracy incentives impact grades. Section 2 reports differences from other grading behaviors associated with effort (i.e., time spent grading, usage of a grading rubric or class notes, and clicks per page). Section 3 includes results from the end-of-experiment survey about the participants' experience.

Impact on Grades.

Are average grades different in a non-blind or blind grading environment?. Table 3 summarizes the average grade differences by treatment status (Panel A) and the treatment heterogeneity by the students' putative racial/ethnic (Panel B) and gender identities (Panel C) in the non-blind grading environment. The estimated coefficients in Columns 1 through 5 are based on the controls used in the regressions. Column 5 is our preferred (and pre-registered) specification.[12]. I use essay-fixed effects to control for unobserved heterogeneity of the essays' quality. The omitted variable is the unknown identity in the blind grading environment. The estimated coefficients are the differences in the average grades given to all of the students in the non-blind grading environment compared to the students in the blind-grading environment.

The results in row 1 show that depending on the controls in the regressions, the difference in grades ranges between 0.067 and 0.116 points. Given that the average score for the unknown identity in the blind-grading environment is 2.87 points, the maximum grade differences are between 2 and 4%, only statistically significant at the ten % level in column 2 and statistically insignificant in column 5. The results shown in this row suggest that, on average, there are no differences in grades between the non-blind or blind grading environments.

Panel B and Panel C report the treatment heterogeneity of the results by the students' putative race/ethnicity and gender observed in the non-blind grading environment. Similar to the results in panel A, the estimated coefficients in column 5 are not statistically significant, suggesting that participants did not discriminate in how they assigned grades in the non-blind grading environment relative to the blind grading environment. Furthermore, hypothesis tests of the differences in the grades between the racial/ethnic and gender identities are not significant, as shown in Table 4.

Column 1 of Table 4 reports the demographic identity compared to the white identity or male identity. Column 2 is the hypothesis test, and the Greek

[12] We also pre-registered column 4. For simplicity, I use column 5's specification for all hypothesis tests.

letters refer to the Greek letters listed in Table 3 as a reference. Columns 3 through 6 report the estimated differences, the standard errors, the t-statistics, and the p-values. The differences in grades between the students' racial/ethnic and gender identities in the treatment group have p-values greater than 0.05.

Do accuracy incentives reduce discrimination in grading?. Table 5 and Table 6 report the difference-in-difference estimates of interacting 'Incentive,' a variable representing the value of the minimum and maximum accuracy incentive, with the students' putative race/ethnicity or gender in the non-blind grading environment.[13]

Rows 1 through 4 in Table 5 and rows 1 through 2 in Table 6 report the estimated average differences in grades given to students' putative identities in the non-blind grading environment relative to the unknown identity in the blind grading environment when participants are assigned an accuracy incentive of five cents. Column 5 reports that all student's racial/ethnic identities except for Black students receive average scores between 0.271-0.459 points higher than the unknown identity in Table 5, statistically significant between the 0.1-10% levels. The effect on the Black identity is 0.114 points larger than the unknown identity but statistically insignificant. The coefficients estimated on the female and male identity are 0.281 and 0.310 points higher and significant at the 1% level. The main effect of 'Incentive' on the unknown identity is statistically insignificant, suggesting that as the accuracy incentive increases, grades remain unchanged in the blind grading environment.

[13] All of the demographic variables technically interact with 'Non-Blind.' However, I removed 'Non-Blind' only for visualization purposes. Visualization of the estimated coefficients from Table 5 can be found in Figure D.1.

Table 3 Main Results: Differences in Grades in the Non-Blind Grading Environment

	(1)	(2)	(3)	(4)	**(5)**
Panel A: Average Treatment Effect					
Non-Blind	0.116^{*}	$0.087^{\cdot}$	0.106^{*}	0.075	0.067
	(0.052)	(0.051)	(0.051)	(0.050)	(0.050)
Panel B: Treatment Heterogeneity by the Students' Putative Race/Ethnicity					
	Relative to the Unknown Identity				
Asian×Non-Blind (α_1)	0.105	$0.132^{\cdot}$	0.097	0.123	0.119
	(0.102)	(0.077)	(0.100)	(0.076)	(0.074)
Black×Non-Blind (α_2)	0.070	0.013	0.059	0.003	0.017
	(0.102)	(0.080)	(0.103)	(0.080)	(0.079)
Hispanic×Non-Blind (α_3)	−0.043	0.059	−0.055	0.044	0.039
	(0.090)	(0.075)	(0.090)	(0.074)	(0.075)
White×Non-Blind (α_4)	0.161^{**}	0.098	0.151^{**}	0.087	0.072
	(0.055)	(0.053)	(0.054)	(0.053)	(0.053)
Panel C: Treatment Heterogeneity by the Students' Putative Gender					
	Relative to the Unknown Identity				
Female×Non-Blind (γ_1)	0.097	0.063	0.086	0.052	0.049
	(0.060)	(0.054)	(0.059)	(0.053)	(0.053)
Male×Non-Blind (γ_2)	0.136^{*}	0.110^{*}	0.126^{*}	0.099	0.084
	(0.061)	(0.055)	(0.060)	(0.054)	(0.055)
Benchmark Score	N	Y	N	Y	N
Participant Control	N	N	Y	Y	Y
Essay FE	N	N	N	N	Y
Mean Unknown Identity	2.83				

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: This table shows the estimates of the differences in assigned grades by grading environment (panel A), the observed putative race of the student in the non-blind grading environment (panel B) and the observed putative gender of the student in the non-blind grading environment (panel C). 'Non-Blind' is an indicator variable if the participants are in the non-blind grading environment or in the blind grading environment. Our registered and preferred specification is denoted by column (5). Standard errors are robust and clustered at the participant level in all columns. Hypothesis tests of the estimated coefficients in column 5 are reported in Table 4

Table 4 Hypothesis Tests of Differences in Average Grades in the Non-Blind Grading Environment

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
		Panel A: Racial/Ethnic Gaps			
		Relative to white			
Asian	$\alpha_1 = \alpha_4$	0.047	0.065	0.713	0.476
Black	$\alpha_2 = \alpha_4$	-0.055	0.067	-0.830	0.407
Hispanic	$\alpha_3 = \alpha_4$	-0.033	0.068	-0.490	0.624
		Panel A: Gender Gap			
		Relative to male			
Female	$\gamma_1 = \gamma_2$	-0.035	0.044	-0.804	0.421

Note: This table reports the hypothesis tests of the estimated coefficients in column 5 from Table 3. Column 1 reports the demographic identity compared to the omitted variable, which is either the white identity or the male identity. Column 2 is the hypothesis tested, using the coefficients listed in Table 3 as a reference. Columns 3 through 6 report the estimated differences, the standard errors, the t-statistics, and the p-values.

Rows 6 through 9 in Table 5 and rows 4 through 5 in Table 6 are the average grade differences assigned to the students' putative identities relative to the unknown identity when participants could earn forty-five cents for grading accurately. In column 5, all of the estimates are negative and statistically significant except for Black students. Asian students, in particular, receive the largest decrease in grades as incentives increase from five to forty-five cents (a decrease of 0.664 points). The estimate on Black students is likewise negative but statistically no different than the effect on the unknown identity in the blind grading environment, suggesting increasing accuracy incentives did not impact how participants assessed Black students, only how participants assessed all of the other demographic groups.

What are the implications of these results? At lower incentives, graders may be more generous in how they assign grades to students when they can see signals of the students' demographics than the unknown identity. However, Black students do not seem to benefit from this generosity. Additionally, the effect of increasing the value of the accuracy incentives on the racial/ethnic and gender identities in the non-blind grading environment suggests that graders may become increasingly more critical in how they assign grades as they are paid more for their accuracy, except participants do not change how they assess Black students.

To further evaluate these results, I conduct hypothesis tests of the differences in the estimated treatment effects on the students' putative identities in Table 7. Panel A reports the hypothesis tests for the racial/ethnic identities, and Panel B reports the tests for the gender identities. The hypothesis tests in Panel A compare how average grades are predicted to differ between the students' putative racial/ethnic identities if participants receive the lowest accuracy incentive (5 cents). The only significant difference in average grades is between Black and white identities (p-value=0.055). Black students receive the lowest grade by 0.21 points, a difference of 4.2 percentage points (0.21/5). At the maximum accuracy incentive (45 cents), there are no differences in average grades between racial/ethnic or gender demographic groups. Panel B reports that there are no differences in grades between female and male students at any accuracy incentive amount.

Does essay order matter?. Grading experience may change how participants assign grades. Do grade gaps decrease as graders become more familiar with the material and how to assign grades? I explore this question by evaluating how grade gaps change from grading the first to the last essay or the order effects. Tables 8 and 9 report the estimated coefficients from estimating the order effects of

Table 5 Impact of Accuracy Incentives on Grades by the Students' Putative Race/Ethnicity

	Dependent variable:				
	Score				
	(1)	(2)	(3)	(4)	(5)
	Relative to the Unknown Identity				
Asian (α_1)	0.478^{**}	0.451^{***}	0.515^{**}	0.482^{***}	0.459^{***}
	(0.159)	(0.131)	(0.158)	(0.130)	(0.124)
Black(α_3)	0.105	0.060	0.130	0.079	0.095
	(0.169)	(0.130)	(0.172)	(0.132)	(0.131)
Hispanic (α_3)	0.122	0.223	0.162	$0.251^{\cdot}$	$0.271^{\cdot}$
	(0.162)	(0.139)	(0.161)	(0.136)	(0.139)
White (α_4)	0.346^{***}	0.307^{**}	0.380^{***}	0.334^{***}	0.305^{**}
	(0.099)	(0.100)	(0.098)	(0.098)	(0.098)
	Main effect on the Unknown Identity				
Incentive	0.083	0.132	0.112	0.161	0.154
	(0.099)	(0.101)	(0.099)	(0.100)	(0.099)
	Marginal effects relative to the unknown identity				
Asian × Incentive (β_1)	-0.731^{**}	-0.621^{**}	-0.819^{**}	-0.701^{***}	-0.664^{***}
	(0.271)	(0.207)	(0.265)	(0.205)	(0.198)
Black × Incentive (β_2)	−0.062	−0.084	−0.135	−0.142	−0.147
	(0.300)	(0.211)	(0.303)	(0.211)	(0.205)
Hispanic × Incentive (β_3)	−0.320	−0.315	−0.424	$-0.399^{\cdot}$	-0.448^{*}
	(0.261)	(0.227)	(0.261)	(0.222)	(0.227)
White × Incentive (β_4)	-0.360^{*}	-0.403^{**}	-0.448^{**}	-0.477^{**}	-0.450^{**}
	(0.153)	(0.150)	(0.149)	(0.146)	(0.146)
Observations	3,977	3,977	3,977	3,977	3,977
R^2	0.007	0.476	0.015	0.483	0.513
Benchmark Score	N	Y	N	Y	N
Participant Control	N	N	Y	Y	Y
Essay FE	N	N	N	N	Y
Mean of Unknown Identity	2.75				

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: This table reports the marginal effects of the accuracy incentives on standardized grades in the non-blind grading environment relative to the blind grading environment. 'White', 'Asian', 'Black', 'Hispanic' are indicator variables for whether the observed identity is the putative race/ethnicity or the unknown identity in the blind grading environment. The omitted variable is the unknown identity in the blind grading environment. 'Incentive' is a value between zero and one, using the following equation $\frac{AssignedIncentive-0.05}{0.45-0.05}$. Hypothesis tests of the estimated coefficients in column 5 reported in Table 7

Table 6 Impact of Accuracy Incentives on Grades by the Students' Putative Gender

	Dependent variable:				
	Score				
	(1)	(2)	(3)	(4)	(5)
	Relative to the Unknown Identity				
Female (γ_1)	0.298^{**}	0.270^{**}	0.332^{**}	0.296^{**}	0.281^{**}
	(0.105)	(0.098)	(0.104)	(0.097)	(0.097)
Male(γ_2)	0.315^{**}	0.302^{**}	0.349^{***}	0.328^{**}	0.310^{**}
	(0.106)	(0.103)	(0.105)	(0.101)	(0.101)
	Main effect on the Unknown Identity				
Incentive	0.083	0.132	0.112	0.161	0.154
	(0.099)	(0.101)	(0.099)	(0.100)	(0.099)
	Marginal effects relative to the Unknown Identity				
Female × Incentive (δ_1)	-0.391^{*}	-0.398^{**}	-0.479^{**}	-0.472^{**}	-0.450^{**}
	(0.163)	(0.149)	(0.160)	(0.146)	(0.147)
Male× Incentive (δ_2)	-0.347^{*}	-0.369^{*}	-0.435^{**}	-0.443^{**}	-0.436^{**}
	(0.166)	(0.154)	(0.164)	(0.151)	(0.150)
Observations	3977	3977	3977	3977	3977
R^2	0.00	0.48	0.01	0.48	0.51
Benchmark Score	N	Y	N	Y	N
Participant Control	N	N	Y	Y	Y
Essay FE	N	N	N	N	Y
Mean of Unknown Identity	2.75				

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: This table reports the marginal effects of the accuracy incentives on standardized grades in the non-blind grading environment relative to the blind grading environment. 'Male' and 'Female' are indicator variables for whether the observed identity is male or unknown; and female or unknown. The omitted variable is the unknown identity in the blind grading environment. 'Incentive' is a value between zero and one, using the following equation $\frac{AssignedIncentive - 0.05}{0.45 - 0.05}$. Hypothesis tests of the estimated coefficients from column 5 reported in Table 7

Table 7 Hypothesis Tests: Impact of Accuracy Incentives on Grades in the Non-Blind Grading Environment

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
	Panel A: Students' Racial/Ethnic Gaps				
	Relative to white students: differences at 5 cents				
Asian	$\alpha_1 = \alpha_4$	0.154	0.111	1.386	0.166
Black	$\alpha_2 = \alpha_4$	-0.210	0.109	-1.916	0.055
Hispanic	$\alpha_3 = \alpha_4$	-0.034	0.123	-0.277	0.781
	Relative to white students: differences at 45 cents				
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	-0.060	0.103	-0.581	0.561
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	0.094	0.107	0.877	0.380
Hispanic	$\alpha_3 + \beta_3 = \beta_3 + \beta_4$	-0.032	0.118	-0.270	0.788
	Panel B: Students' Gender Gaps				
	Relative to male students: differences at 5 cents				
Female	$\gamma_1 = \gamma_2$	-0.029	0.075	-0.388	0.698
	Relative to male students: differences at 45 cents				
Female	$\gamma_1 + \delta_1 = \gamma_2 + \delta_2$	-0.042	0.073	-0.577	0.564

Note: This table reports the hypothesis tests of the estimated coefficients in column 5 from Table 5 for the racial/ethnic effects and Table 6 for the gender effects. Bootstrap standard errors of $\alpha_2 = \alpha_4$ are shown in Figure D.3

grading homework submissions sequentially. The estimates shown for the students' putative identities in the non-blind grading environment are the differences in the average effects relative to the unknown identity in the blind grading environment. 'Order' is the order in which participants graded the essays, re-scaled so that values range from zero to one, where zero is the first essay and one is the eighth essay.

Regardless of the specification, there are no essay order effects on grades in the non-blind grading environment compared to the blind grading environment. Nor does order impact the unknown identity in the blind grading environment.

Furthermore, there are no statistically significant differences in grades between the students' demographic groups. Table 10 shows the results of the hypothesis tests.

Table 9 Impact of Essay Order on Grades by the Students' Putative Gender

	Dependent variable:				
	Score				
	(1)	(2)	(3)	(4)	(5)
Relative to the Unknown Identity					
Female (γ_1)	0.090	0.078	0.075	0.063	0.070
	(0.093)	(0.082)	(0.092)	(0.081)	(0.080)
Male (γ_2)	0.099	0.109	0.091	0.100	0.093
	(0.094)	(0.082)	(0.094)	(0.082)	(0.080)
Main effect on the Unknown Identity					
Order	−0.011	−0.030	−0.012	−0.030	−0.029
	(0.093)	(0.071)	(0.093)	(0.071)	(0.069)
Marginal effects relative to the Unknown Identity					
Female × Order (δ_1)	0.013	−0.028	0.023	−0.022	−0.041
	(0.155)	(0.123)	(0.155)	(0.122)	(0.120)
Male × Order (δ_2)	0.075	0.002	0.069	−0.003	−0.018
	(0.158)	(0.121)	(0.159)	(0.122)	(0.116)
Num. obs.	3977	3977	3977	3977	3977
R^2 (full model)	0.00	0.47	0.01	0.48	0.51
Benchmark Score	N	Y	N	Y	N
Participant Control	N	N	Y	Y	Y
Essay FE	N	N	N	N	Y
Mean of Unknown Identity	2.77				

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: This table reports the estimated marginal effects of essay order on grades. 'Order' is a value between zero and one, using the following equation $\frac{Order-1}{8-1}$, where the value of zero is the first essay and a value of one is the last essay graded by each participant. Hypothesis tests of the coefficients in column 5 are reported in Table 10.

Do accuracy incentives impact the order effects?. The results up to this point indicate that accuracy incentives have some effect on how participants assign grades and may reduce discrimination in grading. However, the lack of change in average grades across essay order is puzzling.

Table 8 Impact of Essay Order on Grades by the Students' Putative Race/Ethnicity

	Dependent variable:				
	Score				
	(1)	(2)	(3)	(4)	(5)
Relative to the Unknown Identity					
Asian (α_1)	-0.010	0.083	-0.007	0.088	0.050
	(0.175)	(0.137)	(0.175)	(0.139)	(0.131)
Black (α_2)	0.082	-0.008	0.084	-0.007	0.003
	(0.163)	(0.126)	(0.162)	(0.125)	(0.120)
Hispanic (α_3)	-0.135	0.019	-0.149	-0.001	0.024
	(0.155)	(0.124)	(0.155)	(0.125)	(0.125)
White (α_4)	$0.168^{\cdot}$	$0.133^{\cdot}$	$0.152^{\cdot}$	0.117	0.116
	(0.089)	(0.080)	(0.088)	(0.079)	(0.078)
Main effect on the Unknown Identity					
Order	-0.011	-0.030	-0.012	-0.030	-0.029
	(0.093)	(0.071)	(0.093)	(0.071)	(0.069)
Marginal effects relative to the unknown identity					
Asian × Order (β_1)	0.238	0.101	0.217	0.071	0.140
	(0.297)	(0.216)	(0.298)	(0.219)	(0.203)
Black × Order (β_2)	-0.025	0.042	-0.052	0.019	0.027
	(0.278)	(0.198)	(0.276)	(0.197)	(0.186)
Hispanic × Order (β_3)	0.183	0.081	0.187	0.089	0.029
	(0.264)	(0.193)	(0.263)	(0.194)	(0.187)
White × Order (β_4)	-0.014	-0.067	-0.003	-0.059	-0.085
	(0.145)	(0.114)	(0.146)	(0.115)	(0.111)
Num. obs.	3977	3977	3977	3977	3977
R^2 (full model)	0.00	0.47	0.01	0.48	0.51
Benchmark Score	N	Y	N	Y	N
Participant Control	N	N	Y	Y	Y
Essay FE	N	N	N	N	Y
Mean of Unknown Identity	2.77				

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: This table reports the estimated marginal effects of essay order on grades. 'Order' is a value between zero and one, using the following equation $\frac{Order-1}{8-1}$, where the value of zero is the first essay and a value of one is the last essay graded by each participant. Hypothesis tests of the coefficients in column 5 are reported in Table 10.

Table 10 Hypothesis Tests: Impact of Essay Order on Grades in the Non-Blind Grading Environment

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
Panel A: Hypothesis tests from Table 8					
Relative to white students: differences on the first essay graded					
Asian	$\alpha_1 = \alpha_4$	-0.065	0.126	-0.520	0.603
Black	$\alpha_2 = \alpha_4$	-0.113	0.123	-0.912	0.362
Hispanic	$\alpha_3 = \alpha_4$	-0.091	0.125	-0.728	0.466
Relative to white students: differences on the last essay graded					
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	0.160	0.120	1.334	0.182
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	-0.001	0.125	-0.005	0.996
Hispanic	$\alpha_3 + \beta_3 = \beta_3 + \beta_4$	0.022	0.118	0.191	0.849
Panel B: Hypothesis tests from Table 9					
Relative to male students: differences of the intercept terms					
Female	$\gamma_1 = \gamma_2$	-0.023	0.084	-0.276	0.782
Relative to male students: differences of the marginal effect of the essay order					
Female	$\gamma_1 + \delta_1 = \gamma_2 + \delta_2$	-0.046	0.078	-0.594	0.553

Note: This table reports the hypothesis tests of the estimated coefficients in column 5 from Table 8 for the racial effects and Table 9 for the gender effects.

The experiment is designed so that there are no differences in the quality of the student's work, regardless of the students' putative race/ethnicity or gender. Participants may have some initial beliefs about the different demographics coming into the experiment, but if they statistically discriminate, they should change how they assign grades as they learn the true distribution. This updating, in theory, should be expressed by a decrease in grade gaps as they grade each subsequent essay. However, the results in Tables 8 and 9 indicate that the participants are not learning the underlying distribution of the quality of the essays. Instead, the evidence seems to indicate taste-based discrimination as the source of the differences in grades in this study, especially for the Black-white gap. That is, paying people to grade Black students more favorably reduces discrimination.

To investigate further whether the grading behavior follows a statistical discrimination or a taste-based discrimination model, I will investigate if the value of the assigned accuracy incentives could mask the order effects. Given my hypothesis that larger accuracy incentives increase attention towards learning the quality of the student's work in a statistical discrimination framework, there should be less discrimination in grading at the beginning and the end of the grading sequence for those assigned larger incentives. However, discrimination in grading should be more pronounced at the beginning than at the end of the grading sequence for those assigned a lower incentive. If this is the case, then there are two conclusions - (1) graders exhibit behavior associated with rational inattentiveness, and (2) discrimination in grading is likely due to statistical discrimination.

I first evaluate the average grade differences in four groups: the first four and last four essays and small and large accuracy incentives. Figure 5 shows the average grades de-meaned by essay for each putative identity on the first and last four essays at low incentive and high incentive amounts. A small incentive is an accuracy incentive that is less than or equal to the median incentive value (25 cents). A large incentive is an accuracy incentive greater than or equal to the median incentive value. The point estimates and the confidence intervals are based on the average de-meaned score for each essay and the standard deviation of those de-meaned values.

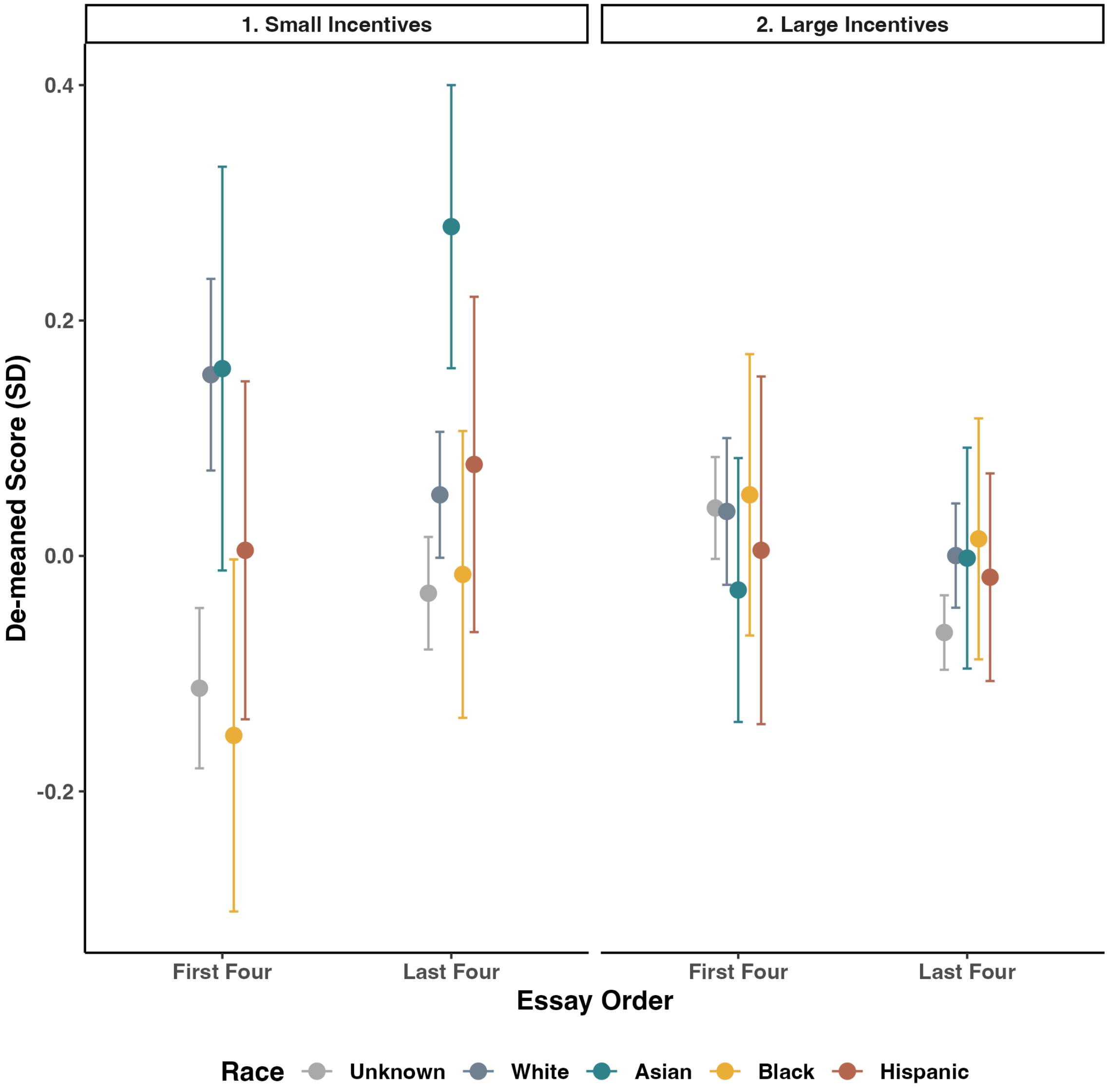
This figure shows that if participants are assigned a low accuracy incentive, grade gaps between the racial/ethnic identities are more pronounced in the first four submissions. However, those grade gaps disappear in the last four submissions, except for the Asian identity. For participants assigned a larger accuracy incentive, there are no significantly obvious grade gaps on the first or last four submissions.[14]. To formally investigate if order effects matter at different accuracy incentive values, I evaluate a triple difference of the interaction between the essay order, the accuracy incentive, and the putative racial/ethnic identities of the students in the non-blind grading environment.

Table 11 reports the estimates of the triple difference and Table 12 reports the hypothesis tests of the differences between the racial/ethnic identities. The coefficients are split into four sections, denoted by the incentive amount and the first or last essay graded. 'Incentive' and 'Order' are variables that range between zero and one (the same variables used in the previous evaluations). These variables are re-scaled so that zero is equal to five cents on the first essay and one is equal to forty-five cents and the eighth essay.

The results indicate that graders may behave more in a statistical discrimination framework than a taste-based one, and that they may be choosing what to pay attention to while they grade. At five cents, both white and Asian students received better grades on the first essay relative to the unknown identity. White and Asian students earned almost half a point larger than the unknown identity, significant at the .1 % level for white students and 10% for Asian students. There is no effect on Black or Hispanic students relative to the unknown identity. The Black-white gap is also large on the first essay. The gap is -0.525 points (p-value=0.012), about a 10 pp difference, and significant at the 1% level. However, by the last essay, there are no differences in grades between Black and white students (0.114 points, p-value=0.558). For Asian students, they continue to receive

[14] I conduct regressions to evaluate the order effects on the first and last four essays and incentive amounts in Table C.1 and Table C.2

Figure 5 Accuracy Incentives and Order Matters in Reducing Grading Discrimination



This figure shows the differences in grades on the first and last four essays for each putative racial/ethnic identity. The grades are de-meaned by essay. Panel 1 and Panel 2 show the average scores and confidence interval that participants who are assigned an incentive less than or equal to 25 cents ('Small Incentives') or greater than or equal to 25 cents ('Large Incentives').

better grades than all other racial/ethnic groups. Relative to white students, they earn a grade larger by 0.374 points (p-value=0.071), significant at the 10% level. Yet, when participants could receive 45 cents for grading accurately, the pattern changes. On the first and the last essay, there are no statistically significant grade gaps.

Assuming participants are rationally inattentive, the hypothesis tests in Table 12 indicate participants rely on and are slow to update their prior beliefs when grading. However, when graders are compensated for their work, they quickly learn how to distribute the quality of the homework submissions. Even if their prior beliefs are biased, i.e., they initially believe there are differences in the quality of work from students of different demographics, they learn the true distribution earlier in the grading session than if they had been paid lower incentives.

Other Behaviors Related to Grading Effort. In this section, I investigate other grading behaviors relating to grading effort: time spent grading, usage of the grading rubric or class notes, and clicks per page. In general, I find no differences between the non-blind and blind grading environments, nor any treatment heterogeneity.

Differences between non-blind and blind grading environments. Table 13 reports the average differences in the behavioral indicators of effort between the non-blind and blind grading environments. Table 14 reports the hypothesis tests of the treatment heterogeneity between the students' putative racial/ethnic and gender identities in the non-blind grading environment.
The dependent variable for time and clicks are logarithms due to the skewed distributions. The estimates of information seeking are based on a PLM OLS regression, so they are in percentage point differences. In both tables, I find no statistically significant differences in the coefficients.

Table 11 Triple Difference of the Impact of Accuracy Incentives and Essay Order on Grades by the Students' Putative Race/Ethnicity

5 cents & First Essay		**5 cents & Last Essay**	
		Main Effect on Unknown Identity	
		Order (β_0)	0.062
			(0.130)
Relative to Unknown Identity		*Relative to Unknown Identity*	
Asian (α_1)	$0.411^{\cdot}$	Asian × Order (β_1)	0.096
	(0.244)		(0.378)
Black (α_2)	-0.042	Black × Order (β_2)	0.289
	(0.201)		(0.289)
Hispanic (α_3)	0.264	Hispanic × Order (β_3)	0.014
	(0.213)		(0.339)
White (α_4)	0.483^{***}	White × Order (β_4)	$-0.350^{\cdot}$
	(0.139)		(0.188)
45 cents & First Essay		**45 cents & Last Essay**	
Main Effect on Unknown Identity		*Main Effect on Unknown Identity*	
Incentive (δ_0)	$0.239^{\cdot}$	Order × Incentive (ζ_0)	-0.169
	(0.136)		(0.197)
Relative to Unknown Identity		*Relative to Unknown Identity*	
Asian × Incentive (δ_1)	$-0.680^{\cdot}$	Asian × Order × Incentive (ζ_1)	0.034
	(0.355)		(0.559)
Black × Incentive (δ_2)	0.101	Black × Order × Incentive (ζ_2)	-0.527
	(0.330)		(0.495)
Hispanic × Incentive (δ_3)	-0.476	Hispanic × Order × Incentive (ζ_3)	0.054
	(0.365)		(0.563)
White × Incentive (δ_4)	-0.710^{***}	White × Order × Incentive (ζ_4)	$0.512^{\cdot}$
	(0.209)		(0.302)
Num. obs.	3977		
R^2 (full model)	0.514		
Benchmark Score	N		
Participant Control	Y		
Essay FE	Y		
Mean of Omit. Var	2.72		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: This table reports estimates from a triple difference between the students' racial putative identities, the accuracy incentives and the essay order. All coefficient estimates are from one regression. For visibility and readability, I split the difference-in-difference effects into four panels based on the lowest and highest value of the accuracy incentives (5 and 45 cents) and the first and last essay graded. Hypothesis tests of the treatment heterogeneity are conducted based on the estimated coefficients in Table 12. Further regressions evaluating the effect of the accuracy incentives and essay order can be found in Table C.1

Table 12 Hypothesis Tests of Accuracy Incentives and Essay Order on the Differences in Grades in the Non-Blind Grading Environment

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
	Differences relative to white students				
	Panel A: Impact of Small Incentives				
	First Essay				
Asian	$\alpha_1 = \alpha_4$	-0.072	0.227	-0.319	0.750
Black	$\alpha_2 = \alpha_4$	-0.525	0.209	-2.513	0.012
Hispanic	$\alpha_3 = \alpha_4$	-0.219	0.209	-1.048	0.295
	Last Essay				
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	0.374	0.207	1.805	0.071
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	0.114	0.195	0.586	0.558
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	0.145	0.230	0.631	0.528
	Panel B: Impact of Large Incentives				
	First Essay				
Asian	$\alpha_1 + \delta_1 = \alpha_4 + \delta_4$	-0.042	0.191	-0.220	0.826
Black	$\alpha_2 + \delta_2 = \alpha_4 + \delta_4$	0.286	0.199	1.434	0.152
Hispanic	$\alpha_3 + \delta_3 = \alpha_4 + \delta_4$	0.015	0.235	0.064	0.949
	Last Essay				
Asian	$\alpha_1 + \beta_1 + \delta_1 + \zeta_1 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	-0.074	0.197	-0.374	0.709
Black	$\alpha_2 + \beta_2 + \delta_2 + \zeta_2 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	-0.113	0.201	-0.565	0.572
Hispanic	$\alpha_3 + \beta_3 + \delta_3 + \zeta_3 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	-0.078	0.203	-0.387	0.699

Note: This table reports the hypothesis tests from Table 11. The hypothesis tests are the differences in the treatment effects between the students' racial identities relative to white students. Panel A shows the results for those assigned 5 cents. Panel B shows the results for those assigned 45 cents. The coefficients listed in the second column are the coefficients listed in Table 11. Bootstrap standard errors of $\alpha_2 = \alpha_4$ are shown in Figure D.4.

Table 13 Main Results: Differences in Effort

	Time (sec)	Information Seeking	Clicks
Panel A: Average Treatment Effect			
Non-Blind	0.05	-0.01	0.65
	(0.05)	(0.03)	(0.51)
Panel B: Treatment Heterogeneity by the Students' Racial Identities			
		Relative to the Unknown identity	
Asian (α_1)	0.00	0.00	1.27
	(0.07)	(0.04)	(0.83)
Black (α_2)	0.13*	0.00	0.54
	(0.06)	(0.04)	(0.55)
Hispanic (α_3)	0.01	-0.05	0.07
	(0.07)	(0.07)	(0.48)
White (α_4)	0.04	-0.01	0.68
	(0.05)	(0.03)	(0.56)

Table 14 Hypothesis Tests of Differences in Effort

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
		Panel A: Racial effects			
	Relative to White: differences in ***Time***				
Asian	$\alpha_1 = \alpha_4$	-0.04	0.05	-0.76	0.45
Black	$\alpha_2 = \alpha_4$	0.09	0.05	1.58	0.11
Hispanic	$\alpha_3 = \alpha_4$	-0.03	0.06	-0.51	0.61
	Relative to White: differences in ***Information Seeking***				
Asian	$\alpha_1 = \alpha_4$	0.01	0.04	0.38	0.71
Black	$\alpha_2 = \alpha_4$	0.02	0.03	0.44	0.66
Hispanic	$\alpha_3 = \alpha_4$	-0.04	0.03	-1.09	0.28
	Relative to White: differences in ***Clicks per Page***				
Asian	$\alpha_1 = \alpha_4$	0.58	0.74	0.79	0.43
Black	$\alpha_2 = \alpha_4$	-0.14	0.61	-0.24	0.81
Hispanic	$\alpha_3 = \alpha_4$	-0.61	0.46	-1.35	0.18
		Panel A: Gender effects			
	Relative to Male: differences in ***Time***				
Female	$\gamma_1 = \gamma_2$	0.00	0.04	0.10	0.92
	Relative to Male: differences in ***Information Seeking***				
Female	$\gamma_1 = \gamma_2$	-0.02	0.02	-0.86	0.39
	Relative to Male: differences in ***Clicks per Page***				
Female	$\gamma_1 = \gamma_2$	-0.01	0.04	-0.36	0.72

Note: This table reports the hypothesis tests from Table 13

The appendix goes into further details of the differences in the behavioral variables across accuracy incentives and essay order. For the impact of the accuracy incentives on these outcomes, please see Tables B.2 and B.1. Hypothesis tests are found in Table B.3. In terms of time grading, participants in both the non-blind and blind grading environments spent the same amount of time grading each essay. However, they all spent significantly more time on the first essay than the last. This result is shown in Tables B.4 and B.5, with the hypothesis tests shown in

Table B.6. The interaction between the accuracy incentives and essay order can be seen in Table B.7, with the hypothesis tests shown in Tables B.8, B.9, B.10.

End-of-survey Questions. At the end of the survey, I ask participants to rate their agreement with four questions related to their grading experience using 5-level Likertscale responses indicating their level of agreement.

The questions are: (1) whether they found grading exhausting, the usefulness of the (2) grading rubric and (3) class notes, and (4) whether they felt like grading became easier throughout the experiment. The question on grading exhaustion measures differences in learning how to grade and exhaustion from the participants' viewpoint. The questions on the grading rubric's usefulness and class notes measure differences in how participants feel about the provided information. Although it cannot measure whether they feel they reference the grading rubric or class notes more often, it suggests differences in how participants feel about the material, indicating that they refer to the rubric and notes more often. The last question on the ease of grading over the study captures any differences in how participants felt about their ability to allocate points, either from learning how to grade each essay or from fatigue.

Most of the participants answer these questions. Of the 500 participants in this study, there are between 497 and 499 answers for each question. Because there are so few missing observations, I do not impute missingness or worry about sample selection bias in the estimates. To calculate the differences in grading experiences between the blind and non-blind, I regress the participants' answers to these questions on a treatment indicator and included control variables for the participants' demographics. The results are all statistically insignificant.

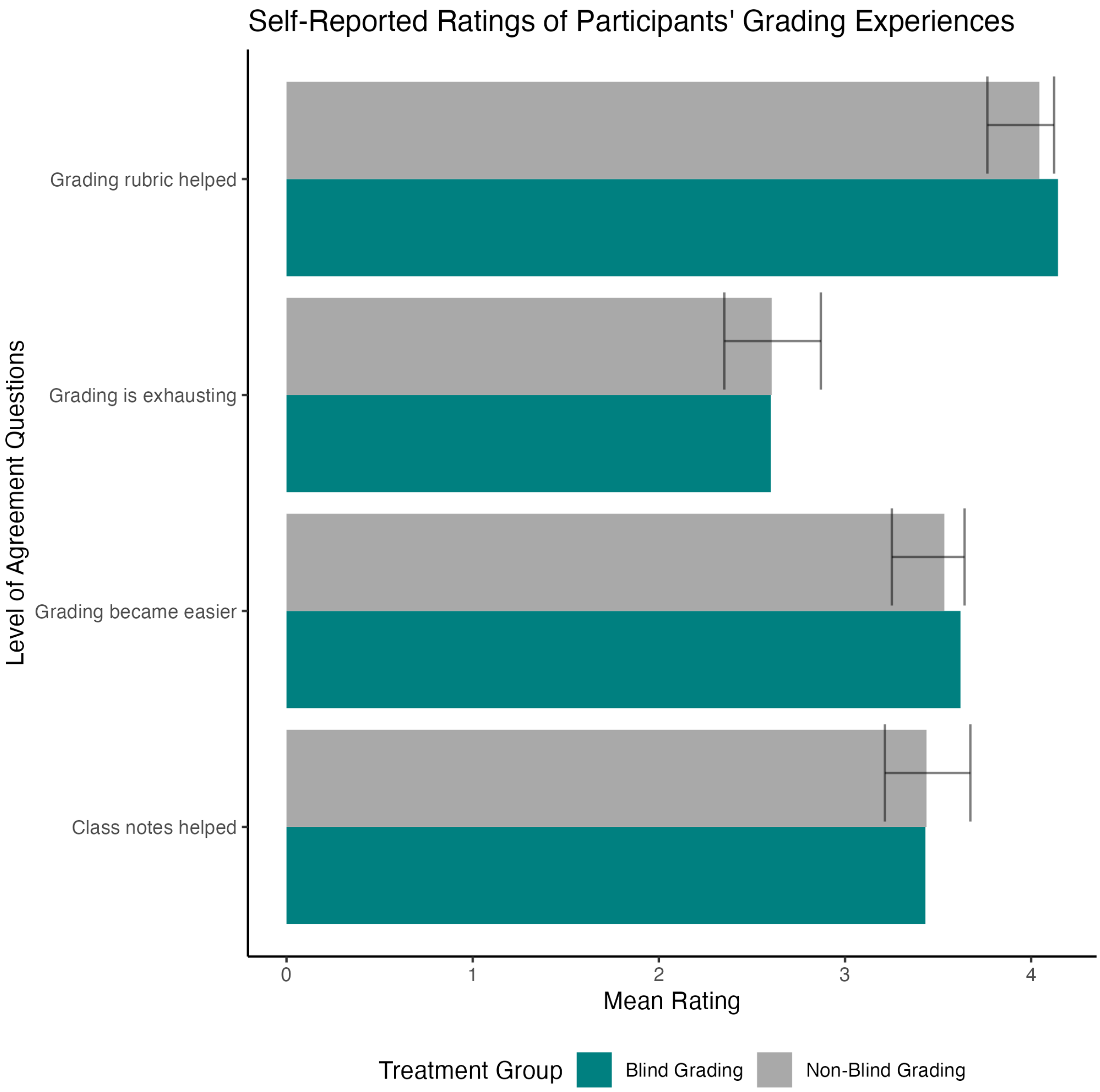
Figure 6 illustrates the point estimates and the confidence intervals. The gray bar is the estimated treatment mean with the confidence interval shown on it. The teal bar is the control mean. Analysis of the answers to these questions suggests that despite some differences in scores across the treatment (non-blind grading) and control groups (blind grading), I cannot reject the hypothesis that there are no differences in participants' attitudes about their grading experiences in the study.

2.7 Conclusion

The results from this experiment provide insights into the mechanisms driving bias in grading in a non-blind grading environment and the potential for blind grading to reduce biases in grades.

I find evidence that graders may rely more on their beliefs about the quality of work from students of different demographic groups while grading subjective homework, like an essay. Even though graders may rely less on students' observable demographics as they assess the material, the magnitude of the bias is sensitive to whether graders are compensated for their effort. Order also seems to reduce grade gaps, but only for those assigned smaller accuracy incentives. De-identifying students' information may be a way to circumvent grades being correlated with students' demographics. However, grades given in a blind grading environment may be lower, on average, than those in a non-blind environment. This grade difference may negatively affect students if lower grades impact their morale and willingness to invest in their education. For example, Ahn, Arcidiacono, Hopson, and Thomas (2019) find that female students may opt out of STEM degrees in higher education partly because STEM-based classes tend to give out low grades, even if female students are the highest performers.

Figure 6 End of Survey Question Results



However, the results from this paper indicate no substantial differences in grading behaviors between the blind grading and non-blind grading environments. The time on task and references to the grading rubric and class notes are nearly identical. One of this study's hypotheses is that there should be differences in effort, and I use time grading and the usage of a rubric/class notes as a proxy for effort. However, I find no evidence that behaviors related to effort are different. Instead, I show that all indicators of grading behaviors unanimously decrease while participants finish grading each subsequent essay, with no difference between grading environments. Additionally, the self-reported ratings from the end-of-survey questions provide evidence that participants do not measurably behave differently while grading. Participants in both grading environments report similar levels on questions about their grading experiences.

This study does have limitations that may need to be addressed for a full analysis of the effect of grading in a blind environment on the equity of grades. First, the results reflect the types of participants in the study. This study's population is drawn from the general population on Prolific. How teachers may grade in an educational setting could be different. Second, participants in the experiment are not allowed to revisit and change their grades. While this ensures the measurement of the treatment effect, it may not align with real-life teacher practices, where grade adjustments can occur. Further research is needed to investigate if teachers make such adjustments and whether they impact reducing grade gaps. Third, the results may be a lower bound of the true effect in the real world for two primary reasons. Participants know the essays are not from real students, and they may discriminate based on the students' skin color. If the essays are from real students, they may care more about giving partial credit or spending more time assessing the essays. I do not test for this, but skin color discrimination could impact how participants view or perceive students, especially if they believe the students are of a different race/ethnicity (i.e., Indian or Black) than intended in the experiment. Third, there may be non-linear relationships in behaviors and grades that I do not have the power to estimate because of the sample size. Fourth, this chapter's focus is grading behaviors related to time spent on each essay and the use of the grading rubric or class notes. Although I find no differences across treatment groups, participants may interact with the material differently. Unfortunately, I cannot measure how participants use the rubric or class notes.

Notwithstanding the limitations, the implications of this study's findings are profound. They suggest that discrimination in grading aligns with statistical discrimination where graders choose the information they deem important in forming their grading heuristics. This underscores the need for schools using Learning Management Systems (LMSs) to consider implementing a blind grading policy. Furthermore, institutions may want to pay teachers for their effort to grade accurately, by using some auditing mechanism that rewards teachers with bonuses for being fair in their grading.

In this chapter, I show in a laboratory experiment that decision-makers can engage in racial/ethnic and gender discriminate when they have to assign a value of quality to homework submissions. The experiment also shows that discrimination can decrease if decision-makers are paid for their accuracy. In the next chapter, I report results from a field experiment that assesses for racial/ethnic and gender discrimination in email communications, when signals of quality in the text also differ. The experiment is on discrimination in inquiries to law enforcement agencies about purchasing a firearm.

CHAPTER III
TO WHOM DOES THE SECOND AMENDMENT APPLY? EXPERIMENTAL EVIDENCE OF LAW ENFORCEMENT'S IMPACT ON EQUITABLE ACCESS TO FIREARMS

This chapter is co-authored with Garrett Stanford, who significantly contributed to this work by providing the initial experimental design, the names used in the experiment, the literature review, law enforcement data and code from previous research, writing the first draft with preliminary results and the research grant, cleaning the preliminary data and providing the initial data analysis, interviewing Research Assistants, and working on the power analysis. I was the primary contributor in the writing of this chapter and the pre-analysis plan, the data analysis, the data cleaning of the full data set, managing and working with the Research Assistant, gathering sheriff's offices data, conducting the experiment, coming up with the research question, and building on Garrett's experimental design with varying signals in the emails.

3.1 Introduction

In 2021, the United States witnessed 48,000 firearm-related deaths, with a concerning rise in such deaths among children by 50% between 2019 and 2021.[1]. This issue has re-sparked debates about addressing gun violence and enhancing public safety with stricter firearm laws. These discussions are often shaped by a delicate balance between protecting constitutional rights, such as the Second Amendment, and addressing the societal impact of gun violence. Without comprehensive federal policies, many states have begun implementing their own laws. One notable example is Oregon's Measure 114, passed in 2022 as

[1] PEW Research: "What the data says about gun deaths in the U.S." and PEW Research: "Gun deaths among U.S. children and teens rose 50% in two years"

a public referendum to bolster gun control. This program mandates that citizens must obtain a permit from local law enforcement agencies to buy firearms legally. Advocates argued that this law would help curb firearm-related deaths, given that past Research suggests that stricter gun laws can reduce firearm deaths.[2] However, Measure 114 faced steep opposition due to concerns not only about constitutional rights infringement but also about law enforcement potentially discriminating against minority racial groups, specifically Black and Hispanic individuals, in permit issuance.

Motivated by the tension between the evidenced efficacy of stricter gun control laws and the potential negative distributional consequences of empowering law enforcement agencies as gatekeepers of gun ownership, our study examines whether law enforcement agencies exhibit bias in their decisions to assist citizens in legally purchasing a firearm. More precisely, we test for a causal relationship between a citizen's race/ethnicity and gender and the response behavior of law enforcement agencies to a request for information concerning purchasing a firearm. To estimate this effect, we turn to the audit study design, using a similar methodology as Bertrand and Mullainathan (2004). In our experiment, we send email requests for assistance in purchasing guns to law enforcement agencies—both sheriff's offices and local police departments. Using distinctive names suggesting the demographic identities of the senders, we create six fictitious identities with three different racial/ethnic groups (Black, Hispanic and White) and two genders (male and female). We then randomly assign the racial/ethnic and gender identity shown on the email to each law enforcement agency. In addition

[2] For example, Rudolph, Stuart, Vernick, and Webster (2015) find the passage of a permit-to-purchase program in Connecticut was followed by decreases in homicides. Webster, Crifasi, and Vernick (2014) and Williams Jr (2020) illustrate the efficacy of permit-to-purchase laws by demonstrating an increase in homicides after Missouri ended its permit-to-purchase program.

to these racial/ethnic and gendered identities, we also randomly varied signals implying the quality of the sender (neutral, positive, and negative), following a similar design as Ewens, Tomlin, and Wang (2014).

Our results indicate an overall response rate of 46.75%. Response rates for emails from Black or Hispanic email senders are, respectively, 3 and 7 percentage points (pp) lower than the response rates for emails from white identities, statistically significant only for Hispanic email senders (at the 0.01% level). Response rates for emails assigned a female identity are 4 pp higher than the male identity. Among our six identities, white female email senders are the most likely to receive a response, and Hispanic email senders are the least likely to receive a response from law enforcement agencies. Responses from law enforcement are between 18-28% longer for female email senders, especially for Black and white female identities. We do not find any differences in the sentiment of the responses from law enforcement agencies.

In analyzing the impact of signals on response rates, we evaluate whether there are differences across the email sender identities based on the U.S. state requiring a firearm permit and on the signal in the email. States that require firearm permits responded to emails 12 pp more (significant at the 0.01% level) with text that is more positive in sentiment (0.14 units, significant at the 0.01% level). Generally, we do not find any gender effects based on the state laws. We do find that in U.S. states that do not require firearm permits, Hispanic email senders receive fewer responses than white email senders. However, in states that do require firearm permits, we do not find any difference in response rates between Hispanic and white email senders.

The effects of the signals in the emails on response rates, email length and sentiment show that response rates are lower for male senders when the assigned signal is negative and more significant when the assigned signal is positive. However, for the female identity, response rates are larger for both the negative and positive signals. For the racial/ethnic identities, the results are primarily insignificant, except that response rates for the Hispanic identity are lower for the neutral signal by 9 pp than the white identity. Furthermore, Black male email senders receive fewer responses when the signal is negative and more responses when the signal is positive. We also evaluate whether the content of the responses changes depending on the signal in the sent emails, and we find that the Hispanic identity receives longer emails for the positive signal than the neutral signal. Lastly, text sentiment is more negative for emails assigned a negative signal in general and for the Hispanic, white, female, and male identities.

The results of this study suggest that law enforcement may communicate differently with different groups of people interested in purchasing a firearm, especially between male and female email senders. We do find some racial/ethnic differences in the responses. Most notably, Hispanic email senders received fewer emails, and white females were more likely to receive a response from law enforcement. Our results do not explicitly indicate whether there is discrimination in firearm ownership.

Although this study examines the potential for discrimination in legally accessing firearms through law enforcement, it remains untested whether law enforcement denies civilians access to firearms on account of discrimination.

However, over-policing has been studied and confirmed in many different settings and contexts against Black and Hispanic communities.[3]

Our study examines several important areas of study. First, we contribute to a growing body of Research that seeks to causally test the existence and extent of bias in law enforcement practices. We examine whether there is racial/ethnic and gender discrimination while exploring the underlying mechanisms driving those biases. Second, we provide one of the first estimates of racial/ethnic and gender discrimination in the context of purchasing a firearm. Finally, our study provides a descriptive assessment of whether and how law enforcement agencies help civilians (independent of bias) in the process of purchasing firearms—likely an essential measure given the complexity of the ever-changing state-level gun laws in the United States and law enforcement's frequent central role in enforcing these laws.

In this paper, we will first discuss the current state of gun control and biased policing in the U.S. and motivate our decision to conduct an audit study. We then walk through our experimental design and then the empirical specification. Next, we report the main results of the experiment. We conduct a deep dive into an analysis of the heterogeneity in the results, specifically, how the results differ based on the email signals and whether U.S. states require a firearm permit. We conclude the paper with a brief discussion of the model of discrimination.

[3] Research finds that Black and Hispanic communities and individuals are more likely to experience a higher level of police presence and be stopped by the police (e.g., Bulman, 2019; Chen, Christensen, John, Owens, & Zhuo, 2021; Gelman, Fagan, & Kiss, 2007; Pierson et al., 2020). Additionally, during police-civilian interactions, they are more likely to experience excessive use of force, receive citations, and be targeted for asset forfeiture (e.g., Goncalves & Mello, 2021; Makowsky, Stratmann, & Tabarrok, 2019; Nix, Campbell, Byers, & Alpert, 2017; Ross, 2015; Sances & You, 2017; West, 2018). Furthermore, in civilian-initiated interactions, Black individuals are less likely to receive assistance Giulietti, Tonin, and Vlassopoulos (2019); Stanford (2023).

3.2 Background

Conversations surrounding firearms and firearm regulation play an increasingly important role in American politics. The firearm-regulation debate and research concerning firearms have both evolved over time in focus and framing (Carlson, 2020; Steidley & Yamane, 2022). However, the central tenets of proponents and opponents of firearm regulation remain relatively unchanged. Those in favor of regulation often cite the large number of injuries and deaths related to firearms each year. On the other hand, opponents of regulation frequently argue that access to firearms is a constitutional right and that firearms are an important tool for citizens to keep themselves safe. Carlson (2020) highlights that both sides argue that "evidence matters" and the other side "ignore[s] the facts."

Currently, firearms are regulated at both the national and state levels. Since the National Firearms Act of 1934, the expansion of federal regulation has been modest.[4] Until the passage of the Bipartisan Safer Communities Act (2022) under the Biden administration, the most recent national-level firearm regulation policies were the Brady Handgun Prevention Act (1993) and the Federal Assault Weapon Ban (1994–2004). In the latter half of the 20th century and the early 21st century, federal-level firearm regulation has often been catalyzed by either high-profile firearm-related incidents or by rising crime rates.

In contrast to the few changes in federal regulation, there have been significant changes to state-level firearm laws in the United States over the last 20 years. Some states have enacted more-stringent regulations on firearm ownership, while others have loosened their laws. A thorough discussion of firearm laws is

[4] Vizzard (2015) provides a good overview of firearm policy in the United States.

beyond the scope of this paper. However, we briefly discuss two firearm-related laws relevant to this paper: background check laws and permit-to-purchase laws.

Background Checks: Federal law requires a background check be performed through the FBI's National Instant Criminal Background System (NICS) for all purchases from a federal firearms licensee (FFL), manufacturer, or importer. There are several modifications states have made to this federal mandate. First, depending on state law, FFLs either communicate with the NICS directly or alert a state-designated point of contact that then communicates with the NICS. Second, states have adopted laws that expanded background check requirements beyond FFL sales. For example, states with Universal Background Check laws (UBC) require background checks for all firearm sales and transfers (e.g., sales between private parties, sales at gun shows, or gifts). Third, states may run their own background checks that access state records that are not included in a NICS background check. Finally, state laws may allow for the substitution of an ATF-qualified alternate permit that can act in place of an NICS background check. However, an individual must undergo a background check to obtain a permit.

Permit-to-Purchase: A permit-to-purchase (P2P) law requires individuals to have a permit or license before purchasing a firearm. In contrast to background checks, P2P laws have been implemented only at the state level. An individual must apply for a permit at a local agency to obtain a permit. Typically, the local agency is a law enforcement, and the application must be completed in person. Depending on state law, successfully obtaining a permit may requirements in addition to a background check (e.g., a gun-safety course). Like UBC laws, P2P laws ensure that all gun owners are subject to a background check. Some states have adopted both UBC laws and P2P laws, meaning that a background check is performed for an individual at the time of the permit application and an additional time at the point of sale.

Gun laws in Oregon In November 2022, nearly two million Oregon voters narrowly passed Measure 114, an initiative aimed to curb access to firearms and high-capacity magazines. Measure 114 would require buyers to receive a permit from local law enforcement before purchasing a firearm. In contrast to previous background check requirements, Measure 114 would result in permits being denied more easily based on concerns over an individual's psychological state. Additionally, permits are contingent on demonstrated completion of a firearm safety course. Permits under Measure 114 are valid for five years. Beyond the gun-permitting requirements, Measure 114 would make it a criminal offense to possess magazines capable of holding ten or more rounds.

The passage of Measure 114 was controversial throughout Oregon and followed a decade of gradually strengthening gun legislation in the state. Between 2015 and 2021, Oregon expanded its background check data infrastructure, implemented a "red flag" law enabling judges to order the removal of firearms from at-risk individuals upon a petition from a household member or law enforcement agency, extended gun restrictions for those under a restraining order or convicted of stalking, and mandated safe firearm storage practices. With the added provisions of Measure 114, Oregon rose in Everytown's gun law strength rankings from 11th in the country to 9th in January 2023.

A slight majority of Oregonians celebrated the passage of Measure 114 and its promise for improved gun safety. Opposition fell into two visible camps. Perhaps unsurprisingly, one group was comprised of right-leaning conservatives from rural

Oregon. However, opposition also came from progressives who argued that law enforcement agencies could not be trusted to enforce Measure 114 in a neutral way. In Oregon's November-2023 voter pamphlet, one of the opposing arguments read:

> Gun violence is an urgent problem that needs effective action; however, Measure 114 has serious potential to harm some of our most vulnerable communities. It is important to consider POC (people of color) and people of marginalized genders who will be unfairly targeted by poorly written, misleading, and unclear laws, which is why we urge you to vote NO on this bill.

The conception of our study is motivated by Oregon's Measure 114. However, law enforcement agencies' involvement in the process of legally obtaining a firearm extends beyond Oregon's borders.

Law Enforcement Biases A growing body of research highlights the disproportionate burden that policing can place on people of color in these various contexts. For example, Chen et al. (2021) find that neighborhoods with larger Black populations experience a considerably higher police presence. Similarly, there is evidence that people of color are more likely to be stopped by police (e.g., Bulman, 2019; Gelman et al., 2007; Pierson et al., 2020). There is also evidence that the result of a police-citizen interaction depends on the citizen's race/ethnicity. Numerous studies find that people of color are more likely to experience the use of force by police (e.g., Edwards, Lee, & Esposito, 2019; Fryer, 2020; Nix et al., 2017; Ross, 2015). People of color are also more likely to be targeted for traffic citations and asset forfeitures (e.g., Goncalves & Mello, 2021; Makowsky et al., 2019; Sances & You, 2017; West, 2018).

Research remains limited, however, on biases in civilian-initiated interactions. Stroube (2021) documents that, in Chicago, formal complaints made by Black residents were less likely to be sustained than formal complaints made by White residents. Stanford (2023) finds that local police departments are less likely to respond to requests for assistance in making formal complaints when the requests are signed with distinctively Black or Hispanic names compared to white names. Similarly, Giulietti et al. (2019) find lower response rates from sheriff's offices for emailed inquiries about a "lost and found" when the request comes from a Black-sounding email address. In the case of automobile accidents, which are to some degree citizen-initiated, West (2018) finds that law enforcement officers are more lenient when interacting with same-race civilians.

Establishing causality in the context of biased policing is difficult. First, differences in the frequency of interaction do not necessarily reflect biased policing. There is the possibility of a systemic selection problem. Consider the scenario where sociodemographic groups participate in criminal activity at different frequencies. In this case, unbiased policing could still result in heterogeneous rates of police-citizen interactions across sociodemographic groups (Fridell, 2017). Second, measuring biased policing by comparing outcomes for citizens conditional on an interaction with police does not permit causal inference. Suppose the motivation for police initiating an interaction with a citizen is biased, even though outcomes for all police-citizen interactions are similar. In that case, a naive analysis can obscure the presence of biased policing (e.g., Knox, Lowe, & Mummolo, 2020; Ross, Winterhalder, & McElreath, 2018). Consequently, researchers remain divided on the existence and extent of biased policing (Fridell, 2017; Smith, Rojek, Petrocelli, & Withrow, 2017). Furthermore, while some researchers employ research designs that permit causal inferences, many of these studies rely on self-reported police data. Such reliance on police-reported data can lead to inconclusive or incorrect conclusions if police departments strategically or unintentionally underreport or misreport (e.g., Luh, 2020).

Audit studies We use this particular experimental design for several reasons. The challenge of causally identifying discrimination is not unique to the context of law enforcement. Over the last decade, correspondence studies, a type of randomized controlled trial (RCT), have become an increasingly popular tool for researchers studying the presence of discrimination (Bertrand & Duflo, 2017).[5] Emulating the seminal work of Bertrand and Mullainathan (2004), researchers have used correspondence studies to identify a variety of types of discrimination (e.g., gender, age, or race/ethnicity) in various contexts (e.g., housing, medical services). To date, most correspondence studies focus on Black versus White discrimination, primarily in the context of hiring practices. There are very few audit or correspondence studies that focus on discrimination in the provision of public services in the United States (Butler & Broockman, 2011; Einstein & Glick, 2017; Oberfield & Incantalupo, 2021; White, Nathan, & Faller, 2015). Considering that marginalized groups, on average, are more likely to depend on public services, discriminatory practices are of utmost concern for social planners.

To the best of our knowledge, the only prior correspondence studies that concern law enforcement agencies are Giulietti et al. (2019) and Stanford (2023).

[5] In a correspondence study, individuals (often fictitious)—who are identical in terms of all observable characteristics other than the characteristic of interest—apply for a job, service, or good. The researcher then examines whether the experimentally varied characteristic of interest affects the outcome of the application or request (Bertrand & Duflo, 2017). The present study uses email instead of the traditional approach of "snail mail," and explained below—requests assistance from law enforcement agencies instead of applying for jobs or making purchases.

Giulietti et al. (2019) conduct a correspondence study with a wide range of public institutions. Included in their list of public institutions are sheriff's offices. In their study, the authors email the various public institutions with benign requests for information. The authors vary the identity of the requesters, using two distinctively black male names and two distinctively white male names. The authors find that these public institutions (ranging from public libraries to sheriff's offices) are less likely to respond to emails from individuals with distinctively black names. Furthermore, among the various institutions, this effect is most pronounced for sheriff's offices.

The current study is motivated by the results of Stanford (2023), and recreates much of its research design. Stanford (2023) tests for law enforcement bias by requesting assistance in making a complaint against an officer from a local police department. The results of the study show that police are significantly less likely to respond to requests from emails signed with Black or Hispanic names. Our study deviates from Stanford (2023) in scope and the content of the emails. First, we contact sheriff's offices in addition to local police departments. Second, we ask for assistance in purchasing a gun rather than assistance in filing a complaint. In addition, we include a criminal background signal that is varied across emails—a treatment dimension not included in Stanford (2023).

By using a correspondence study, we overcome two of the main challenges in studying discrimination in the context of law enforcement: (1) finding causal estimates (as opposed to mere associations) and (2) avoiding potentially compromised self-reporting of data collected or provided by law enforcement agencies. Estimates from properly randomized correspondence studies can be reasonably assumed to be causal. As mentioned above, the primary obstacles to causal inference in the study of potentially biased policing arise from systematic selection—by types of people into criminal activities and stemming from police discretion concerning with whom they interact. We avoid these challenges by creating a citizen-initiated police interaction not predicated on a crime taking place and by designating our outcome of interest as the police department's decision to interact with the potential complainant.[6]

Avoiding the use of administrative data provided by law enforcement agencies has significant advantages. First, intentionally or inadvertently, departments can have ongoing difficulties reporting accurate data (e.g., Luh, 2020). Second, police data can be a product of subjective reporting by individual officers and department-specific classification conventions.[7] Even when police officers honestly record officer conduct, decisions made in the heat of the moment during any given citizen-officer interaction could influence how events are recorded. Finally, agencies may be unwilling to disclose "sensitive information." [8]

A correspondence study also allows us to use a national sample of police departments. We use a sample of approximately 4,000 law enforcement agencies representing all states except Hawaii and Alaska.[9] As a result, the measures of biased policing from this study represent policing in the mainland portion of the United States on average rather than for any specific state, county, or city.

[6] We use "citizen" as shorthand for "member of the community" and not imply any formal citizen/ resident alien/ illegal alien distinction.

[7] For instance, PolicingProject.org describes the discrepancies across states in reporting requirements for officer-initiated stops. RevealNews.org finds that the Washington D.C. police department has a comparatively loose definition of "resisting arrest.".

[8] Weisburst (2019) does not find evidence of racially biased policing in Dallas. However, Weisburst hypothesizes that the department's willingness to disclose its data to researchers might stem from the fact that the Dallas police do not appear to have a problem with biased policing.

[9] Hawaii's exclusion was a result of random selection. Hawaii only has four distinct police departments and five sheriff's offices.

Consequently, inferences made in this study are more likely to reflect systemic nationwide behavior patterns rather than specific department cultures.

3.3 Experimental Methodology and Data

In this section, we describe the design and implementation of our correspondence study. The objective of the study is to test whether law enforcement agencies exhibit signs of racial/ethnic or gender discrimination to citizen-initiated requests for help purchasing a firearm. In broad strokes, this study collects contact information for a nationwide sample of police departments and Sheriff's offices in the U.S. and then emails each department using one randomly assigned instance from a specially designed set of "identities" that we create.[10]

3.3.1 Experimental Methodology. To test our hypotheses that inquiries about firearm purchases differ across race/ethnicity and gender, we experimentally manipulate race/ethnicity, gender, and signals inquiring law enforcement agencies in the United States whether a permit is required to purchase a firearm.

Experimental Subjects: Law Enforcement Agencies (LEAs) The law enforcement agencies included in this study are a stratified sample of 4,000 agencies. For inclusion in the study, we select police departments that are associated with a local government (i.e., no state police) and that serve a population of at least 7,500 people and Sheriff's offices that operate in counties with populations of 10,000 or more. We designed the experiment so that half of the agencies are police departments and half are Sheriff's offices.

[10] We preregistered this experiment at the AEA RCT Registry, and the pre-analysis plan can be found here. Additionally, we have IRB approval for conducting this experiment, which can be found HERE.

To assemble our list of Sheriff's offices, we use the 2018 Census of State and Local Law Enforcement Agencies (CSLLEA) to identify sheriff agencies in counties with at least 10,000 individuals. The CSLLEA survey is conducted every 4 years. It reports information on the characteristics of the agencies, including the number of employees and the type of tasks the agencies do, such as whether the LEA conducts background checks. This list of Sheriff's offices included 2,354 sites. Unfortunately, the CSLLEA does not include contact information. To identify email addresses for each Sheriff's office, we searched the internet for an email address for the corresponding Sheriff's office. Some sheriffs did not have publicly available email addresses, or they only had an online form to contact the office. We recorded each of these outcomes when they occurred and dropped the county in question from the study if there was no email address or form. After going through each agency, we identified 1905 sheriff's offices in which 228 of those had only online forms for their contact information. Local police departments included in this study come from Stanford (2023), and followed a similar process.[11] Through our sampling process, we identified 1,906 sheriff's offices and 2,080 local police departments as both eligible and able to receive emails, representing 49 states.[12]

Identity Creation of the Fictional Email Senders We use the names of the putative email senders to signal race/ethnicity and gender. We created six broad categories of identities for this study: Black female, Black male,

[11] Please refer to Stanford (2023) for details on the selection process. One difference between the agencies is that the police departments did not include agencies with only online forms for their contact information.

[12] We excluded both Hawaii and Alaska from our sample, so our experiment represents only law enforcement agencies in the mainland. As noted above, Hawaii's exclusion from the study was an unintentional result of the sampling process, and Alaska's was deliberate, as they had no sheriff's offices.

Hispanic female, Hispanic male, White female, and White male. Sixty unique first-name/last-name combinations are specified for each type of identity.

For this study, we borrow the names used in Stanford (2023). The last names are drawn from the "Frequently Occurring Surnames in the 2010 Census" dataset. The names are selected based on the criteria that they are racially distinctive while also commonly occurring. We select six last names for each race/ethnicity to ensure our results reflect the effect of race/ethnicity rather than a fixed effect for a particular name.

First names are also borrowed from Stanford (2023). These names were motivated by Gaddis (2017b) and Gaddis (2017a). Gaddis conducts two experiments that explicitly test which first and last names are racially and ethnically distinctive for African-American (Gaddis, 2017b) and Hispanic (Gaddis, 2017a) individuals. We use the ten most distinctive first names for the respective identities from these two studies. In total, we created 360 unique names (6 identities × 6 last names × 10 first names). After selecting the names for each identity, We created a unique email address for each last name used in the study (e.g., *snyderrr.1992@examplemail.com*). We then created a unique email address profile for each identity (e.g., *Dustin Snyder ¡snyderrr.1992@examplemail.com¿*). As a result, the full names of the identities were visible in the email inboxes of the police departments (e.g., *Dustin Snyder ¡snyderrr.1992@examplemail.com¿*).

The complete list of names can be inferred from Tables G.1 and G.2 (360 unique name combinations). We ultimately omitted six potentially recognizable celebrity names from our set of names: Denzel Washington, Tyra Banks, DaShawn Jackson, Seth Meyer(s), Katelyn Olson, and Pedro Martinez. These names have widespread recognition, and during the testing process, respondents noted that they strongly associated these names with celebrities who have the same name.

Email Each law enforcement agency receives one email from a single randomly assigned identity. We create three types of emails: *negative*, *neutral*, and *positive*. The different email types reflect varying degrees of lawfulness and willingness to comply with law enforcement. As seen below, a single line referencing background checks is included in the *positive* and *negative* emails. We include a reference to a background check because it is relevant to purchasing a firearm and insinuates the "quality" of the applicant in terms of compliance and lawfulness.[13] Emails also vary by the identity of the sender. We randomly assigned the email type. We use the email sender's name twice in each email to increase the salience of this information. The three email types can be seen in Table 3.3.1.

The **bold** words indicate that these words changed across emails. As described above, we created profiles for the email accounts so that agencies would see the sender's full name twice and their first name three times. Law enforcement agencies were addressed directly with a hello.

[13] Federal law requires a background check to be conducted at the time of purchase when purchasing firearms from vendors with Federal Firearm Licenses. Many states have adopted universal background checks. Given the confusing patchwork nature of firearm laws in the U.S., it is believable that a civilian would need clarification on the matter.

Assigned Email Signals		
NEGATIVE	NEUTRAL	POSITIVE
”Hello, My name is [**first name varying by email**]. I am interested in purchasing a gun, but I’m not sure where or how to start the process. Do I need a permit? If so, how do I apply for one? *If a background check is required, how far back does it go into one’s history?* Best, [**Name, varying by email**]”	”Hello, My name is [**first name varying by email**]. I am interested in purchasing a gun, but I’m not sure where or how to start the process. Do I need a permit? If so, how do I apply for one? Best, [**Name, varying by email**]”	”Hello, My name is [**first name varying by email**]. I am interested in purchasing a gun, but I’m not sure where or how to start the process. Do I need a permit? If so, how do I apply for one? *I don’t mind going through a background check if I need to do that.* Best, [**Name, varying by email**]”

Timing We conducteded the study over an eight-week period, from mid-October 2023 to early December 2023.[14] We sent roughly 500 weekly emails, split across Tuesday, Wednesday, and Thursday. We automatically sent emails from 8 a.m. to 3 p.m., except for the emails we had to manually send on each Sheriff’s

[14] Due to technical difficulties, we had to pause the experiment during the third week of the rollout. Consequently, the experiment ran for nine weeks with one bye-week.

webpage, which were manually sent throughout the day. We decided on an eight-week window to minimize the chance that a single unanticipated news event or holiday would compromise the generalizability of the results. We split the emails across days of the week and time day to reduce the logistical difficulty of sending the emails. We decided not to send emails on Thanksgiving or weekend days to give departments at least two full weeks to respond to the inquiry.

Treatment Assignment The "treatment" here is the identity (race/ethnicity and gender). Although we randomly assign the email type (neutral, positive, and negative) law enforcement sees, the sample size for each signal is underpowered, so we treat the results for those as more descriptive. We stratify treatments by week, state, and agency type (i.e., local police department or Sheriff's office). As a result, the number of departments for each agency type for each state is balanced each week. Treatment is then randomly assigned across agencies within each week-state-agency type stratum.

Balance Table Data. The study includes data for several other observable department characteristics. These characteristics, ex-ante, seemed to be potentially important determinants of the response behaviors of police departments: numbers of officers and civilian employees for each department, local crime levels, county-level income information, and county-level racial/ethnic composition. We use the data to identify whether the characteristics of the populations we treated are similar across treatment.[15].

We compile income and race/ethnicity data from the 2021 American Community Survey (U.S. Census Bureau, 2021). Police departments selected for

[15] Note that the final data set is smaller than the initial data set. After finishing the experiment, we had to drop some data. More information is in Appendix F

the study are associated with governments smaller than counties. However, it is not clear with exactly which population each department would interact. If we use data for a geography that is too precise (e.g., the zip code of the department), we risk mischaracterizing a department's local context. Accordingly, we use county-level data to characterize the economic and racial composition of a department's environs. We sacrifice some precision with this approach but avoid inaccuracy. This dilemma is not relevant for sheriff's offices, although some police departments and sheriff's offices share the same county. We show the distribution of those agencies in Table H.1.

We use the Uniform Crime Reporting (UCR) Program data compiled by Kaplan (2023b) for employee counts for each agency. The UCR dataset includes employee counts through 2020. We use arrest data from the NIBERS/Uniform Crime Reporting (UCR) Program data compiled by Kaplan (2023a) as a proxy for crime in the area. This data is in terms of total crime per county for the years between 2017-2021.

We also include two more gun-specific data variables. First, we use data from the K-12 School Shooting Database for information about school shootings back until the 1970s. This data includes the news article of the school shooting, the date, the information about the shooter, and a longitude/latitude indicator of the school. However, it did not include the county in which the shooting took place. We used the longitude/latitude data to identify the county. After which, the data we use in our balance table is the total number of school shootings that took place in each county from the year 2000 through 2023, divided by the population in each county. Because the count of shootings is so small, the numbers presented in our balance table are in 1000s of a percent. Finally, we use data from Every Town Research to identify the US state that requires a concealed carry firearm permit.[16]

Table 15 shows relevant department characteristics using the whole data set. Column (1) of the table is the mean value for each different characteristic for agencies that received emails from White-male identities. Columns (2) through (6) are the differences between the White-male mean value and the mean values for other identities. The final column indicates the number of counties that had available data. Several differences were statistically significant; for example, the average median Black income was higher for agencies contacted by Hispanic men than for agencies contacted by white men. However, some correlations should exist in expectation, and Table 15 largely confirms that the treatment was successfully randomized across the most obvious department and county characteristics relevant to this study.

[16] Some states require a permit to just purchase a firearm. A concealed carry permit is more stringent, as it allows owners to carry a gun into a public space.

Table 15 Balance Table - Differences in state,county, and agency level characteristics on the assigned race and gender

	Male			Female			
	White (Intercept)	Black	Hispanic	Black	Hispanic	White	N
Firearm Permit Required in State (1 = Yes)	0.56***	0	0.01	0	0	-0.01	3762
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	
Median Household Income - County level							
All	58161.93***	167.67	824.58	758.01	420.25	17.17	3762
	(1917.46)	(816.39)	(806.44)	(850.08)	(849.34)	(835.68)	
Black	37584.03***	1449.72	2614.18**	533.97	908.5	1591.47	3119
	(2041.13)	(1147.86)	(1212.17)	(1166.26)	(1170.81)	(1178.34)	
Hispanic	50191.04***	622.69	-5.96	1744.34*	1101.42	1145.45	3474
	(2692.12)	(997.43)	(916.55)	(1030.31)	(1000.07)	(1018.49)	
White	69872.04***	-366.5	495.11	127.6	-524.52	-1110.94	3762
	(2241.13)	(880.41)	(844.85)	(909.85)	(881.21)	(880.8)	
Population (1000s) - County level							
All	501.91***	-54.84	-23.14	-45.9	-60.34	-48.94	3762
	(87.99)	(51.29)	(41.73)	(48.02)	(48.77)	(38.9)	
Black	104.23***	-9.3	-3.36	-10.71	-11.36*	-12.95**	3762
	(25.84)	(6.01)	(6.85)	(8.45)	(6.67)	(5.55)	
Hispanic	100.69***	-19.48	-5.3	-15.28	-23.57	-18.57	3762
	(34.81)	(24.32)	(18.27)	(19.05)	(22.53)	(16.17)	
White	248.21***	-17.93	-6.91	-15.75	-14.16	-10.11	3762
	(31.33)	(15.68)	(14.47)	(17.39)	(15.55)	(14.84)	
Law Enforcement Agency - Agency Level							
Agency Type (1=Police)	0.32***	0.01	0	0.03	0.01	0.02	3762
	(0.08)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	
Officers	81.08***	8.2	12.55	10.88	0.28	-7.63	3753
	(25.68)	(12.85)	(17.94)	(10.97)	(10.04)	(9.84)	
Civilian Workers	15.82	8.39	10.24	11.8	2.75	0.85	3753
	(13.39)	(6.99)	(7.9)	(7.42)	(5.35)	(5.36)	
Assaults on Officers	5.54***	0.15	-0.76	1.52	0.32	0	3753
	(1.82)	(1.37)	(1.06)	(1.38)	(1.66)	(1.19)	
Firearm Assaults on Officers	0.32*	0.14	0.06	0.21	-0.03	0.02	3753
	(0.16)	(0.16)	(0.14)	(0.13)	(0.12)	(0.12)	
Background Checks (1=Yes)	0.58***	-0.01	0	0.01	0	0.01	3621
	(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	
Crime - County level							
Crime per Person (2017-2021)	4.53***	0.16	0.21	-0.22	0.12	-0.13	3762
	(0.66)	(0.38)	(0.4)	(0.37)	(0.37)	(0.35)	
School Shootings per Person (2000-2023)	0.63***	-0.05	-0.07	-0.07*	-0.06	-0.05	3762
	(0.15)	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	

Note: Column 1 is the county-level characteristic. All regressions contain week fixed effects, state-level fixed effects, and agency type fixed effects. We do not include state-level fixed effects for variable 'Firearm Permit Required' or agency fixed effects for 'Agency Type'. Column 2 is the average number of assigned white male identities with the fixed effects. Columns 3 through 6 are the mean differences between each putative identity and the white male identity. Column 7 is the number of observations in each regression. Clustered SEs at the county level as the data includes both sheriffs (at the county level) and police departments (within the county).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

3.4 Empirical Specification and Summary Statistics

Empirical Specification. In this section, we detail how we analyze our data. We aim to identify if there are differences in how law enforcement responds to inquiries about how to purchase a firearm and whether a permit is necessary based on the putative race/ethnicity/gender of the identities we assigned to each email. Furthermore, we also explore how the signals assigned to each email we sent contributed to the responses.

Outcome Variables. The outcomes of interest, (Y), in our study, include (i) response rates, (ii) the word count of the responses, and (iii) the sentiment of the text from law enforcement. The variable for (i) response rates is a dummy variable equal to 1 if the law enforcement agency had responded at least once and zero otherwise. We use the linear probability model (OLS) regression, as the regressor is dichotomous. For (ii) word count and (iii) sentiment, we analyze only the responses. For these, we use a standard OLS regression in analyzing these regressors, except word count is the logarithm of the word count due to the density of word count is highly skewed towards zero (see Figure H.4). Our results for the word count must be exponentiated to fully understand how word count differs across treatment. The (iii) sentiment of the word text is further explained in detail in our Data section. We do not transform the data, as the data ranges between -1 and +1 and the logarithm of zero is indefinite. However, the data is skewed more towards +1, reducing the precision of our estimates (see Figure H.5).

Estimator. In all of our analysis, we use the difference-in-means estimator. Our independent variables are dummy variables or factor variables of the putative race/ethnicity and gender of the assigned identities and the assigned signals and state permit laws. We are interested in the mean responses for each factor variable, as they reflect the difference in outcomes relative to the omitted variable (the omitted demographic variables are either white or male). In some cases, we interact multiple factor variables together, such as how the outcome variables change depending on the assigned signal and racial identity. With interactions between factor variables, we conduct separate hypothesis tests to identify whether the differences are statistically significant.

Controls. To improve the precision of our estimates, given that we stratified treatment by week, state, and agency type, we included fixed effects for the week we sent the emails, the US state of the agency, and the agency type. Including or excluding these variables does not change the mean/average outcome but reduces standard errors. Although we gathered other data to illustrate the balance of our treatment, we do not need to include those variables as controls. However, we could assess how our outcomes change depending on those variables (such as county-level income or total crime).

Clustering Standard Errors. Because treatment is at the agency level, many law enforcement agencies are in the same county. In all of our regressions, we cluster the standard errors at the county level to eliminate the correlation in our error terms associated with law enforcement agencies being in the same county or jurisdiction.

Regression of the Main Results. The following regressions are what we use to analyze our outcome variables. We separately analyze how race/ethnicity and gender impact the outcomes interest in the following specifications:

$$(1) \quad Y = \beta_0 + \alpha_1 \times \text{Black} + \alpha_2 \times \text{Hispanic} + \gamma \times \text{FE} + \epsilon$$

$$(2) \quad Y = \beta_0 + \beta_1 \times \text{Female} + \gamma \times \text{FE} + \epsilon$$

In regression (1), Black and Hispanic are indicator variables if the assigned identity is Black or Hispanic. The omitted variable is the white identity. α_1 and α_2 are the coefficients of interest as they show the average difference of outcome variable for the Black and Hispanic identity relative to the white identity. In regression (2), our dependent variable is a dummy variable if the assigned identity is female. β_1 is the average difference of the female identity relative to the male identity.

We separately analyze the following hypothesis tests that α_1, α_2, and β_1 are equal to zero. If we reject our hypotheses, then it implies that there are significant differences in average outcomes between the specific demographic of interest and the white or male identity.

Regression of the Heterogeneity Results. To understand how we analyze the heterogeneity in our data, we show the specification that we use to analyze how the outcome of interest may differ by the assigned gender and email signal in the regression below:

$$(1) \quad Y = \delta_0 + \delta_1 \text{Female} + \delta_2 \text{Pos} + \delta_3 \text{Neg} + \delta_4 \text{Female} \times \text{Pos} + \delta_5 \text{Female} \times \text{Neg} + \gamma \text{FE} + \epsilon$$

In this regression, we interact the identity of the signal on an indicator of the assigned gender (female). ‘Pos’ is a dummy variable for whether the signal is positive or not and ‘Neg’ is the same except the signal is negative or not. The omitted variable in this case is male and neutral signal.

To understand how the signals (or another variable such as the agency type) impact our outcome variable, Y, we conduct several hypothesis tests. [17] In this

[17] Note that the number of hypothesis tests we conduct increases in the number of factors being interacted.

specific example, we test nine different hypotheses on how the outcomes differ between the female and male identity, conditional on each signal, and conditional on the gender, how the outcomes change across signals. We list the nine specific hypothesis tests in the Appendix: Table G.6.

3.5 Main Results

In this section, we discuss how our outcomes of interest differ by the putative race/ethnicity and gender assigned to each email we sent to law enforcement agencies. We show the main results in Table 16 for each outcome of interest. We also show the results by race/ethnicity by gender identities in Table 17 to illustrate which identities may be driving our results in Table 16.

Response Rates. In column (1) and column (2) of Table 16, we present the results from regressing an indicator variable that the law enforcement agency responded at least once on indicator variables of (1) race/ethnicity or (2) gender. We include fixed effects for week, US state, and agency type. All standard errors are clustered at the county level.

The omitted variable in column (1) is white, and in column (2) is male. The point estimates are the differences by percentage points (pp) in the response rates. Hence, the point estimate on female in column 2 can be translated as “the difference in the response rates between the female and male identity is 4 pp (standard error of 2 pp), significant at the 5% level.” Both the Hispanic identity and the female identity show statistically significant different response rates relative to the white and male identity, respectively. The Hispanic identity received fewer responses by 7 pp (standard error of 2 pp), significant at the 0.1% level.

We also show the response rates by “race/ethnicity by gender” indicator in column (1) of Table 16. In this table, we compare the differences between the various combinations of race/ethnicity and gender—Black female, Black male, Hispanic female, Hispanic male, and white female—relative to the omitted variable, white male. We find that all differences relative to white male are insignificant except for the white female identity (6 pp difference with a standard error of 3 pp, significant at the 5% level) and the Hispanic male (-5 pp with a standard error of 3 pp, significant at the 10% level).

Hence, the results in columns (1) and (2) of Table 16 could be partially explained by white females receiving significantly more responses than white males and both Hispanic males and Hispanic females receiving fewer responses than white males (although we fail to reject the hypothesis that this is significantly different from zero with a p-value of 0.05).

Table 16 Main Results

	Response Rate		Length		log(Length)		Sentiment	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black	-0.03		-0.55		0.04		0.02	
	(0.02)		(3.74)		(0.05)		(0.03)	
Hispanic	-0.07^{***}		-8.64^{**}		-0.07		-0.01	
	(0.02)		(2.98)		(0.05)		(0.03)	
Female		0.04^{*}		6.22^{*}		0.17^{***}		0.02
		(0.02)		(2.73)		(0.04)		(0.02)
Num. obs.	3762	3762	3762	3762	1759	1759	1758	1758
R^2 (full model)	0.06	0.06	0.03	0.03	0.06	0.06	0.08	0.08
Omit. Var.	White	Male	White	Male	White	Male	White	Male
Mean Omit. Var.	0.50	0.45	40.48	34.54	4.01	3.92	0.32	0.32

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

All regressions include week, US state, and agency type fixed effects. Standard errors are robust and clustered at the county level. In columns 1 and 2, we regress a dummy variable on whether law enforcement responded to the sent emails that are assigned a putative race or gender. Columns 3 through 6 are analysis of the responses from law enforcement. In columns 3 and 4, we took the logarithm the word count of the responses and regressed it on indicators of race or gender on the assigned sent emails. In columns 5 and 6, we use the Vader R package to create a sentiment score of the responses from law enforcement that range between -1 (most negative) to +1 (most positive) and regressed those values on the assigned indicators.

Table 17 Main Results: Race by Gender

	Response Rate	Length	log(Length)	Sentiment
	(1)	(2)	(3)	(4)
Black Female	0.02	6.01	0.24^{***}	0.05
	(0.03)	(4.23)	(0.07)	(0.04)
Black Male	-0.01	5.56	0.11	0.04
	(0.03)	(5.85)	(0.08)	(0.04)
Hispanic Female	-0.03	0.59	0.12	0.02
	(0.03)	(4.13)	(0.08)	(0.04)
Hispanic Male	$-0.05^{\cdot}$	-5.31	0.02	0.01
	(0.03)	(3.81)	(0.07)	(0.04)
White Female	0.06^{*}	12.45^{**}	0.26^{***}	0.04
	(0.03)	(4.57)	(0.07)	(0.04)
Num. obs.	3762	3762	1759	1758
R^2 (full model)	0.06	0.03	0.07	0.08
Omit. Var.	White Male	White Male	White Male	White Male
Mean Omit. Var.	0.47	34.36	3.86	0.29

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Description of the regressions can be read in Table 16.

Email Length. We also explore how the word count in the responses from law enforcement differs across the putative identities, conditional on receiving an email from law enforcement. In columns (3)-(6) of Table 16, we report the results from regressing the word count of the email response on the treatment variables. Columns (3) and (4) report the results when treating no response as an email with zero words. Columns (5) and (6) report the logarithm of the word count conditional on receiving a response from law enforcement. We find in column (6) that emails to the female email sender are longer by about 17%, significant at the 0.1% level).

Column (3) of Table 17 reports the results based on the race/ethnicity and gender of the identity. We find that word length in the responses from law enforcement is statistically much longer for both Black and white females, about 25% (statistically significant at the 1% level) and 28% (statistically significant at the 0.1% level) longer than white males.

Email Sentiment. Lastly, we assess how the sentiment of the text differs relative to the white identity, the male identity, and the white male identity. The data is reported as an index between -1 and +1, with -1 being the text that is the most negative and +1 being the text that is the most positive. Anything above 0 is generally considered positive. To determine the sentiment, we use the Vader package in the R platform.[18] We report the results in columns (7) and (8) in Table 16 and in column (4) in Table 17. We find no significant effects. There are a few possible reasons for this. Firstly, the Vader package we use to identify the sentiment of the text in R treats any word about guns and firearms as negative. So, even if an email is beneficial and long, the algorithm does not properly capture that sentiment. Secondly, the distribution of the sentiment text is skewed towards +1 (see Figure H.5), reducing the precision of our estimates. Our results remain inconclusive that law enforcement responds more or less favorably across the identities, and we fail to reject any hypotheses that there are differences in text sentiment.

3.6 Heterogeneity

In this section, we analyze the heterogeneity in our main results based on the identities and signals assigned to the emails. We also explore other heterogeneity—the differences in the outcomes based on the legal status of firearm permits at the state level, the differences based on the law enforcement agency

[18] The details on the package can be found here.

type, and then the differences based on whether the agency conducts background checks. However, we leave that analysis within our appendix.

State Permit Laws. Not all US states require firearm permits. Twenty-one states require permits to carry a concealed firearm. To see if there are differences in the results based on whether a state requires firearm permits, we use those twenty-one states to create an indicator variable equal to 1 if the state requires a permit and 0 otherwise. Table 18 shows the differences in the outcome variables based on whether the state requires a firearm permit. All columns, except for column (3), indicate that states that require a firearm permit are more likely to respond to the emails (12 pp, significant at the .1% level), send more positive responses (0.14 units more positive, significant at the 0.1% level), and possibly send longer emails (8.41 words longer on average in column 2, significant at the 1% level).

We evaluate the heterogeneity of the treatment effects based on whether the state requires a firearm permit. This is shown in Table 19. Note that because we use state-fixed effects, there is no main effect on 'Permit.' Columns (1) and (2) show no difference in response rates between Black and female email senders relative to white and male email senders. Hispanic email senders receive fewer responses from law enforcement (9 pp, significant at the 0.01% level). However, the interaction of 'Permit' on Hispanic is not significant. The result of conducting the hypothesis test, $\beta_2 + \alpha_2 = 0$, that is, there is no difference in response rates between Hispanic email senders to White email senders in states that require a firearm permit, is statistically insignificant (-0.04 pp, SE 0.029, p-value = 0.122). We do not find any significant effects on Black email senders. Lastly, we find marginally significant differences in responses between female and male senders: $\gamma_1 + \delta_1 = 0$

shows that there is a difference in response rates between female email senders to male email senders in states that require firearm permits at the 10 percent level: 0.04 pp, SE 0.0232, p-value 0.0885.

Table 18 Differences in Main Results by Firearm Permit State Laws

	Response Rate (1)	Length (2)	Log(Length) (3)	Sentiment (4)
Permit Required	0.12^{***}	8.41^{**}	0.03	0.14^{***}
	(0.02)	(2.96)	(0.04)	(0.02)
Num. obs.	3762	3762	1759	1758
R^2 (full model)	0.02	0.00	0.00	0.04
Mean No Permit	0.41	33.3	3.98	0.24

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Description of the regressions can be read in Table 16, although we do not include state-level fixed effects in these regressions. ‘Permit Required’ is equal 1 if the US state requires a firearm permit and zero otherwise. There are 21 US states that require firearm permits.

Table 19 Heterogeneity of the Main Results by US States with Firearm Permit Laws

	Response Rate		Length		log(Length)		Sentiment	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black (α_1)	-0.01		1.92		-0.04		-0.00	
	(0.03)		(5.90)		(0.07)		(0.04)	
Hispanic (α_2)	-0.09^{***}		-8.88^{*}		-0.08		-0.04	
	(0.03)		(3.87)		(0.07)		(0.05)	
Black × Permit (β_1)	-0.03		-5.13		0.14		0.04	
	(0.04)		(7.52)		(0.10)		(0.06)	
Hispanic × Permit (β_2)	0.05		0.47		0.01		0.05	
	(0.04)		(5.98)		(0.10)		(0.06)	
Female (γ_1)		0.03		6.21		0.24^{***}		0.04
		(0.02)		(4.24)		(0.06)		(0.04)
Female × Permit (δ_1)		0.01		0.02		-0.13		-0.04
		(0.03)		(5.40)		(0.08)		(0.05)
Num. obs.	3762	3762	3762	3762	1759	1759	1758	1758
R^2 (full model)	0.06	0.06	0.03	0.03	0.06	0.06	0.08	0.08
Omit. Var.	White	Male	White	Male	White	Male	White	Male

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $\cdot p < 0.1$

Description of the regressions can be read in Table 16. Permit is an indicator variable equal to 1 if the state requires a firearm and zero otherwise. Hypothesis tests of the response rates (column 1 and 2) are not significant for Black email senders. Because of the significance on α_2 and the significant difference in response rates for female senders in general, I conduct two hypothesis tests of the results in columns 1 and 2:

1. $\beta_2+\alpha_2 = 0$ shows there is no difference in response rates between Hispanic email senders to white email senders in states that require a firearm permit: -0.04 pp, SE 0.029, and p-value = 0.122.
2. $\gamma_1 + \delta_1 = 0$ shows that there is a difference in response rates between female email senders to male email senders in states that require firearm permits at the 10 percent level: 0.04 pp, SE 0.0232, p-value 0.0885.

Assigned Signals. To understand how the assigned signal on the emails impacts our results, we first regress the main outcomes of interest on just the signals and report the results in Table 20. The main effects for the signals are denoted by "Positive Signal" and "Negative Signal." There are no statistically significant differences at the 5% level between the positive and negative signals relative to the neutral signal. However, emails with a positive signal seem to incur more positive sentiment than the neutral signal (significant at the 10% level), as shown in column (4). The difference between the positive and negative signal in column (4) is 0.0866 (SE of 0.0291) and significant at the 0.2% level. This result suggests that irrespective of the sender's identity, law enforcement responded more positively to emails with a positive signal and less positively to those with a negative sentiment. However, these values are an index, and text that receives values above zero are considered positive in sentiment. Emails with a negative signal still received positive responses on average from law enforcement.

To understand how these results vary depending on the race/ethnicity or gender of the email sender, we interact the race/ethnicity and gender identities with the signals and report the results in Table 21. The main effects for racial/ethnic or gender identity are denoted by "Black," "Hispanic," and "Female." The results for these variables are the differences in the outcomes for the Black, Hispanic, or female identity relative to the white or male identity for emails with the neutral signal. The interaction terms between the main effects give the differences between the positive and negative signals and the different racial and gender identities.

In general, we do not find any differences in the interactions of the identity and the signals. These coefficients, denoted as the "[Identity] × [Signal]", show that the differences in slopes relative to the omitted variable are statistically insignificant. For instance, we do not find that the change in responses between the positive and neutral signal for the white identity is any different than the change in responses between the positive and neutral signal for the Black and the Hispanic identities. In column 2, we do show a difference for the "Female ×

Table 20 Heterogeneity of the Main Results by Email Signal

	Response Rate (1)	Length (2)	log(Length) (3)	Sentiment (4)
Positive Signal (λ_1)	0.03	$6.45^{\cdot}$	0.07	$0.05^{\cdot}$
	(0.02)	(3.82)	(0.05)	(0.03)
Negative Signal (λ_2)	0.00	4.79	0.08	-0.04
	(0.02)	(3.03)	(0.05)	(0.03)
Num. obs.	3762	3762	1759	1758
R^2	0.06	0.03	0.06	0.09
Mean of Neutral Signal	0.46	34.03	3.95	0.32

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Description of the regressions can be read in Table 16. ‘Positive Signal’ is a binary variable equal to one if the email includes a positive signal; ‘Negative Signal’ is a binary variable equal to one if the email includes a negative signal

Hypothesis tests of the differences in outcomes between the positive and negative signal are done by testing the differences in the estimates. The difference is only significant for the sentiment of the text in column (4):

1. Test: $\lambda_1 = \lambda_2$. Results: 0.09, SE 0.0291, and p-value =0.003

Negative Signal” of 8 pp, significant at the 5% level. We interpret this coefficient as the slope between the negative and neutral signal for the male identity is smaller than for the female identity, which can be observed in Figure H.6 Panel (b).

To compare differences in levels (i.e., how the responses compare across the demographic identity, conditional on a specific signal, or how the responses compare across the signal, conditional on a specific identity), we conduct hypothesis tests for the outcomes: response rates (see Tables 23 and 22), word count (see Tables 25 and 24), and email sentiment (see Tables 27 and 26). We also illustrate the predicted outcomes with confidence intervals in Figures H.6, H.7, H.8.

Impact on Response Rates. We report the results of the differences in response rates based on the signal and demographic identity in columns (1) and (2) of Table 21.

Table 21 Heterogeneity of the Main Results by the Assigned Signal and Email Sender Demographics

	Response Rate		Length		log(Length)		Sentiment	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black (α_1)	−0.04		−1.70		0.05		0.06	
	(0.04)		(5.47)		(0.09)		(0.05)	
Hispanic (α_2)	-0.09^{**}		-13.13^{**}		-0.19^{*}		−0.00	
	(0.03)		(4.79)		(0.09)		(0.05)	
Positive (λ_1)	0.01	0.02	−0.25	7.74	−0.00	0.09	$0.08^{\cdot}$	0.03
	(0.03)	(0.03)	(5.09)	(6.51)	(0.08)	(0.07)	(0.05)	(0.04)
Negative (λ_2)	−0.02	−0.04	5.97	−1.68	0.07	0.09	−0.02	−0.05
	(0.03)	(0.03)	(6.57)	(3.85)	(0.09)	(0.07)	(0.05)	(0.04)
Black × Positive (β_1)	0.02		8.14		0.00		−0.11	
	(0.05)		(9.42)		(0.12)		(0.07)	
Hispanic × Positive (β_2)	0.04		$11.89^{\cdot}$		$0.23^{\cdot}$		0.01	
	(0.05)		(6.81)		(0.12)		(0.07)	
Black × Negative (β_3)	0.02		−4.85		−0.05		−0.02	
	(0.05)		(8.07)		(0.12)		(0.07)	
Hispanic × Negative (β_4)	0.04		1.61		0.10		−0.04	
	(0.05)		(7.45)		(0.13)		(0.07)	
Female (γ_1)		0.00		2.83		0.19^{**}		0.01
		(0.03)		(3.89)		(0.07)		(0.04)
Female × Positive (δ_1)		0.02		−2.61		−0.04		0.03
		(0.04)		(7.61)		(0.10)		(0.06)
Female × Negative (δ_2)		0.08^{*}		12.79^{*}		−0.03		0.03
		(0.04)		(5.86)		(0.10)		(0.06)
Num. obs.	3762	3762	3762	3762	1759	1759	1758	1758
R^2 (full model)	0.06	0.06	0.03	0.03	0.06	0.06	0.09	0.09
Omit. Var.	White	Male	White	Male	White	Male	White	Male
Mean Omit. Var.	0.50	0.45	40.48	34.54	4.01	3.92	0.32	0.32

Description of the regressions can be read in Table 16.

Relative to the White or Male Identity. We find that relative to response rates for emails assigned a neutral signal, there are no significant differences in response rates for the negative and positive signal for the white and male identity (see Tables 23 and 22). For the neutral signal, we show that only the Hispanic identity received fewer responses than the white identity (-9 pp, standard error of 3 pp, significant at the 1% level); in column (2), we find no difference between the female and male identity. However, we find that for emails assigned the negative signal, the female identity received far more responses than the male identity (8.42 pp with a standard error of 2.75, significant at the 1% level).

Relative to the OWN Identity. We compare how the response rates differ across the signals for each identity. We find no effects on the racial identity in column (1), as shown in Table 22, suggesting that we do not have enough evidence to reject the hypothesis that law enforcement treats different racial groups differently based on a signal of their lawfulness.

We do find evidence that these results may treat male email senders differently depending on the signal in the email, as shown in Table 23. In fact, there is a -5.7 pp drop in response rates (SE of 2.75 pp, significant at the 5% level) for emails assigned a negative signal relative to the positive signal for the male identity. However, we do not find any significant differences between these signals relative to the neutral signal.

We also find evidence for why the female identity received far more emails than the male identity in our main results (see Table 16). Regardless of whether the assigned signal is positive or negative, the response rates are nearly identical (different by 0.003 pp) and greater than the neutral signal. For the male identity, response rates increase or decrease depending on the assigned signal.

Our results suggest there is no difference in behaviors towards the Black identity. However, when investigating the race/ethnicity by gender effects by signal, we find two counter effects for the Black identity. Conditional on the Black female identity, we find an increase in response rates for the negative signal relative to the neutral or positive signal, although statistically insignificant from zero. However, response rates do significantly differ between the negative and positive signal for the Black Male identity (estimate is -9.5 pp with a standard error of 4.6 pp, significant at the 5% level).

Table 22 Hypothesis Tests of the Heterogeneity in the *Response Rates* by the Email Signal and the Senders' Putative *Race/Ethnicity*

Signal	Test	Estimate	SE	T-stat	P-value
Relative to white email sender					
Black email sender					
Neu. to Neu.	$\alpha_1 = 0$	-0.042	0.034	-1.227	0.220
Pos. to Pos.	$\alpha_1 + \beta_1 = 0$	-0.024	0.034	-0.697	0.486
Neg. to Neg.	$\alpha_1 + \beta_2 = 0$	-0.019	0.034	-0.557	0.577
Hispanic email sender					
Neu. to Neu.	$\alpha_2 = 0$	-0.092	0.033	-2.829	0.005
Pos. to Pos.	$\alpha_2 + \beta_2 = 0$	-0.054	0.034	-1.597	0.110
Neg. to Neg.	$\alpha_2 + \beta_4 = 0$	-0.054	0.034	-1.607	0.108
Relative to OWN Identity					
Black email sender					
Neg. to Pos.	$\beta_3 + \lambda_2 = \beta_1 + \lambda_1$	-0.023	0.034	-0.671	0.503
Pos. to Neu.	$\beta_1 + \lambda_1 = 0$	0.028	0.035	0.818	0.413
Neg. to Neu.	$\beta_3 + \lambda_2 = 0$	0.006	0.034	0.162	0.872
Hispanic email sender					
Neg. to Pos.	$\beta_4 + \lambda_2 = \beta_2 + \lambda_1$	-0.027	0.033	-0.820	0.412
Pos. to Neu.	$\beta_2 + \lambda_1 = 0$	0.049	0.033	1.467	0.142
Neg. to Neu.	$\beta_4 + \lambda_2 = 0$	0.021	0.033	0.644	0.519
White Email Sender					
Neg. to Pos.	$\lambda_2 = \lambda_1$	-0.028	0.034	-0.810	0.418
Pos. to Neu.	$\lambda_1 = 0$	0.010	0.034	0.305	0.760
Neg. to Neu.	$\lambda_2 = 0$	-0.017	0.034	-0.515	0.607

Hypothesis tests from column (2) in Table 21

Table 23 Hypothesis Tests of the Heterogeneity in the *Response Rates* by the Email Signal and the Senders' Putative *Gender*

Signal	Test	Estimate	SE	T-stat	P-value
	Female relative to male email sender				
Neu. to Neu.	$\gamma_1 = 0$	0.003	0.027	0.117	0.907
Pos. to Pos.	$\gamma_1 + \delta_1 = 0$	0.022	0.028	0.804	0.421
Neg. to Neg.	$\gamma_1 + \delta_2 = 0$	0.082	0.027	2.995	0.003
	Relative to OWN Identity				
	Male email sender				
Neg. to Pos.	$\lambda_2 = \lambda_1$	-0.057	0.027	-2.064	0.039
Pos. to Neu.	$\lambda_1 = 0$	0.019	0.028	0.700	0.484
Neg. to Neu.	$\lambda_2 = 0$	-0.037	0.028	-1.345	0.178
	Female email sender				
Neg. to Pos.	$\delta_2 + \lambda_2 = \delta_1 + \lambda_1$	0.003	0.028	0.107	0.915
Pos. to Neu.	$\delta_1 + \lambda_1 = 0$	0.038	0.028	1.390	0.164
Neg. to Neu.	$\delta_2 + \lambda_2 = 0$	0.041	0.027	1.520	0.128

Hypothesis tests from column (2) in Table 21

Impact on Word Count. We report the results of the differences in word count based on the signal and putative identity in columns (5) and (6) of Table 21. Panel (a) and Panel (b) in Figure H.7 illustrate the predicted results, and Tables 23 and 22 report the hypothesis tests.

Relative to the White or Male Identity. Similar to the results in columns (1) and (2) of Table 21, we find that relative to the neutral signal, there are no significant differences in word length for the negative and positive signal for the white or male identity. For the neutral signal in column (5), we find that only the Hispanic identity received shorter emails than the white identity (-18.5% difference, significant at the 5% level); in column (6), we find a significant difference between the female and male identity (19.5% longer emails to female email senders, significant at the 1% level). If we assess the impact of any other signal (positive or negative) relative to the white identity for the same signal, we find no significant effects. Lastly, we find that regardless of the signal, the female email sender received longer emails than the male email sender.

Relative to the OWN Identity. We compare how the word length differs across the signals for each identity. We find no effects on the racial identity in column (5) of Table 21, except for the Hispanic identity. Emails with the positive signal received longer responses from law enforcement than the neutral signal. For gender identity, we do not find any differences in word length when comparing the outcomes for each signal to each other.

Table 24 Hypothesis Tests of the Heterogeneity in the *Email Word Count* by the Email Signal and the Senders' Putative *Race/Ethnicity*

Signal	Test	Estimate	SE	T-stat	P-value
	Relative to white email sender				
	Black email sender				
Neu. to Neu.	$\alpha_1 = 0$	0.053	0.085	0.623	0.533
Pos. to Pos.	$\alpha_1 + \beta_1 = 0$	0.055	0.080	0.683	0.495
Neg. to Neg.	$\alpha_1 + \beta_2 = 0$	0.001	0.086	0.011	0.992
	Hispanic email sender				
Neu. to Neu.	$\alpha_2 = 0$	-0.186	0.084	-2.210	0.027
Pos. to Pos.	$\alpha_2 + \beta_2 = 0$	-0.088	0.091	-0.967	0.334
Neg. to Neg.	$\alpha_2 + \beta_4 = 0$	0.047	0.080	0.587	0.557
	Relative to OWN Identity				
	Black email sender				
Neg. to Pos.	$\beta_3 + \lambda_2 = \beta_1 + \lambda_1$	0.015	0.082	0.182	0.856
Pos. to Neu.	$\beta_1 + \lambda_1 = 0$	-0.001	0.086	-0.012	0.991
Neg. to Neu.	$\beta_3 + \lambda_2 = 0$	0.014	0.084	0.166	0.868
	Hispanic email sender				
Neg. to Pos.	$\beta_4 + \lambda_2 = \beta_2 + \lambda_1$	-0.066	0.087	-0.756	0.449
Pos. to Neu.	$\beta_2 + \lambda_1 = 0$	0.229	0.083	2.761	0.006
Neg. to Neu.	$\beta_4 + \lambda_2 = 0$	0.163	0.089	1.836	0.066
	White Email Sender				
Neg. to Pos.	$\lambda_2 = \lambda_1$	0.069	0.084	0.818	0.414
Pos. to Neu.	$\lambda_1 = 0$	-0.003	0.079	-0.037	0.970
Neg. to Neu.	$\lambda_2 = 0$	0.066	0.087	0.757	0.449

Hypothesis tests from column (5) in Table 21

Table 25 Hypothesis Tests of the Heterogeneity in the *Word Count* by the Email Signal and the Senders' Putative *Gender*

Signal	Test	Estimate	SE	T-stat	P-value
	Female relative to the male email sender				
Neu. to Neu.	$\gamma_1 = 0$	0.190	0.069	2.756	0.006
Neg. to Neg.	$\gamma_1 + \delta_2 = 0$	0.156	0.071	2.188	0.029
Pos. to Pos.	$\gamma_1 + \delta_1 = 0$	0.151	0.067	2.267	0.023
	Relative to OWN Identity				
	Male email sender				
Neg. to Pos.	$\lambda_2 = \lambda_1$	-0.001	0.072	-0.018	0.986
Pos. to Neu.	$\lambda_1 = 0$	0.087	0.073	1.193	0.233
Neg. to Neu.	$\lambda_2 = 0$	0.086	0.073	1.176	0.239
	Female email sender				
Neg. to Pos.	$\delta_2 + \lambda_2 = \delta_1 + \lambda_1$	0.004	0.066	0.061	0.952
Pos. to Neu.	$\delta_1 + \lambda_1 = 0$	0.048	0.062	0.778	0.437
Neg. to Neu.	$\delta_2 + \lambda_2 = 0$	0.052	0.067	0.774	0.439

Hypothesis tests from column (6) in Table 21

Impact on Sentiment. We report the results of the differences in the sentiment of the text based on the signal and putative identity in columns (5) and (6) of Table 21. Tables 23 and 22 report the hypothesis tests. Panel (a) and Panel (b) in Figure H.8 illustrate the predicted results.

Relative to the White or Male Identity. We find no significant differences in the text sentiment of the responses of each racial or gender email sender relative to the white or male email sender.

Relative to the OWN Identity. When we compare the results of across the signal for each identity, we find that both the Hispanic email sender and the white email sender received more negative sentiment in the responses with the negative signal. We find a similar result for the female and male identities. We also find significant differences in the sentiment of the text for the Hispanic male email sender: email sentiment for negative signals is more negative by 0.253 units relative to the positive signal (significant at the 0.01% level). No other race/ethnicity by gender identity has any significant results.

These results suggest that law enforcement may be more positive in how they interact with individuals if the signal in the initial communication is positive than if the signal is negative, but we lack any substantial evidence to conclude that that is the case.

Table 26 Hypothesis Tests of the Heterogeneity in the *Email Sentiment* by the Email Signal and the Senders' Putative *Race/Ethnicity*

Signal	Test	Estimate	SE	T-stat	P-value
Relative to white email sender					
Black email sender					
Neu. to Neu.	$\alpha_1 = 0$	0.064	0.048	1.336	0.181
Pos. to Pos.	$\alpha_1 + \beta_1 = 0$	-0.047	0.047	-0.984	0.325
Neg. to Neg.	$\alpha_1 + \beta_2 = 0$	0.047	0.053	0.890	0.374
Hispanic email sender					
Neu. to Neu.	$\alpha_2 = 0$	-0.003	0.048	-0.065	0.948
Pos. to Pos.	$\alpha_2 + \beta_2 = 0$	0.006	0.048	0.131	0.896
Neg. to Neg.	$\alpha_2 + \beta_4 = 0$	-0.040	0.054	-0.739	0.460
Relative to OWN Identity					
Black email sender					
Neg. to Pos.	$\beta_3 + \lambda_2 = \beta_1 + \lambda_1$	-0.011	0.050	-0.212	0.832
Pos. to Neu.	$\beta_1 + \lambda_1 = 0$	-0.030	0.048	-0.613	0.540
Neg. to Neu.	$\beta_3 + \lambda_2 = 0$	-0.040	0.051	-0.794	0.427
Hispanic email sender					
Neg. to Pos.	$\beta_4 + \lambda_2 = \beta_2 + \lambda_1$	-0.150	0.052	-2.900	0.004
Pos. to Neu.	$\beta_2 + \lambda_1 = 0$	0.090	0.048	1.870	0.062
Neg. to Neu.	$\beta_4 + \lambda_2 = 0$	-0.060	0.052	-1.148	0.251
White Email Sender					
Neg. to Pos.	$\lambda_2 = \lambda_1$	-0.104	0.050	-2.068	0.039
Pos. to Neu.	$\lambda_1 = 0$	0.081	0.047	1.703	0.089
Neg. to Neu.	$\lambda_2 = 0$	-0.023	0.050	-0.458	0.647

Hypothesis tests from column (7) in Table 21

3.7 Discussion & Conclusion

3.7.1 Discussion. We use signals in the emails to explore how responses may differ depending on the information available to law enforcement.

Table 27 Hypothesis Tests of the Heterogeneity in the *Email Sentiment* by the Email Signal and the Senders' Putative *Gender*

Signal	Test	Estimate	SE	T-stat	P-value
	Female relative to male email sender				
Neg. to Neg.	$\gamma_1 + \delta_2 = 0$	0.035	0.044	0.789	0.430
Pos. to Pos.	$\gamma_1 + \delta_1 = 0$	0.036	0.039	0.916	0.360
Neu. to Neu.	$\gamma_1 = 0$	0.008	0.039	0.204	0.839
	Relative to OWN Identity				
	Male email sender				
Neg. to Pos.	$\lambda_2 = \lambda_1$	-0.088	0.042	-2.069	0.039
Pos. to Neu.	$\lambda_1 = 0$	0.033	0.039	0.827	0.408
Neg. to Neu.	$\lambda_2 = 0$	-0.055	0.042	-1.314	0.189
	Female email sender				
Neg. to Pos.	$\delta_2 + \lambda_2 = \delta_1 + \lambda_1$	-0.089	0.041	-2.185	0.029
Pos. to Neu.	$\delta_1 + \lambda_1 = 0$	0.060	0.039	1.535	0.125
Neg. to Neu.	$\delta_2 + \lambda_2 = 0$	-0.028	0.041	-0.688	0.492

Hypothesis tests from column (8) in Table 21

Although we do find some significant differences in average response rates and the content of the responses, the marginal effects of the signals do not seem to have a large effect on outcomes.

Whether the results suggest the differences are due to statistical (Phelps, 1972) or taste-based (Becker, 1971) discrimination remains inconclusive. We can directly test this by conducting hypothesis tests on the interaction between demographic identities and signals. Following Ewens et al. (2014), who showed that if the interaction terms on the negative signals for Black (and in our case Hispanic) identities are positive and if the interaction terms on the positive signals are ambiguous, our results suggest statistical discrimination. If the coefficients on the interaction term between the positive signals and the non-white identities are negative, then the results suggest taste-based discrimination (see Table 21, column 1). However, our estimates and hypothesis tests for these interaction terms on the racial/ethnic identities are inconclusive.

For the effects on gender identity, our study provides puzzling results. We show that response rates are lower for the emails with negative signals relative to the positive signals for the male email sender. Yet for the female email sender, we find that response rates actually increase for the negative signal relative to the neutral signal (the point estimate on the positive signal is also positive), and that the length of the responses from law enforcement, regardless of the signal, are greater than for the male email sender.

It is plausible that the type of discrimination occurring in our study is a combination of several types of behaviors. Law enforcement may statistically discriminate against men and minorities, for instance, but pay more attention to emails and put in more effort in communicating with women.

3.8 Conclusion

This paper suggests that when individuals send inquiries about how to legally purchase a firearm, law enforcement could treat groups of people differently based on signals of their demographic identities and background history. However, it is important to note that our results do not conclusively indicate discrimination in firearm ownership or accessibility. Instead, our results suggest the potential for unfair treatment towards citizens working with law enforcement to purchase a gun. This unfairness raises consumer protection concerns, as it indicates there could be unequal treatment in firearm accessibility, particularly as law enforcement gains more discretion in who can legally purchase firearms.

The email servers law enforcement use may have impacted our results. Response rates for Hispanic email senders are lower on average, potentially due to the email servers flagging those emails as spam. We caught our own email server directly treating two of the email addresses for the Hispanic email sender as spam starting the second week of our experiment (we fixed this and resent the appropriate emails). If the email servers are systematically flagging emails associated with a specific race/ethnicity (or gender) as spam, then our results are not due to law enforcement ignoring emails but a technological problem. Yet, when we evaluate the content of law enforcement's email responses, we find evidence suggesting that law enforcement responded differently to emails assigned to the Hispanic identity. Although technology could impact the results, the results show systematic differences in communications from law enforcement by the putative race/ethnicity and gender of the email senders.

Up to this point in the dissertation, I show in both a laboratory and field experiment that decision-makers discriminate when they observe demographic information in a digital setting. The next chapter pivots to explore how researchers can access more robust data on population-wide internet searches. Although not an experiment, this chapter shows how one can leverage online technology such as Google Trends to better understand the beliefs, interests, or general sentiment of populations across time and geographical space.

CHAPTER IV

ESTIMATING PANEL GOOGLE TRENDS DATA

4.1 Introduction

Google Trends is an online platform that provides data on the intensity of internet searches made on the Google search engine. The data has been widely used in the social sciences to understand how populations in different regions search the internet and the implications of those searches on different outcomes. For example, it has been used to show how racial animus cost President Obama votes in the 2012 elections (Stephens-Davidowitz, 2014). To access the data, one can download the data from the Google Trends website or access more granular data with an API. The data is an index index (i.e., Google does not provide the actual number of searches made for a search term) of the search volumes, constructed in a two-step process that "normalizes [the search volumes] to the time and location of a query."[1] As a result of the two-step process, all Google Trends data range between zero and 100, with 100 corresponding to the time-period and location with the maximum search intensity.

However, there are limitations to the data, specifically accessing panel data (i.e., the index of search volumes that are relative across locations and time) for more than five regions and a single search term. By design, users of Google Trends data cannot compare the raw data from different queries with (1) different search terms, (2) different geographic regions, and (3) different time periods. But given the available Google Trends data, one might ask if it is possible to construct panel data for more than five regions and one search term with the raw data. In this

[1] Google Trends FAQ

paper, I provide an algorithm that estimates panel search volumes for the sub-regions in a geographic region (such as panel variation for provinces in a country).

The algorithm constructs panel data by reverse engineering the indexing process and using only Google Trends data. The two main data-sets used in the algorithm are the search volumes over time (i.e., time series data) of a geographic region and its sub-regions' cross-sectional search volumes. Across the time window in the parent region's time series, I retrieve the sub-regions' cross-sectional search volumes for each time interval (i.e., days, weeks, or months). For example, to estimate weekly panel data across a full year, the algorithm would require fifty two sets of cross-sectional search volumes for each week in the year. To estimate panel data, the algorithm combines the sub-regions' search volumes at each time interval with the parent region's search volumes in order to transform the sub-regions' values so they are relative across the weeks. In this way, the algorithm estimates panel data for as many sub-regions that exist in a geographic location, which number of sub-regions is typically more than five.

A key assumption in the algorithm is that the sub-regions' total search volumes for any search term must add up to the parent region's search volumes. This also means that all the sub-regions' internet searches must add up to the internet searches in the parent region. I make this assumption in order to combine the sub-regions' cross-sectional search volumes with the the parent region's time series data. Without the assumption and the parent region's search volumes, the sub-regions' cross-sectional search volumes across every time interval in a given time window would not contain any cardinal information—it would only show which subregion searched the internet the most at any point in time. In other words, the search volumes would not be a panel data set.

Given this assumption and the mechanics of the algorithm, the algorithm can be summarized in three general steps:

(1) *Gather data.* For a given time window, retrieve Google Trends search volumes over time for a parent region and the sub-regions' cross-sectional search volumes for each time interval (daily, weekly, or monthly).

(2) *Create naive estimates.* Adjust the cross-sectional values to be proportional to the parent region's search volumes in each time interval, so the sub-regions' search volumes are relative to the parent region's maximum search volume (i.e., the search volume in the time interval with a 100).

(3) *Construct adjusted estimates.* Adjust the naive estimates with each subregion's time series data from a different Google Trends query to reflect the temporal variation at the subregion level. These values continue to remain relative to the parent region's maximum search intensity.

To showcase the efficacy of the algorithm, I provide a case-study that estimates panel data for the UK's four sub-regions—Northern Ireland, England, Scotland and Wales—for a single search term "weather" over the 2021 year. The United Kingdom was specifically chosen as it contains fewer than five regions, so I could retrieve the panel data for those regions from Google Trends and test how well the estimates predict the actual Google Trends search volumes. I use OLS regressions to calculate the R^2 from regressing the panel volumes directly queried from Google Trends on the estimated values for each subregion. The R^2s are near 100%: for the naive estimates, the R^2 ranges between 96.31% and 97.61%, and, for the adjusted estimates, the R^2 ranges between 97.90% and 99.58%.

One might wonder if the precision demonstrated by the R^2s is related to the United Kingdom having fewer than five sub-regions. Unfortunately, I cannot answer that question as there is no way to provide empirical evidence of the precision of our algorithm for a region with more than five sub-regions. However, the empirical evidence from the United Kingdom suggests the algorithm nearly perfectly estimates Google Trends search volumes.

Another concern one might have is why there is a range of R^2s with Northern Ireland having the smallest R^2. One possible reason is that there are some inconsistencies in the data retrieved from Google Trends. The ordering of the sub-regions' cross-sectional search volumes is not perfectly aligned with the ordering of the sub-regions' actual panel search volumes in some of the weeks in 2021, i.e., in those weeks, there were times that Wales had the highest value in the cross-sectional data when Northern Ireland had the highest value in the panel data. An explanation for this is that Google uses different samples in constructing the cross-sectional and the panel search volumes. However, Google does not provide how it samples its data, so without information from Google, the reason for the inconsistency remains a conjecture.

The last concern is whether the R^2 are close to 100% due to the search term used, i.e., "weather." Considering the search term "weather" was consistently searched similarly across all the United Kingdom's sub-regions, one might think the high R^2 is related to the underlying characteristics of the search terms. To alleviate this concern, I am continuing to investigate how well the algorithm estimates search volumes for terms with different patterns of search volumes across the sub-regions.

Even with these concerns, I offer the algorithm to the general public (please see Appendix J for the details) to contribute to the ongoing efforts to improve the accessibility of Google Trends data. The results from estimating panel data for the United Kingdom's sub-regions are indicative that the outputs from the algorithm do predict Google Trends panel data for those sub-regions. As far as I know our paper is the first to rigorously demonstrate precision in the mechanics of the panel data estimation process.

The remaining paper is structured as follows. In Section 2 I describe some relevant background information, inclusive of Google's indexing process and the relationship between a region's Google Trends search volume and its sub-regions' search volumes. In Section 3 I discuss how I overcome the limitations in the API. In section 4, I demonstrate the algorithm's precision by testing both the accuracy of the estimated panel data for the United Kingdom's sub-regions for one search term. In section 5, I provide a proof of the algorithm and then offer concluding remarks.

4.2 Background on the Limitations of Accessing Panel Data

This section delves further into the limitations that prevent users from retrieving panel data for more than five regions for a single search term.

The Google Trends Indexing Process. Google Trends offers information on how frequently search terms are entered into Google's search engine, although the quantity of searches made by people are never provided. Instead, search volumes are first aggregated and anonymized before Google Trends indexes the values on a relative scale. The indexing process can be summarized in two steps:

1. *Calculate search intensities.* For a given time period, region and search term, the volume of searches for the search term is divided by the total volume of searches completed for all search terms. An example of how search intensities are constructed is if people in the United Kingdom searched "weather" 500

times and all other words, inclusive of “weather,” 1000 times, then the search intensity for “weather” is 0.5.

2. *Construct the index of search volumes.* The search intensities calculated from Step 1 are scaled by the maximum search intensity observed over the given time frame and geographic regions, rounded to the nearest integer, and multiplied by 100. These values are the observed Google Trends data, where the lower number implies the term was searched less relative to the region and time period with a value of 100.

All Google Trends data range between 0 and 100—a value of 100 represents the location and point in time with the highest search intensity, not the highest searches. This is intentional as it allows the comparison of search patterns from locations with various sized populations and internet usage.[2]

The Google Trends API Region and Search Term Restrictions. Users can download Google Trends data directly from the website or through the Google Trends API. In most usage, user-written packages allow users to engage with the Google Trends API more effectively (e.g., ‘gtrendsR‘ or ‘PyTrends‘). In particular, they allow users to import Google Trends data directly into R or Python.

Users interested in accessing panel data from Google Trends can interact with the API in two fundamental ways. First, for a given region (e.g., the world, a country, a state or province), users can retrieve search volumes over time (i.e., time-series) on how people in the region search with different intensities for up to

[2] Note that Google uses a representative sample of Google searches from their database when indexing the search volumes, which can be further read here. I do not know with certainty how Google samples their data.

five search terms. In this way, both time-series and cross-sectional variation (i.e., across search terms, anyway) is retrieved. Second, for a given search term, users can retrieve time-series and cross-sectional search volumes (i.e., now across regions) on how people in up to five regions search for that search term.

Outside the scope of these two interactions, users cannot directly import panel data from the API. In fact, the packages 'gtrendsR' or 'PyTrends' will automatically quit querying the API and return an error code if a user attempts to retrieve panel data for more than five regions for a single search term or for more than five search terms for a single region. As a result of these restrictions and the indexing process, panel data from Google Trends is constrained, limiting meaningful usage of the data to primarily time-series or cross-sectional studies.

4.3 Overcoming the Limitations of Accessing Panel Data

In the following sections, I provide information on cross-sectional data from Google Trends and how it could be used to estimate panel data.

Cross-sectional Search Volumes: What are They. When users query the Google Trends API for time series or panel data for a geographic region or regions, the API also returns cross-sectional search volumes for the sub-regions in the query's inputted regions, i.e., retrieving search volumes for the United Kingdom over time will include the aggregated search volumes for Northern Ireland, Wales, Scotland, and England.

And like all Google Trends data, the cross-sectional search volumes are a relative index. Unlike Google Trends time series or panel data, these cross-sectional search volumes are first aggregated over the span of the inputted time window before they are indexed in the two-step process outline above. By aggregating the search volumes over the span of the time window, there is only one value between 0 and 100 for each subregion and search term in any sub-regions' cross-sectional data.

There is uncertainty in how the search intensities are calculated in general and specifically for these sub-regions. Google does not offer the exact process, so I do not know if there are any other steps involved. When constructing the algorithm, I made several assumptions about the indexing process, which I highlight in the proof of the algorithm. For now, note that I make the assumption that Google constructs the sub-regions' search intensities from the raw search volumes pulled from their database without any additional steps.[3]

Cross-sectional Search Volumes Over Time. A main attribute of the Google Trends API is that users can request the sub-regions' cross-sectional data for any time interval of interest. If one wants the daily sub-regions' cross-sectional search volumes over a month, the user can query the API for each day in the month. If one wants weekly or monthly cross-sectional search volumes, again the user can independently query the API, although Google limits number of subsequent queries one can make within a 24 hour window.

I illustrate the cross-sectional data one can retrieve from multiple queries over sequential time intervals for a single search term and region in Figure 7. I plot in Figure 7 the Google Trends search volumes for the search term "weather" in the United Kingdom (UK) over the 2021 year with weekly cross-sectional search volumes for the UK's sub-regions—Northern Ireland, Scotland, England, and Wales. The solid line is the UK's search volumes over time and the bars are the

[3] I do not know if there are any other steps in the indexing process (i.e., does Google weight the search intensities when comparing different geographic regions?). I make an assumption that the indexing process outlined in the previous section, given the available information from Google, is how Google indexes the search volumes.

individual cross-sectional search volumes for each subregion, stacked in descending order.

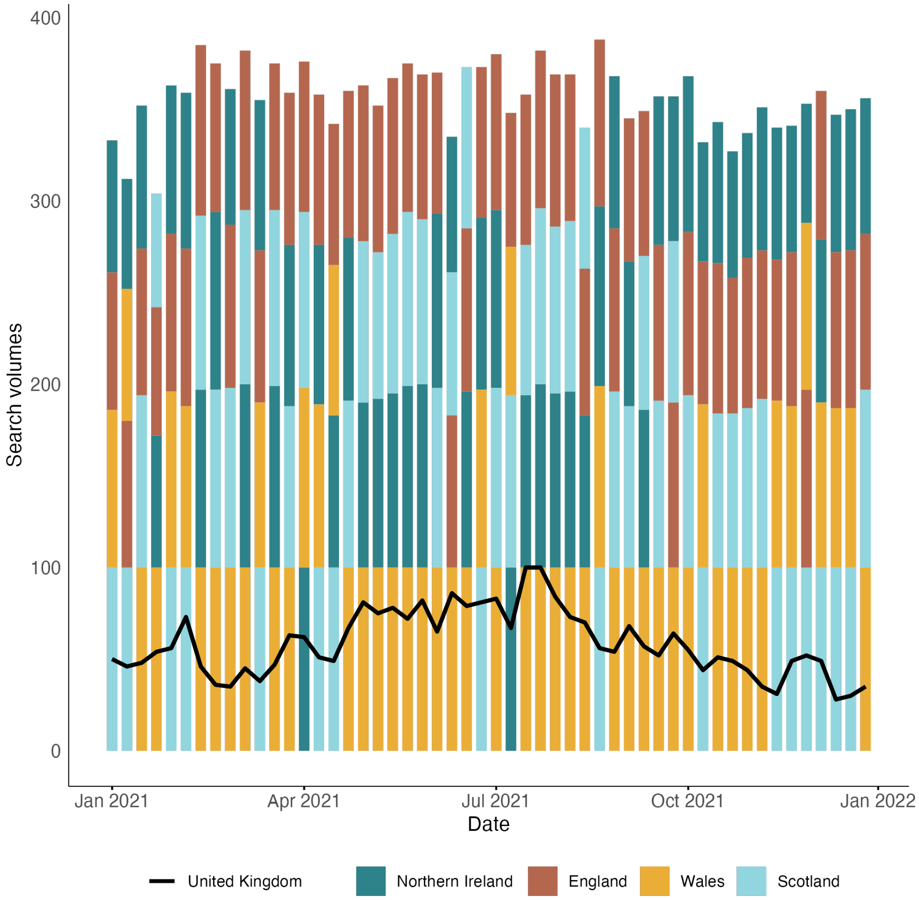
What Figure 7 shows is that the ranking of the sub-regions' search volumes change across the weeks. However, changes in the ranking does not suggest how much each subregion searched for "weather" relative to any other week. Instead, these weekly volumes provide information on the relative position of the search volumes for each subregion at a specific week.

Time Series Search Volumes are necessary to Estimate Panel Data. To extract the cardinal information from sub-regions' cross-sectional search volumes across different time intervals, I make an assumption that the parent region's—the location inputted as a parameter in the query to the Google Trends API—raw search volume is an aggregate of its sub-regions' raw search volumes. For example, if there are only two sub-regions in a parent region (e.g., A and B), then the sum of the total searches for a term in subregion A and subregion B is the total number of searches for that word in the parent region. Similarly, the sum of the total number of searches in each subregion is the total number of searches for all search terms in the parent region.

Given this assumption, Panel A and C of Figure 8 shows that for a given time window, the cross-sectional search volumes retrieved from Google Trends at every time interval can be scaled so the values are linked to the parent region's search volumes.

In Panel A of Figure 8, the sub-regions' cross-sectional values from Figure 8 are proportional to the United Kingdom's maximum search intensity, reflected by the stacked bars lining up perfectly with the United Kingdom's values. These values are labeled as naive estimates, as the construction of them might cause some

Figure 7 Search Volumes Retrieved from the Google Trends API, United Kingdom, 2021



Note: Given a single query of Google Trends API for searches for the word "weather" in the United Kingdom in 2021, the graph is a plot of the weekly search volumes (the line) and the relative search volumes within each week for the four sub-regions of the United Kingdom. The bars are stacked by the values of the subregions' search volumes in ascending order. The first colored bar on the horizontal axis has a value of 100. All other bars stacked on it have smaller values than 100.

loss of the search intensities' temporal information. In order to calculate these naive estimates, I make the assumption that during each time interval, the total internet searches for all terms in each subregion are the same, which is likely not true considering there is population and likely internet usage heterogeneity across the sub-regions.

To introduce temporal information into the naive estimates, I first query Google Trends for each of the sub-regions' time series search volumes separately and plot those values in Panel B of Figure 8. The values for each of these four lines range between zero and 100, meaning that each sub-regions' search volumes are indexed to their own maximum search intensity.[4]

The naive estimates are then readjusted with ratios of the temporal variation from the lines in Panel B of Figure 8 to allow for heterogeneity in internet searches across the sub-regions. In Panel C of Figure 8, I plot those adjusted search volumes, stacking the values again in descending order against the United Kingdom's search volumes. Unlike the naive case, these adjusted search volumes do not necessarily add up to the United Kingdom's search volumes at every time interval, although mathematically, these values are still proportional to the parent region's maximum search intensity.

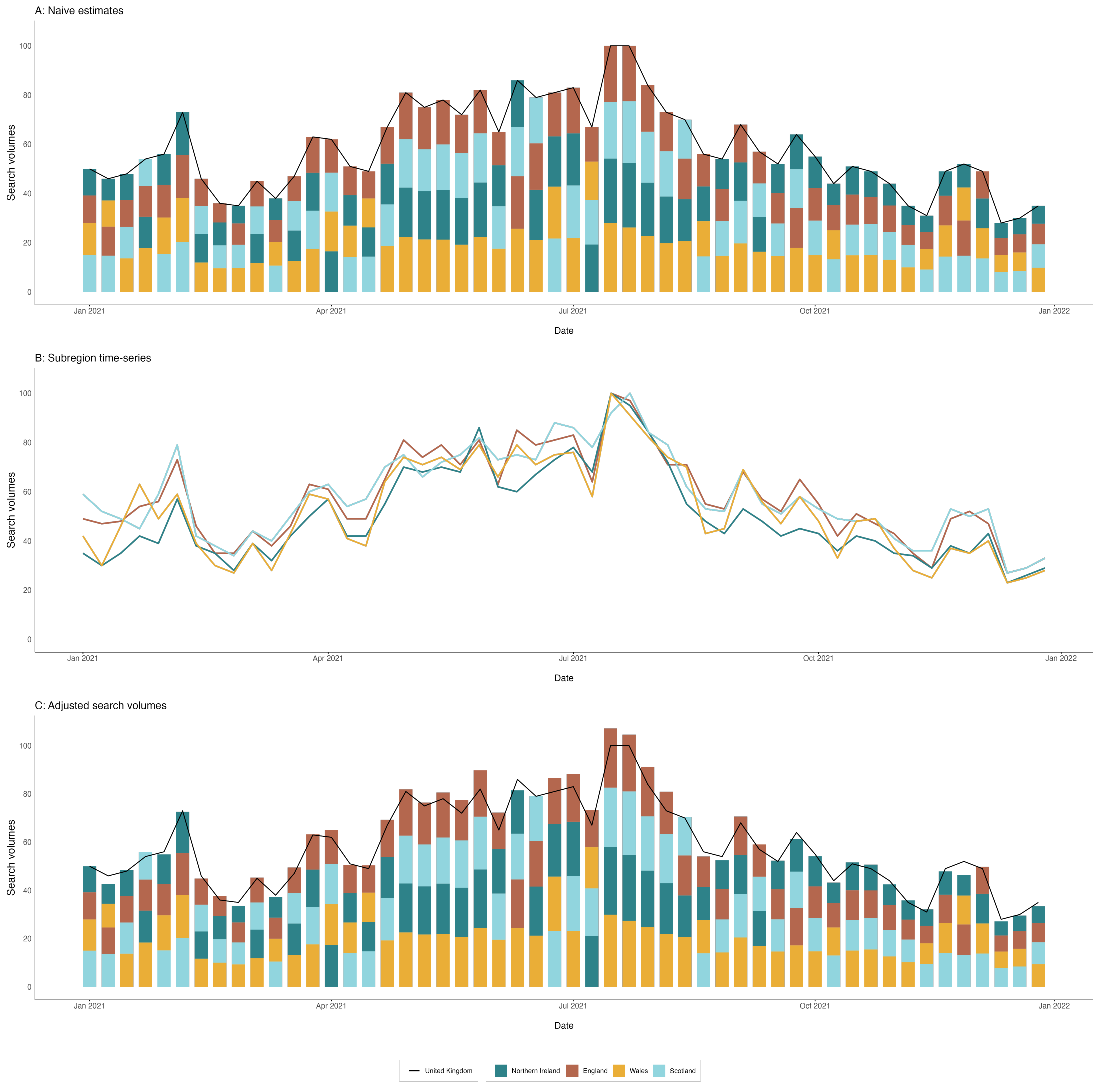
In both Panel A and Panel C of Figure 8, the estimated search volumes do not range between zero and 100. These values represent the proportion of each subregion's search intensity to the parent region—if one of them has a value of 1, then it would imply that that subregion was the only region to search "weather," which in this case did not happen. Considering these values are a proportion, re-indexing the values to the value of the subregion and week with the maximum value mathematically cancels out the parent region's search intensity and leaves relative search volumes for the sub-regions that range between zero and 100.

4.4 Precision of the Algorithm

I now pivot to a discussion about the precision of the naive and adjusted estimates produced from the algorithm.

[4] Note that these lines are not exhibiting panel variation of the regions' search intensities as one cannot make meaningful cross-sectional comparisons across these lines.

Figure 8 Refining the Panel Variation to Reflect the sub-regions' Overall Search Volumes Overtime


A: Naive estimates

B: Subregion time-series

C: Adjusted search volumes

Note: In Panel A, the weekly subregion volumes are scaled to the time-series of the parent region in each week, adjusted to reflect the parent region's total searches relative to the maximum search volume (i.e., 100 on July 17 and 25). In Panel B, the sub-region volumes are the Google Trends search volumes over time—each line is a separate query for the four subregions in the United Kingdom. In Panel C, I adjust the weekly sub-region volumes (in A) to reflect the temporal variation at the subregion level (in B).

OLS regressions. To test the precision of these estimated search volumes, Google Trends panel data for the United Kingdom's sub-regions are compared with the estimated volumes.

I retrieve the panel data by jointly querying the Google Trends API for the word "weather" across Wales, England, Northern Ireland, and Scotland. These values range from zero to 100, with the 100 occurring in Wales on July 18, 2021. The lowest value is 21 occurring in Northern Ireland on December 12, 2021, meaning that the search intensity for "weather" in Northern Ireland was about a fifth of Wales' search intensity for "weather" on July 18. All the other values for all the sub-regions and time periods are similarly defined.

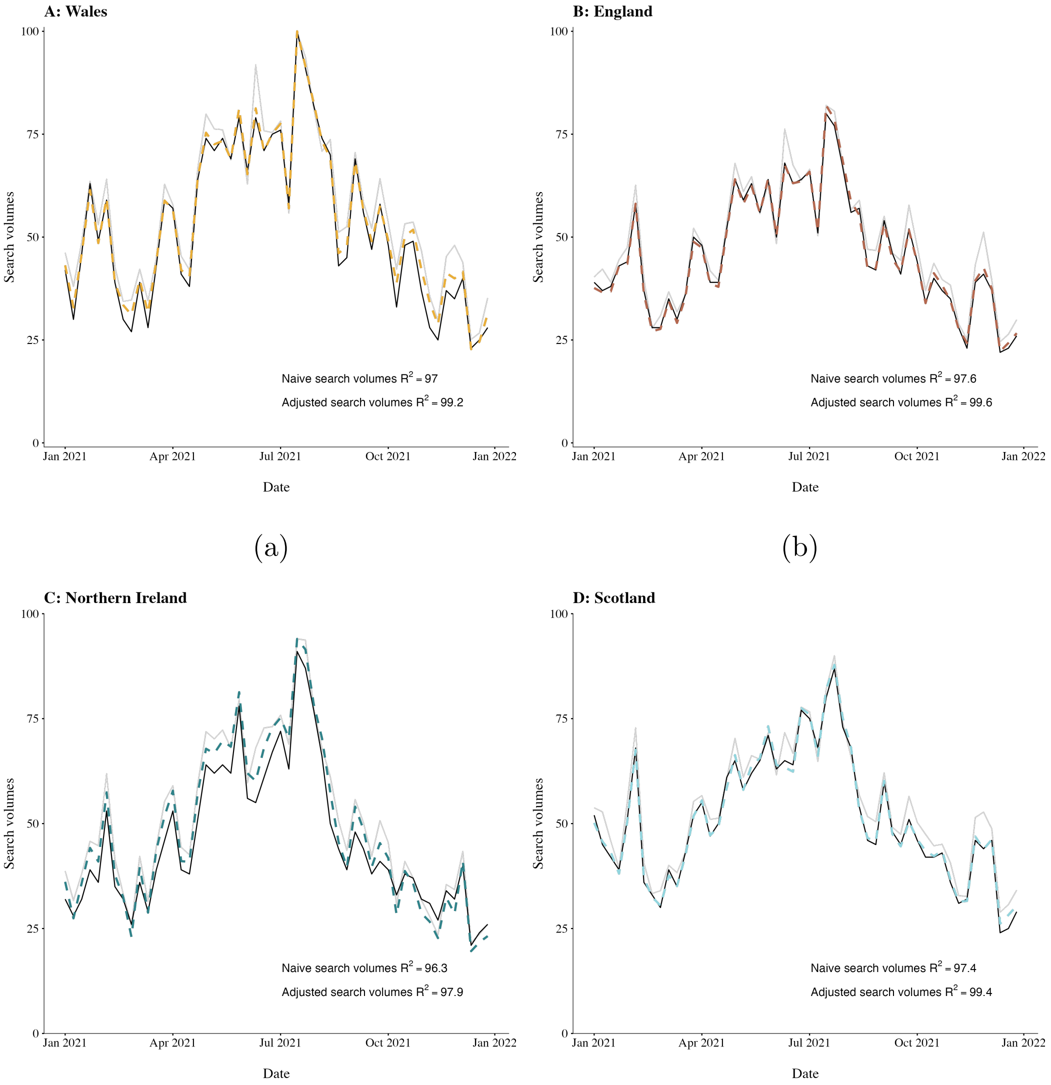
I plot the estimated search volumes with the actual Google Trends search volumes in Panel A through Panel D of Figure 9. The solid black lines are the search volumes retrieved from Google Trends; the colored lines are the adjusted estimates; and, for comparison, the gray lines are the naive estimates.

To test how well the estimated search volumes explain the variation in the panel data retrieved from Google Trends, I use OLS regressions to calculate the R^2. For each subregion, I ran two regressions with the search volumes from Google Trends as the dependent variable: one regression using only the naive estimates as the independent variable and the other regression using only the adjusted estimates as the independent variable.

In each panel, I plot the R^2s from the regressions. The naive estimates explain between 96.31% and 97.61% of the variation in the actual search volumes from Google Trends. The adjusted estimates do a little better, explaining between 97.90% and 99.58% of the variation. In both cases, Northern Ireland has the lowest R^2. Although it seems a 2% improvement with the adjusted estimates is not much, it may indeed be significant for locations with more than five sub-regions.

Discussion of the R^2s. The naive and the adjusted estimates do a great job in explaining the variation of the actual Google Trends search volumes I

Figure 9 Refining the Panel Variation to Reflect the Subregions' Overall Search Volumes Overtime



(a) (b)

(c) (d)

Note: I jointly query the Google Trends API for the word "weather" across the four regions of the United Kingdom (i.e., Wales, England, Northern Ireland, and Scotland). In each panel, I plot these search volumes (the solid black lines) and the adjusted approach that reflects temporal variation in regions' search intensities (colored lines) as shown in panel C of Figure 8. For comparison, I also plot the search volumes estimated by the naive approach (gray lines) as shown in panel A of Figure 8. (This exercise is possible given that there are only five or fewer sub-regions in the United Kingdom. This would not be demonstrable, for example, for a country with more than five sub-regions or states.)

queried (above 95% for each subregion). However, there is some variation across the R^2 that needs to be addressed.

A critique of this case study is that it uses a search term that is too consistently searched across regions and time. One reason why the R^2s are all significantly high but not varied across the naive and adjusted estimated is due to the homogeneity of how people search "weather." Further evaluations of the algorithm's power is to use search terms with more heterogeneity across regions at different time periods and spikes in search volumes. I do not do that for this paper, but will continue to investigate in ongoing works.

Northern Ireland in both cases has the lowest R^2. I cannot say for sure what the reason is for this. However, when I compare the cross-sectional search volumes in Figure 7 with the cross-sectional variation in the panel data I retrieved from Google Trends, there is some discrepancy in the ranking of the sub-regions' search volumes, notably between Wales and Northern Ireland. Because of there does exist some discrepancy between the two data sets, it would affect how well the estimated search volumes explain the variation of the panel search volumes. It does seem to be a small impact, but I cannot say for sure with the available information.

4.4.1 Conclusion. Google collects a vast amount of proprietary data on queries to its search engine, including people's geographic locations and the timing of their searches. This data is publicly available via the Google Trends platform. While Google allows fluid access to Google Trends data on their website or through their Google Trends API, Google does restrict the number of regions and search terms that can be inputted in a single request. By restricting the number of inputs, the API limits the ability to query panel search volumes in a single to five or fewer regions for a single search term(as well as five or fewer search terms for a single region).

I provide an algorithm that reweighs the Google Trends cross-sectional search for sub-regions in a given geographic region for each time period in a time window to be proportional the parent region's Google Trends search volumes. There are two estimates—the naive and the adjusted—that I calculate. The naive estimate is the simple reweighing of the cross-sectional search volumes so the values are a percentage of the parent region's maximum search intensity. The adjusted estimates are those naive values but adjusted to contain temporal information from the subregion's individual time series data. Both estimates are comparable across time and space.

To test the validity of the algorithm, I estimate the search volumes of the United Kingdom's sub-regions and show those estimates explain more that 96% of the variation in the panel data directly retrieved from Google Trends. Unfortunately, I cannot show how well the algorithm estimates search volumes for regions with more than five sub-regions, as that data is not publicly available. However, given the results from the United Kingdom, I conclude with confidence that our algorithm can estimate panel search volumes for all the sub-regions in a geographic location.

However, more work needs to be conducted to identify whether the results are from the search term choice, i.e., "weather". As it stands, the naive estimates and the adjusted estimates are fairly similar in their predictive power, although the adjusted estimates have a higher R^2. It might be the case that search terms with search volume heterogeneity across the sub-regions and time periods results in the adjusted estimates doing better. Considering "weather" is searched consistently, the results from this paper's case-study may be masking the predictive power of the algorithm. The next step moving forward in my attempts to estimate panel data from Google Trends is to identify how well the algorithm predicts search volumes for terms that are not so homogeneously searched.

Even with the need for further evaluations, the algorithm provides promising results for users, researchers, or casual data investigators to feasibly estimate panel data from Google Trends search volumes. Although the best solution would be for Google to provide the data directly to the end user, this paper's algorithm provides a solid solution for users until then.

CHAPTER V

CONCLUDING SUMMARY

This dissertation encompasses three chapters of research. Two chapters use experiments to explore causal relationships, and the last chapter shows how to access more robust and better non-experimental data on population-wide sentiment and interests.

Experiments are powerful tools for studying causal relationships. When experiments are designed well with random assignment into a treatment or a control group, the estimated effects can be considered a direct effect of the intervention. In this dissertation, I use experiments to causally study racial/ethnic and gender discrimination when decision-makers can observe identifying information in digital platforms as decision-makers assess the quality of work (i.e., homework or tests) or decide to interact with individuals. I find evidence that decision-makers can discriminate when they observe identifying information suggesting an individual's race/ethnicity or gender. By design, the experiments also allow the exploration of the mechanisms driving discrimination in the first place and how to leverage the technology to prevent it from occurring. In particular, I show that financial incentives can effectively reduce discrimination. Another effective strategy is to use blind assessments, where identifying information is hidden from the decision-maker.

However, non-experimental data is also essential for research in the social sciences. Non-experimental data can provide observational evidence, even causal evidence, depending on the data and the empirical methodology. Although experiments are the gold standard for assessing causality, they can be expensive and infeasible. Researchers studying topics that rely on population-wide sentiment are often limited to non-experimental data, such as survey data. Survey data can be especially problematic in identifying causal relationships due to selection bias (and people may lie in surveys). Data on population-wide internet searches may be better than survey data for understanding interests and sentiments, as people cannot lie about their internet searches, and selection bias may be minimal in regions with high internet usage. However, data on internet searches can be limited and potentially too costly. To circumvent these limitations, I provide a novel algorithm that leverages the free Google Trends digital platform to access more robust, non-experimental data on the population's internet searches across geographic space and time.

APPENDIX A

CHAPTER 2 - BALANCE TABLES

Table A.1 Balance table of the participant characteristics and the assigned identities

	Unknown	Female				Male			
Demographic	(Intercept)	Asian	Black	Hispanic	White	Asian	Black	Hispanic	White
Panel A: T-test of equality across student identities									
Age	41.070	1.68 (1.16)	2.64 (1.51)*	1.82 (1.51)	1.76 (1.51)	1.62 (1.58)	1.11 (1.65)	1.5 (1.42)	1.96 (1.18)*
Female	0.480	0.03 (0.05)	0.08 (0.06)	-0.04 (0.06)	0.03 (0.06)	0.04 (0.06)	0.03 (0.06)	0.03 (0.06)	0.02 (0.05)
Black	0.090	0 (0.03)	0 (0.03)	0.02 (0.03)	-0.02 (0.03)	0 (0.04)	-0.02 (0.03)	0.02 (0.03)	0 (0.03)
Asian	0.100	-0.01 (0.03)	0 (0.04)	0 (0.03)	-0.01 (0.03)	-0.01 (0.03)	-0.01 (0.04)	-0.01 (0.03)	0 (0.03)
Hispanic	0.070	-0.04 (0.02)**	-0.04 (0.02)	-0.02 (0.03)	-0.04 (0.02)*	-0.04 (0.02)*	-0.06 (0.02)***	-0.04 (0.02)*	-0.04 (0.02)*
Other ethnicity	0.030	-0.02 (0.01)	-0.02 (0.01)*	-0.02 (0.02)	-0.01 (0.02)	-0.03 (0.01)***	0 (0.03)	-0.02 (0.01)	-0.02 (0.01)
Master's or Higher	0.240	-0.01 (0.04)	-0.02 (0.05)	0.02 (0.05)	-0.01 (0.05)	-0.03 (0.05)	0.01 (0.05)	-0.04 (0.05)	-0.01 (0.04)
Bachelor's Degree	0.740	0.02 (0.04)	0.04 (0.05)	-0.01 (0.05)	0.02 (0.05)	0.05 (0.05)	0.01 (0.05)	0.04 (0.05)	0.02 (0.04)
Sales/Service/Construction	0.220	0 (0.04)	-0.01 (0.05)	0 (0.05)	-0.01 (0.05)	-0.09 (0.04)**	0.02 (0.05)	0.07 (0.05)	0 (0.04)
Teacher	0.070	0.02 (0.03)	0.04 (0.04)	0 (0.03)	0.03 (0.04)	0.02 (0.04)	0.08 (0.05)	-0.03 (0.03)	0.02 (0.03)
Unemployed/Retired/Student	0.360	-0.03 (0.04)	-0.02 (0.06)	0.03 (0.06)	-0.03 (0.06)	-0.05 (0.06)	0.01 (0.06)	-0.08 (0.05)	-0.01 (0.04)
Prolific Experience	2045.290	-89.62 (133.5)	-16.34 (180.99)	28.75 (157.56)	-10.57 (184.07)	-96.49 (184.95)	-224.43 (178.39)	-63.95 (167.65)	-13.76 (133.08)

$^{***}p < 0.01$; $^{**}p < 0.05$; $*p < 0.1$

Note: Panel A shows the control mean μ_B, the difference between the treatment and control mean $x - \mu_B$, the standard error in parentheses, and the p-value. Panel B reports the results from a joint test from regressing treatment status on the control variables.

Table A.2 Balance table of the assigned essays and the assigned identities

	Unknown	Female				Male			
Demographic	(Intercept)	Asian	Black	Hispanic	White	Asian	Black	Hispanic	White
				Panel A: T-test of equality across student identities					
1	0.05 (0)***	-0.01 (0.01)	0.03 (0.02)	0 (0.02)	-0.03 (0.01)***	-0.03 (0.02)*	0.06 (0.03)**	-0.01 (0.02)	0 (0.01)
2	0.05 (0)***	0.01 (0.01)	-0.02 (0.02)	0.01 (0.02)	-0.01 (0.02)	-0.01 (0.02)	0.02 (0.03)	-0.01 (0.02)	-0.02 (0.01)***
3	0.05 (0)***	0.03 (0.01)**	-0.01 (0.02)	-0.05 (0)***	-0.01 (0.02)	0 (0.02)	-0.04 (0.01)***	0 (0.02)	-0.01 (0.01)
4	0.05 (0)***	0 (0.01)	-0.01 (0.02)	0.01 (0.02)	0.02 (0.02)	0 (0.02)	-0.03 (0.01)**	-0.01 (0.02)	-0.01 (0.01)
5	0.05 (0)***	-0.01 (0.01)*	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.03 (0.03)	0 (0.02)	0.02 (0.02)	0 (0.01)
6	0.05 (0)***	-0.01 (0.01)	0.01 (0.02)	-0.05 (0.01)***	0 (0.02)	-0.02 (0.02)	-0.02 (0.02)	-0.01 (0.02)	-0.01 (0.01)
7	0.05 (0)***	0.01 (0.01)	-0.02 (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.01 (0.02)	-0.01 (0.02)	0 (0.02)	0.02 (0.01)**
8	0.05 (0)***	0.02 (0.01)*	0.02 (0.02)	0 (0.02)	-0.04 (0.01)***	-0.01 (0.02)	0 (0.02)	0.03 (0.02)	0 (0.01)
9	0.06 (0)***	-0.01 (0.01)	-0.02 (0.02)	0 (0.02)	0 (0.02)	0 (0.02)	-0.01 (0.02)	-0.01 (0.02)	-0.01 (0.01)
10	0.05 (0)***	-0.01 (0.01)	-0.02 (0.02)	0.01 (0.02)	0.02 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0 (0.01)
11	0.05 (0)***	-0.02 (0.01)**	0.03 (0.02)	0.02 (0.02)	0.04 (0.03)	-0.01 (0.02)	0.03 (0.03)	-0.01 (0.02)	0 (0.01)
12	0.05 (0)***	0 (0.01)	0 (0.02)	0 (0.02)	0 (0.02)	0 (0.02)	0.02 (0.02)	0.02 (0.02)	-0.01 (0.01)
13	0.06 (0)***	0 (0.01)	-0.02 (0.02)	-0.02 (0.02)	-0.04 (0.01)***	0.02 (0.03)	-0.03 (0.02)	-0.03 (0.01)***	0 (0.01)
14	0.05 (0)***	0.01 (0.01)	-0.01 (0.02)	0.02 (0.02)	0 (0.02)	0.01 (0.02)	0 (0.02)	0 (0.02)	0.02 (0.01)**
15	0.05 (0)***	-0.01 (0.01)	-0.03 (0.01)*	-0.01 (0.02)	0.03 (0.02)	-0.02 (0.02)	0 (0.02)	0.01 (0.02)	0 (0.01)
16	0.05 (0)***	0 (0.01)	0.01 (0.02)	0.02 (0.02)	-0.01 (0.02)	-0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)*
17	0.04 (0)***	0.01 (0.01)	0 (0.02)	0.02 (0.02)	0.01 (0.02)	0.05 (0.03)*	0 (0.02)	-0.01 (0.02)	0 (0.01)
18	0.05 (0)***	0 (0.01)	0.02 (0.02)	0.01 (0.02)	0.04 (0.03)*	0.01 (0.02)	-0.04 (0.01)***	0.01 (0.02)	0.01 (0.01)
19	0.04 (0)***	0 (0.01)	0.03 (0.02)	0 (0.02)	0.02 (0.02)	0.02 (0.02)	0.01 (0.02)	0 (0.02)	-0.01 (0.01)
20	0.05 (0)***	0 (0.01)	0 (0.02)	0.01 (0.02)	-0.03 (0.02)*	-0.02 (0.02)	0.02 (0.02)	-0.02 (0.02)	-0.01 (0.01)

$^{***}p < 0.01$; $^{**}p < 0.05$; $*p < 0.1$

Note: Panel A shows the control mean μ_B, the difference between the treatment and control mean $x - \mu_B$, the standard error in parentheses, and the p-value. Panel B reports the results from a joint test from regressing treatment status on the control variables.

APPENDIX B

CHAPTER 2 - EFFORT

Table B.1 Differences in Effort and Accuracy Incentives by Students' Putative Gender Identities

	Time (sec)	Information Seeking	Clicks
Female (γ_1)	0.08	−0.00	0.12
	(0.10)	(0.05)	(0.11)
Male (γ_2)	0.10	0.05	0.15
	(0.10)	(0.05)	(0.11)
Incentive	0.07	0.09	0.18
	(0.10)	(0.05)	(0.11)
Female × Incentive (δ_1)	−0.06	−0.03	−0.17
	(0.15)	(0.08)	(0.18)
Male × Incentive (δ_2)	−0.11	−0.10	−0.20
	(0.15)	(0.08)	(0.18)
Num. obs.	3977	3977	3977
R^2 (full model)	0.08	0.02	0.04

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.2 Differences in Effort and Accuracy Incentives by Students' Putative Racial Identities

	Time (sec)	Information Seeking	Clicks
Asian (α_1)	0.03	0.07	2.37
	(0.11)	(0.07)	(1.60)
Black (α_2)	0.19	-0.00	1.33
	(0.12)	(0.07)	(1.07)
Hispanic (α_3)	0.08	0.02	0.86
	(0.13)	(0.07)	(0.72)
White (α_4)	0.09	0.02	1.65
	(0.10)	(0.05)	(1.18)
Incentive	0.07	0.09	0.88
	(0.10)	(0.05)	(0.59)
Asian × Incentive (β_1)	-0.04	-0.13	-2.13
	(0.18)	(0.11)	(1.95)
Black × Incentive (β_2)	-0.11	0.01	-1.51
	(0.17)	(0.11)	(1.70)
Hispanic × Incentive (β_3)	-0.12	-0.14	-1.51
	(0.20)	(0.11)	(1.36)
White × Incentive (β_4)	-0.08	-0.06	-1.86
	(0.15)	(0.08)	(1.52)
Num. obs.	3977	3977	3977
R^2 (full model)	0.08	0.02	0.03

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.3 Hypothesis Tests of Differences in Effort and Accuracy Incentives

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
	Differences in ***Time***				
Asian	$\alpha_1 = \alpha_4$	-0.06	0.09	-0.69	0.49
Black	$\alpha_2 = \alpha_4$	0.10	0.10	0.97	0.33
Hispanic	$\alpha_3 = \alpha_4$	-0.01	0.10	-0.11	0.91
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	-0.02	0.09	-0.25	0.80
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	0.07	0.08	0.91	0.36
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	-0.05	0.10	-0.49	0.63
	Differences in ***Information Seeking***				
Asian	$\alpha_1 = \alpha_4$	0.05	0.06	0.85	0.39
Black	$\alpha_2 = \alpha_4$	-0.02	0.06	-0.33	0.74
Hispanic	$\alpha_3 = \alpha_4$	0.00	0.06	0.05	0.96
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	-0.02	0.06	-0.39	0.70
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	0.05	0.06	0.85	0.40
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	-0.07	0.06	-1.30	0.19
	Differences in ***Clicks per Page***				
Asian	$\alpha_1 = \alpha_4$	0.72	1.36	0.53	0.60
Black	$\alpha_2 = \alpha_4$	-0.32	1.14	-0.28	0.78
Hispanic	$\alpha_3 = \alpha_4$	-0.79	0.79	-1.00	0.31
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	0.45	0.71	0.63	0.53
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	0.03	0.84	0.03	0.98
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	-0.44	0.63	-0.70	0.48
	Differences in ***Time***				
Female	$\gamma_1 = \gamma_2$	-0.03	0.07	-0.39	0.70
Female	$\gamma_1 + \delta_1 = \gamma_2 + \delta_2$	0.03	0.06	0.52	0.61
	Differences in ***Information Seeking***				
Female	$\gamma_1 = \gamma_2$	-0.05	0.04	-1.40	0.16
Female	$\gamma_1 + \delta_1 = \gamma_2 + \delta_2$	0.01	0.04	0.38	0.71
	Differences in ***Clicks per Page***				
Female	$\gamma_1 = \gamma_2$	-0.03	0.07	-0.44	0.66
Female	$\gamma_1 + \delta_1 = \gamma_2 + \delta_2$	0.00	0.07	0.01	0.99

Table B.4 Differences in Effort and Order Effects by Students' Putative Racial Identities

	Time (sec)	Information Seeking	Clicks
Asian (α_1)	-0.09	0.01	1.25
	(0.11)	(0.06)	(1.21)
Black (α_2)	0.04	-0.03	1.26
	(0.10)	(0.06)	(1.33)
Hispanic (α_3)	-0.03	-0.10	-0.24
	(0.11)	(0.06)	(0.92)
White (α_4)	0.06	-0.03	0.74
	(0.06)	(0.03)	(0.72)
Order	-0.93^{***}	-0.61^{***}	-2.15^{***}
	(0.04)	(0.03)	(0.32)
Asian × Order (β_1)	0.15	-0.04	-0.05
	(0.16)	(0.09)	(1.59)
Black × Order (β_2)	0.15	0.05	-1.63
	(0.15)	(0.09)	(1.77)
Hispanic × Order (β_3)	0.10	0.10	0.62
	(0.14)	(0.09)	(1.09)
White × Order (β_4)	-0.01	0.05	-0.08
	(0.08)	(0.05)	(0.58)
Num. obs.	3977	3977	3977
R^2 (full model)	0.20	0.17	0.04

Table B.5 Differences in Effort and Order Effects by Students' Putative Gender Identities

	Time (sec)	Information Seeking	Clicks
Female (γ_1)	0.07	-0.05	-0.03
	(0.07)	(0.03)	(0.09)
Male (γ_2)	-0.02	-0.02	0.03
	(0.07)	(0.04)	(0.09)
Order	-0.93^{***}	-0.61^{***}	-0.33^{***}
	(0.04)	(0.03)	(0.04)
Female × Order (δ_1)	-0.03	0.06	0.12
	(0.08)	(0.05)	(0.09)
Male × Order (δ_2)	0.12	0.03	0.03
	(0.08)	(0.05)	(0.10)
Num. obs.	3977	3977	3977
R^2 (full model)	0.20	0.17	0.05

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.6 Hypothesis Tests of Differences in Effort and Order Effects

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
	*Differences in **Time***				
Asian	$\alpha_1 = \alpha_4$	-0.15	0.10	-1.42	0.16
Black	$\alpha_2 = \alpha_4$	-0.02	0.10	-0.21	0.83
Hispanic	$\alpha_3 = \alpha_4$	-0.09	0.10	-0.91	0.36
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	0.02	0.10	0.19	0.85
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	0.14	0.09	1.47	0.14
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	0.02	0.09	0.19	0.85
	*Differences in **Information Seeking***				
Asian	$\alpha_1 = \alpha_4$	0.04	0.06	0.74	0.46
Black	$\alpha_2 = \alpha_4$	-0.00	0.05	-0.07	0.94
Hispanic	$\alpha_4 = \alpha_1$	-0.07	0.06	-1.24	0.22
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	-0.05	0.06	-0.82	0.41
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	-0.00	0.06	-0.05	0.96
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	-0.01	0.06	-0.22	0.82
	*Differences in **Clicks per Page***				
Asian	$\alpha_1 = \alpha_4$	0.52	1.51	0.34	0.73
Black	$\alpha_2 = \alpha_4$	0.52	1.49	0.35	0.73
Hispanic	$\alpha_4 = \alpha_1$	-0.98	0.90	-1.09	0.27
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	0.54	1.28	0.42	0.67
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	-1.04	0.76	-1.36	0.18
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	-0.29	0.67	-0.42	0.67
	*Differences in **Time***				
Female	$\gamma_1 = \gamma_2$	0.08	0.07	1.30	0.19
Female	$\gamma_1 + \delta_1 = \gamma_2 + \delta_2$	-0.07	0.06	-1.06	0.29
	*Differences in **Information Seeking***				
Female	$\gamma_1 = \gamma_2$	-0.03	0.04	-0.78	0.44
Female	$\gamma_1 + \delta_1 = \gamma_2 + \delta_2$	-0.00	0.04	-0.12	0.91
	*Differences in **Clicks per Page***				
Female	$\gamma_1 = \gamma_2$	-0.06	0.08	-0.74	0.46
Female	$\gamma_1 + \delta_1 = \gamma_2 + \delta_2$	0.03	0.07	0.49	0.63

Table B.7 Triple Difference of the Impact of Accuracy Incentives and Order Effects on Effort by Students' Putative Racial Identities

	Time (sec)	Information Seeking	Clicks
Asian (α_1)	-0.27	0.04	0.21
	(0.20)	(0.10)	(0.22)
Black (α_2)	0.18	-0.05	0.04
	(0.21)	(0.11)	(0.23)
Hispanic (α_3)	-0.03	-0.01	0.06
	(0.19)	(0.10)	(0.22)
White (α_4)	0.03	-0.04	0.07
	(0.11)	(0.06)	(0.14)
Order	-0.94^{***}	-0.65^{***}	-0.40^{***}
	(0.08)	(0.05)	(0.07)
Asian × Order (β_1)	0.60^{*}	0.06	-0.15
	(0.28)	(0.16)	(0.31)
Black × Order (β_2)	-0.03	0.07	0.17
	(0.29)	(0.17)	(0.35)
Hispanic × Order (β_3)	0.18	0.04	0.08
	(0.26)	(0.15)	(0.30)
White × Order (β_4)	0.14	0.13	0.17
	(0.14)	(0.08)	(0.12)
Incentive	0.06	0.05	0.12
	(0.12)	(0.06)	(0.14)
Asian × Incentive (δ_1)	0.36	-0.05	-0.27
	(0.30)	(0.15)	(0.33)
Black × Incentive (δ_2)	-0.28	0.04	-0.12
	(0.29)	(0.16)	(0.37)
Hispanic × Incentive (δ_3)	0.00	-0.18	-0.23
	(0.31)	(0.17)	(0.37)
White × Incentive (δ_4)	0.06	0.02	-0.10
	(0.18)	(0.10)	(0.23)
Order × Incentive	0.02	0.08	0.13
	(0.12)	(0.08)	(0.11)
Asian × Order × Incentive (ζ_1)	-0.87^{*}	-0.21	0.43
	(0.42)	(0.24)	(0.45)
Black × Order × Incentive (ζ_2)	0.36	-0.04	-0.14
	(0.42)	(0.27)	(0.54)
Hispanic ×Order × Incentive (ζ_3)	-0.16	0.13	0.09
	(0.43)	(0.27)	(0.51)
White × Order × Incentive (ζ_4)	-0.29	-0.17	-0.21
	(0.21)	(0.13)	(0.20)
Num. obs.	3977	3977	3977
R^2 (full model)	0.20	0.17	0.05

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$
Note: Hypothesis tests reported in Tables B.8, B.9, and B.10.

Table B.8 Hypothesis Tests of Accuracy Incentives and Essay Order on the Differences in **Time Grading** in the Non-Blind Grading Environment

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
	Differences relative to white students				
	Panel A: Impact of Low Incentives				
	First Essay				
Asian	$\alpha_1 = \alpha_4$	-0.300	0.182	-1.652	0.098
Black	$\alpha_2 = \alpha_4$	0.154	0.187	0.825	0.409
Hispanic	$\alpha_3 = \alpha_4$	-0.059	0.174	-0.337	0.736
	Last Essay				
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	0.164	0.156	1.048	0.295
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	-0.018	0.162	-0.114	0.909
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	-0.018	0.170	-0.108	0.914
	Panel B: Impact of Large Incentives				
	First Essay				
Asian	$\alpha_1 + \delta_1 = \alpha_4 + \delta_4$	-0.004	0.157	-0.028	0.977
Black	$\alpha_2 + \delta_2 = \alpha_4 + \delta_4$	-0.186	0.141	-1.314	0.189
Hispanic	$\alpha_3 + \delta_3 = \alpha_4 + \delta_4$	-0.117	0.177	-0.663	0.507
	Last Essay				
Asian	$\alpha_1 + \beta_1 + \delta_1 + \zeta_1 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	-0.125	0.160	-0.777	0.437
Black	$\alpha_2 + \beta_2 + \delta_2 + \zeta_2 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	0.285	0.144	1.971	0.049
Hispanic	$\alpha_3 + \beta_3 + \delta_3 + \zeta_3 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	0.051	0.153	0.334	0.739

Note: This table reports the hypothesis tests from column 1 of Table B.8.

Table B.9 Hypothesis Tests of Accuracy Incentives and Essay Order on the Differences in **Information Seeking** in the Non-Blind Grading Environment

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
	Differences relative to white students				
	Panel A: Impact of Low Incentives				
	First Essay				
Asian	$\alpha_1 = \alpha_4$	0.080	0.098	0.820	0.412
Black	$\alpha_2 = \alpha_4$	-0.009	0.100	-0.086	0.931
Hispanic	$\alpha_3 = \alpha_4$	0.031	0.095	0.328	0.743
	Last Essay				
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	0.008	0.101	0.084	0.933
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	-0.072	0.103	-0.697	0.486
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	-0.060	0.099	-0.607	0.544
	Panel B: Impact of Large Incentives				
	First Essay				
Asian	$\alpha_1 + \delta_1 = \alpha_4 + \delta_4$	0.008	0.093	0.091	0.928
Black	$\alpha_2 + \delta_2 = \alpha_4 + \delta_4$	0.003	0.090	0.033	0.973
Hispanic	$\alpha_3 + \delta_3 = \alpha_4 + \delta_4$	-0.172	0.099	-1.730	0.084
	Last Essay				
Asian	$\alpha_1 + \beta_1 + \delta_1 + \zeta_1 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	-0.106	0.091	-1.166	0.243
Black	$\alpha_2 + \beta_2 + \delta_2 + \zeta_2 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	0.063	0.108	0.581	0.561
Hispanic	$\alpha_3 + \beta_3 + \delta_3 + \zeta_3 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	0.035	0.099	0.356	0.722

Note: This table reports the hypothesis tests from column 2 of Table B.8.

Table B.10 Hypothesis Tests of Accuracy Incentives and Essay Order on the Differences in **Clicks per Page** in the Non-Blind Grading Environment

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
	Differences relative to white students				
	Panel A: Impact of Low Incentives				
	First Essay				
Asian	$\alpha_1 = \alpha_4$	0.143	0.236	0.606	0.544
Black	$\alpha_2 = \alpha_4$	-0.028	0.231	-0.122	0.903
Hispanic	$\alpha_3 = \alpha_4$	-0.006	0.207	-0.031	0.976
	Last Essay				
Asian	$\alpha_1 + \beta_1 = \alpha_4 + \beta_4$	-0.184	0.201	-0.912	0.362
Black	$\alpha_2 + \beta_2 = \alpha_4 + \beta_4$	-0.028	0.213	-0.130	0.896
Hispanic	$\alpha_3 + \beta_3 = \alpha_4 + \beta_4$	-0.099	0.174	-0.572	0.567
	Panel B: Impact of Large Incentives				
	First Essay				
Asian	$\alpha_1 + \delta_1 = \alpha_4 + \delta_4$	-0.027	0.178	-0.151	0.880
Black	$\alpha_2 + \delta_2 = \alpha_4 + \delta_4$	-0.047	0.215	-0.219	0.827
Hispanic	$\alpha_3 + \delta_3 = \alpha_4 + \delta_4$	-0.139	0.214	-0.648	0.517
	Last Essay				
Asian	$\alpha_1 + \beta_1 + \delta_1 + \zeta_1 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	0.293	0.168	1.743	0.081
Black	$\alpha_2 + \beta_2 + \delta_2 + \zeta_2 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	0.021	0.165	0.127	0.899
Hispanic	$\alpha_3 + \beta_3 + \delta_3 + \zeta_3 = \alpha_4 + \beta_4 + \delta_4 + \zeta_4$	0.072	0.181	0.397	0.691

Note: This table reports the hypothesis tests from column 3 of Table B.8.

APPENDIX C

CHAPTER 2 - LEARNING: ORDER EFFECTS AND ACCURACY INCENTIVES

Table C.1 reports regressions from evaluating differences in grades given to the racial identities in the non-blind grading environment on the first four and last four essays. The dependent variable are the raw scores. The results in column 1 come from pooling all the data. Column 2 and 3 are based on the accuracy incentive assigned to the participants. The data in column 2 are participants assigned an accuracy incentive *less* than or equal to 25 cents. In column 3, the dataset only includes the participants who are assigned an accuracy incentive *greater* than or equal to 25 cents.[1] I label column 2 as "Low Attention" and column 3 as "High Attention" as the accuracy incentives are used to proxy attention in grading.

Column 1 reports no significant differences in any of the estimates. This is not surprising given the null results from Tables 8 and 9. In column 2, essays assigned the white identity receives better grades than the unknown identity by 0.27 points. Given that the average grade given to the unknown identity on the first four essays is 2.85 points, the white identity received a score about 9% greater than the unknown identity.

To compare the differences in scores on the first and last four essays in the non-blind grading environment, I conduct hypothesis tests as reported in Table C.2. Panel A reports the results from column (1); panel B reports the results from column 2; and panel C reports results from panel C. In panel A, there are

[1] I use 25 cents as the threshold and include participants assigned the median value in both data sets in order to keep the power similar between the regressions.

no differences in grades between the students' racial identities on the first or last four essays, as expected.

The hypothesis tests in panel B show that on the first four essays, the Black students receive a lower grade by 0.34 points than the white identity when participants are assigned smaller accuracy incentives. Yet, on the last four essays, the Black-white gap decreases to 0.06 points, statistically insignificant (p-value of 0.61). Additionally, the Asian-white gap became significant on the last four essays by 0.21 points (p-value of 0.05), primarily due to the average grades for white students dropping on the last four essays by 0.12 points and the average grades for Asian students increasing by 0.11 points.

I also investigate the impact of accuracy incentives and order effects by the students' putative gender and find no effect.

Table C.1 Grade Differences Between the First and Last Four Essays

	Pooled (1)	Low Attention (2)	High Attention (3)
	First Four Essays: Relative to the Unknown Identity		
Asian (α_1)	0.063	0.264	-0.093
	(0.116)	(0.165)	(0.134)
Black (α_2)	-0.020	-0.066	0.012
	(0.108)	(0.143)	(0.135)
Hispanic (α_3)	0.005	0.060	-0.054
	(0.117)	(0.139)	(0.162)
White (α_4)	0.079	0.273^{**}	-0.039
	(0.073)	(0.100)	(0.086)
	Last Four Essays: Main Effect on the Unknown Identity		
Last Four	-0.042	0.006	-0.108^{*}
	(0.045)	(0.063)	(0.052)
	Last Four Essays: Relative to Unknown Identity		
Asian × Last Four (β_1)	0.090	0.092	0.121
	(0.131)	(0.185)	(0.154)
Black × Last Four (β_2)	0.059	0.155	0.030
	(0.124)	(0.166)	(0.160)
Hispanic × Last Four (β_3)	0.054	0.129	0.066
	(0.131)	(0.162)	(0.176)
White × Last Four (β_4)	-0.011	-0.125	0.081
	(0.073)	(0.098)	(0.094)
Num. obs.	3977	2266	2497
R^2 (full model)	0.51	0.51	0.54
Accuracy Incentive	Pooled	Incentive $\leq$ \$0.25	Incentive $\geq$ \$0.25
Benchmark Score	N	N	N
Participant Control	Y	Y	Y
Essay FE	Y	Y	Y

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Note: This table reports the differences in grades on the first and last four essays of the putative racial identities relative to the unknown identity in the blind grading environment. The data in column (1) pools all data; column (2) includes observations from participants who are assigned the median incentive amount (25 cents) and less; column (3) includes observations from participants who are assigned the median incentive amount and more. 'Last Four' is an indicator variable equal to one if the graded essays are the last four seen and zero if the essays the first four essays seen by the participants. Hypothesis tests reported in Table C.2

Table C.2 Hypothesis Tests of Attention and Essay order

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
	Panel A: Pooled Results				
	Relative to White: First Four Essays				
Asian	$\alpha_1 = \alpha_4$	-0.02	0.11	-0.16	0.88
Black	$\alpha_2 = \alpha_4$	-0.10	0.11	-0.93	0.35
Hispanic	$\alpha_3 = \alpha_4$	-0.07	0.11	-0.65	0.51
	Relative to White: Last Four Essays				
Asian	$\beta_1 + \alpha_1 = \beta_4 + \alpha_4$	0.08	0.08	1.03	0.30
Black	$\beta_2 + \alpha_2 = \beta_4 + \alpha_4$	-0.03	0.09	-0.34	0.74
Hispanic	$\beta_3 + \alpha_3 = \beta_4 + \alpha_4$	-0.01	0.08	-0.11	0.91
	Panel B: Low Attention				
	Relative to White: First Four Essays				
Asian	$\alpha_1 = \alpha_4$	-0.01	0.15	-0.06	0.95
Black	$\alpha_2 = \alpha_4$	-0.34	0.14	-2.36	0.02
Hispanic	$\alpha_3 = \alpha_4$	-0.21	0.14	-1.58	0.12
	Relative to White: Last Four Essays				
Asian	$\beta_1 + \alpha_1 = \beta_4 + \alpha_4$	0.21	0.10	1.99	0.05
Black	$\beta_2 + \alpha_2 = \beta_4 + \alpha_4$	-0.06	0.11	-0.52	0.61
Hispanic	$\beta_3 + \alpha_3 = \beta_4 + \alpha_4$	0.04	0.11	0.36	0.72
	Panel C: High Attention				
	Relative to White: First Four Essays				
Asian	$\alpha_1 = \alpha_4$	-0.05	0.13	-0.42	0.67
Black	$\alpha_2 = \alpha_4$	0.05	0.13	0.39	0.70
Hispanic	$\alpha_3 = \alpha_4$	-0.01	0.16	-0.09	0.93
	Relative to White: Last Four Essays				
Asian	$\beta_1 + \alpha_1 = \beta_4 + \alpha_4$	-0.01	0.10	-0.14	0.89
Black	$\beta_2 + \alpha_2 = \beta_4 + \alpha_4$	-0.00	0.11	-0.01	1.00
Hispanic	$\beta_3 + \alpha_3 = \beta_4 + \alpha_4$	-0.03	0.10	-0.29	0.77

Note: This table reports the hypothesis tests from Table C.1. ‘Low Attention’ is defined as being assigned incentives lower or equal to 25 cents. ‘High Attention’ is defined as being assigned incentives greater or equal to 25 cents. Standard errors are robust and clustered at the participant level.

Table C.3 Gender differences in grades between first and last four essays and attention levels

	Pooled	Low Attention	High Attention
	(1)	(2)	(3)
First Four Essays: Relative to the Unknown Identity			
Male (γ_1)	0.09	0.23^{*}	-0.02
	(0.07)	(0.10)	(0.08)
Female (γ_2)	0.02	0.16	-0.06
	(0.07)	(0.10)	(0.09)
Last Four Essays:Main Effect on the Unknown Identity			
Last Four	-0.04	0.01	-0.11^{*}
	(0.05)	(0.06)	(0.05)
Last Four Essays: Relative to the Unknown Identity			
Male × Last Four (δ_1)	-0.00	-0.06	0.08
	(0.08)	(0.11)	(0.09)
Female × Last Four (δ_2)	0.04	0.01	0.07
	(0.08)	(0.10)	(0.10)
Num. obs.	3977	2266	2497
R^2 (full model)	0.51	0.51	0.54

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table C.4 Hypothesis tests of the gender effects: Attention and essay order

Comparison	Hypothesis	Estimate	SE	t-stat	p.value
Panel A: Pooled					
Relative to Male: First Four Essays					
Female	$\gamma_2 = \gamma_1$	-0.06	0.07	-0.86	0.39
Relative to Male: Last Four Essays					
Female	$\gamma_2 + \delta_2 = \gamma_1 + \delta_1$	-0.02	0.05	-0.34	0.73
Panel B: Low Attention					
Relative to Male: First Four Essays					
Female	$\gamma_2 = \gamma_1$	-0.07	0.10	-0.76	0.45
Relative to Male: Last Four Essays					
Female	$\gamma_2 + \delta_2 = \gamma_1 + \delta_1$	-0.01	0.07	-0.13	0.90
Panel B: High Attention					
Relative to Male: First Four Essays					
Female	$\gamma_2 = \gamma_1$	-0.04	0.09	-0.40	0.69
Relative to Male: Last Four Essays					
Female	$\gamma_2 + \delta_2 = \gamma_1 + \delta_1$	-0.04	0.07	-0.64	0.52

Note: This table reports the hypothesis tests from Table C.3. ‘Low Attention’ is defined as being assigned incentives lower or equal to 25 cents. ‘High Attention’ is defined as being assigned incentives greater or equal to 25 cents. Standard errors are robust and clustered at the participant level.

APPENDIX D

CHAPTER 2 - FIGURES

Figure D.1 Effect of Accuracy Incentives on Grades

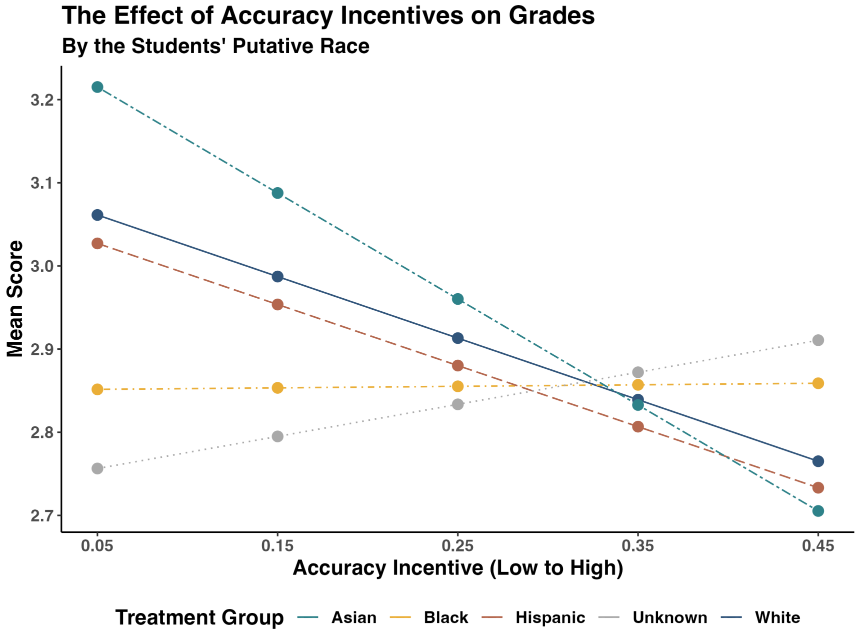


Figure D.2 Average Effect of Accuracy Incentives on Grades

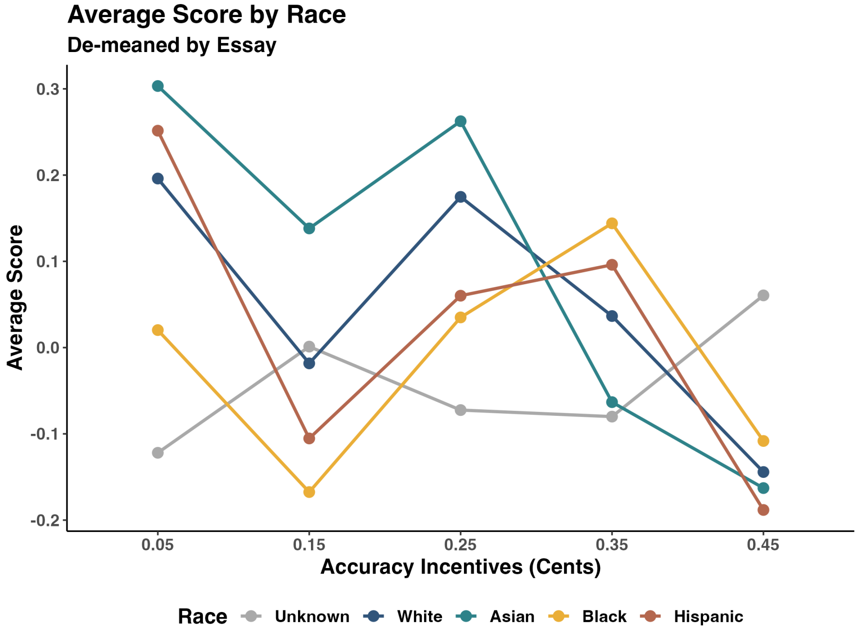
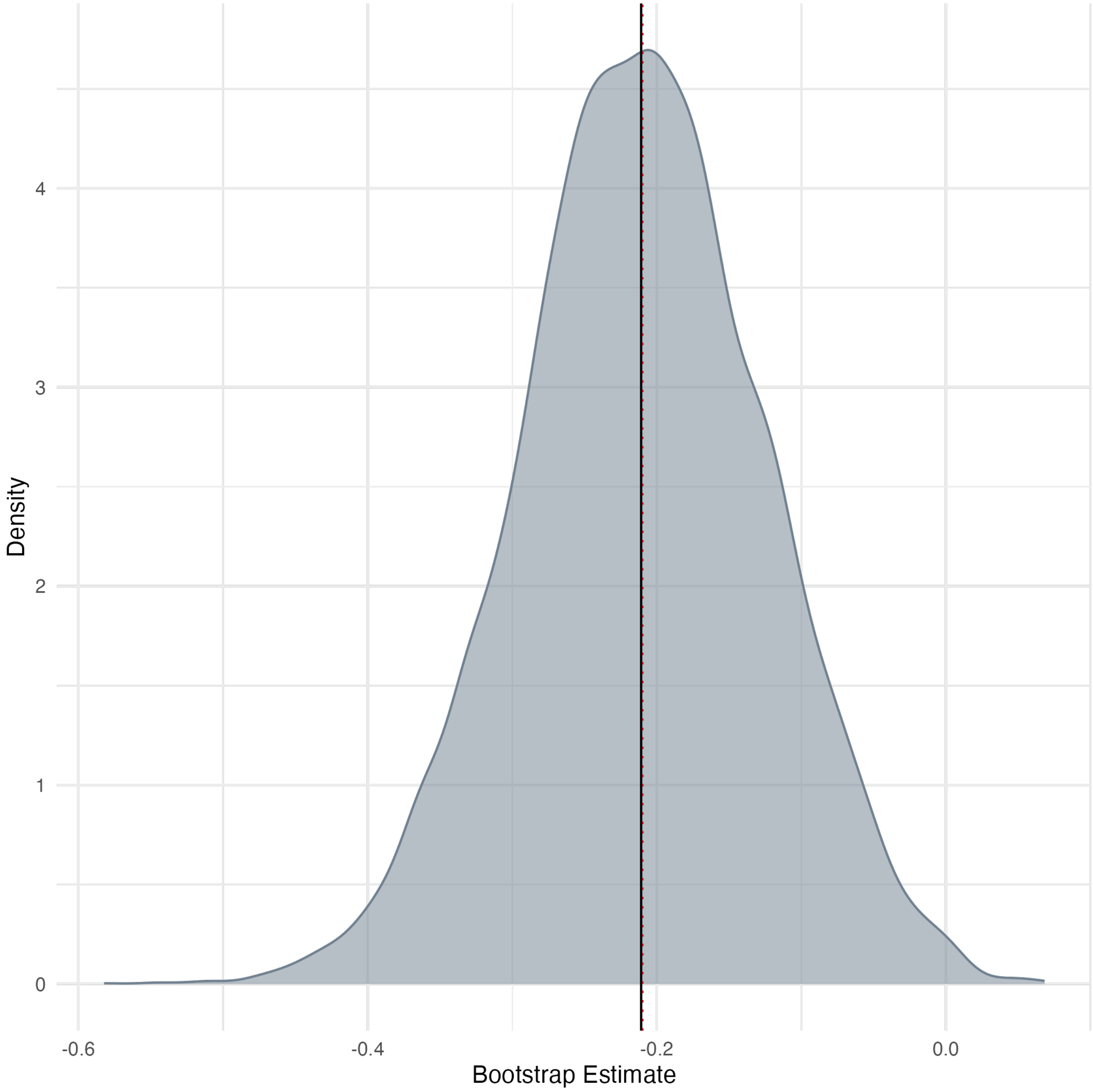
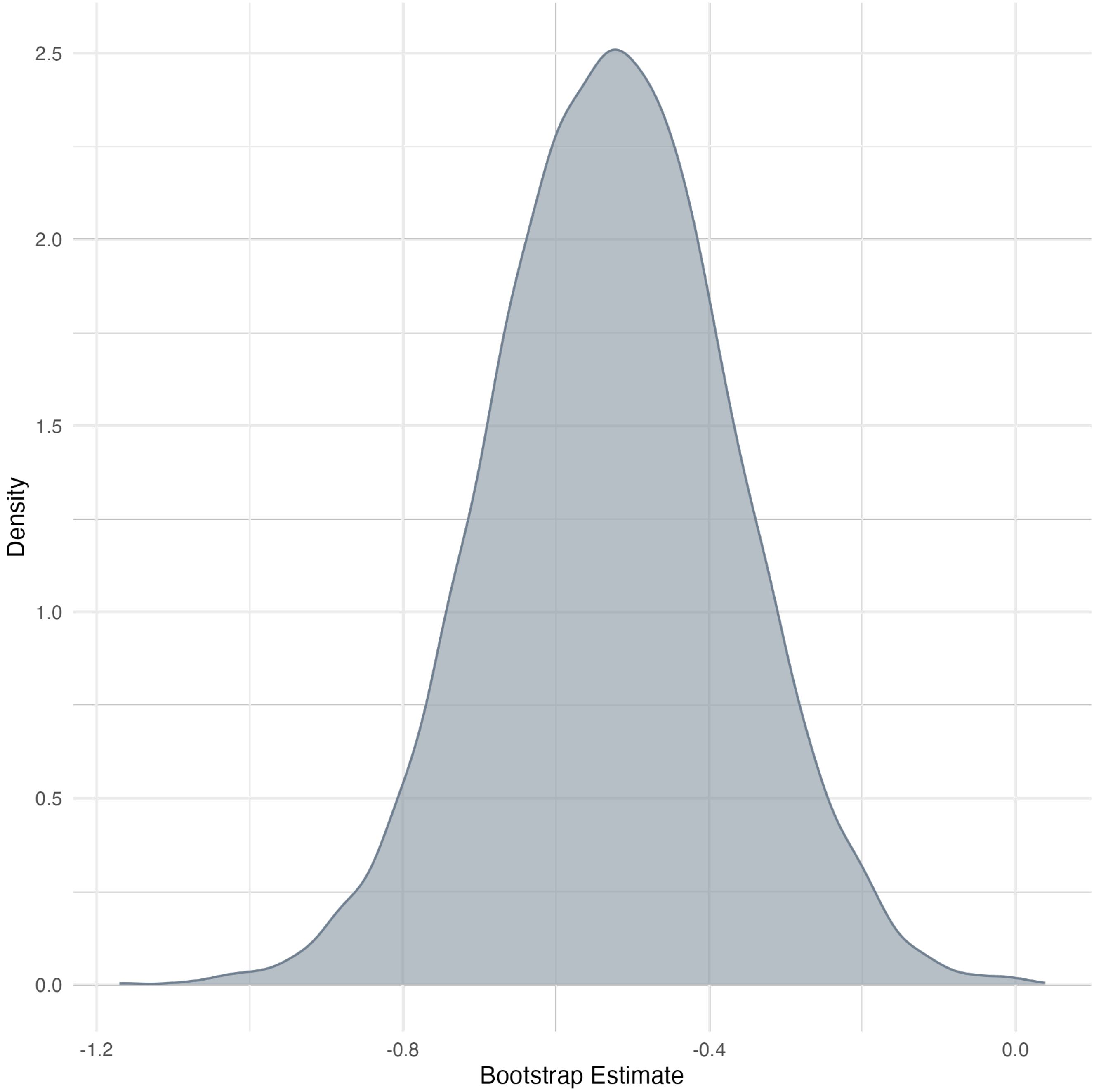


Figure D.3 Bootstrap without Replacement: Impact of 5 Cents on the Black-White Gap in the Non-Blind Grading Environment



This figure shows the density of estimating the Black-white gap at 5 cents as shown in row 2 of Table 7. The standard deviation of the bootstrap results is 0.084 compared to 0.109, the standard error estimated in Table 7.

Figure D.4 Bootstrap without Replacement: Impact of 5 Cents and the First Essay on the Black-White Gap in the Non-Blind Grading Environment



This figure shows the density of estimating the Black-white gap at 5 cents as shown in row 2 of Table 12. The standard deviation of the bootstrap results is 0.16 compared to 0.21, the standard error estimated in Table 12.

APPENDIX E

CHAPTER 2 - EXPERIMENTAL MATERIAL

E.0.1 The Essays.

Submission 1:

In our ever-progressing technological landscape, it becomes crucial to acknowledge and tackle the potential drawbacks that accompany such advancements. Challenges like privacy breaches, biased decision-making, and job displacement demand our attention and proactive solutions. One proposal that aims to hold companies accountable for the negative impacts of their AI algorithms is the implementation of an AI tax. This taxation mechanism serves as an incentive for companies to prioritize safety and fairness, while also generating revenue that can be directed towards supporting those adversely affected by AI disruptions. By adopting a comprehensive approach to this multifaceted issue, we can fully harness the transformative power of AI technologies, ensuring not only their responsible use but also the overall well-being of our society as a whole.

Submission 2:

I'm a little uncertain about an AI tax. I mean, it could potentially provide funding for social programs, which sounds pretty good. I'm wondering if it might also discourage companies from investing in AI research and development. That could result in progress being slowed down and technological advancement coming to a halt, which would be bad news for everyone, I think. I'm really not sure what to think about this whole thing.

Submission 3:

AI technolgy keep getting better and all, but there is some downsides. Like privacys breaches, unfair decisions, and people loosing their job because of it. Have *to make those companys pay for the bad stuff they're AI do, making sure they're doing it safely and fairly. we can use the money we're getting from that tax to help out the folks affected by AI messing things up and teach them about AI. it's like taking action now to create a future where AI do good things for everyone.*

Submission 4:

Possibly. It's difficult to determine whether implementing an AI tax is the best course of action. On one hand, it could potentially provide an avenue for funding social programs and offer support for workers who might be negatively impacted by automation, which are certainly admirable goals. However, on the other hand, there is a risk that such a tax could discourage companies from investing in AI research and development, which could result in slower technological advancement and less innovation. The potential consequences of an AI tax must be weighed carefully before deciding if it is the right choice

Submission 5:

I think an AI tax is good. It could make sure that everyone gets to share the good things that come from AI.

Submission 6:

No, I'm worried that an AI tax might really slow down how fast we can more VR stuff... We got to keep pushing the boundaries of virtual reality. But if they're all like, "Eh, too expensive now," then we're gone be stuck with the same old VR tech, which is not cool. I want to explore new worlds...

Submission 7:

It is essential that companies utilizing AI be subject to taxation. Such a measure would require companies to take responsibility for supporting individuals who are negatively impacted by the increasing use of AI. If companies are deriving *profits from the utilization of AI, then it is only appropriate that they provide assistance to those who have lost their jobs as a result of the technology.*

Submission 8:

As automation continues to disrupt the job market, it's important that we explore solutions that address the economic and social impacts of this trend. The idea of implementing an AI tax has been proposed, which would require corporations to pay a fee for their use of AI technology. While this measure could provide some support to displaced workers, it's important to recognize that it's not a panacea for the complex challenges posed by automation. Moreover, the idea of taxing AI is still controversial, and it's unclear how such a tax would be implemented and enforced in practice. Therefore, we need to carefully consider the potential consequences of an AI tax and explore other options that ensure the wellbeing and dignity of all members of society. Ultimately, we must work towards a comprehensive solution that balances the benefits of automation with the needs of workers and society as a whole.

Submission 9:

I'm really unsure about an AI tax. It could be a good idea because it might help make things more equal by spreading money around. But at the same time, there could be some things that happen that we don't expect, and that might be bad. It's a tough problem, and I just don't know what the best thing to do is.

Submission 10:

An AI tax is when the government takes money from robots. This is a great idea because robots make a lot of money, and they don't need it all. An AI tax will make sure that the money goes to the people who need it, like those who lost their jobs because of robots. Robots are taking over many jobs and making a lot *of money. An AI tax will make them pay their fair share and help people who are affected by automation. This will make sure that everyone benefits from the use of AI and that the money goes to the people who need it most.*

Submission 11:

It behooves one to carefully consider the ramfications of an AI tax when attempting to address the deleterious effects of automation on society. Given the complexity of the matter at hand, it is not immediately clear whethr implementing an AI tax is the optimal solution. Further investigation and deliberation are required to arrive at a definitive conclusion.

Submission 12:

Unclear. It could hold companies accountable for the impact of automation on society. On the other hand, it could prevent progress and slow down technological advancement.

Submission 13

Yes, we should indeed impose an AI tax as it could present a potential source of revenue for the government to utilize in various social programs and support systems for workers whose jobs have been taken over by automation. By taxing companies that benefit from AI, we could ensure that the advantages of automation are fairly distributed among all members of the society. Nevertheless, the implementation of an AI tax could be challenging and could potentially bring about undesirable outcomes such as hindering the progress of technological advancements and triggering job losses in the tech industry, which is something to worry about.

Submission 14:

Imposing an AI tax is an interesting proposal that warrants further discussion. While it could potetially generate additional revenue for the government, we must carefully consider the unintended consequences. For instance, companies may pass on the burden of the tax to consumers, resulting in higher prices for AI-powered products and services. Balancing the benefits and drawbacks is key before making a final determination.

Submission 15:

I think implementing an AI tax is worth considering, but we need to be mindful of its impact on innovation and economic growth. If the tax is too burdensome, it could discourage companies from investing in AI research and development, stifling progress. Striking the right balance between taxation and fostering technological advancements is crucial.

Submission 16:

There are various grounds for considering an AI tax as a suboptimal choice. Taking the international trade perspective into account, an AI tax could create obstacles for companies in terms of competing in the global market, especially if they are subjected to additional taxes. This may result in reduced international trade, dampened innovation, and an increase in trade costs, which may minimize data acquisition. In theory, AI investments should strengthen global competition, so I don't think we should hastily impose taxes on companies that use artificial intelligence.

Submission 17:

An AI tax proposal suggests that the government should impose a tax on robots, which some people believe could help support those who have lost their jobs due to automation. However, there are concerns about the potential negative *impact on the existence of robots. While I can see some benefits to this idea, it is also important to consider the potential consequences and drawbacks. Overall, I think further research and discussion are needed before making a final decision on implementing an AI tax.*

Submission 18:

While the concept of an AI tax seems reasonable on the surface, it's crucial to recognize the pressing financial needs of individuals in our society. As automation and AI technologies continue to advance, we must ensure that the benefits of these advancements are shared equitably among all members of our community. Rather than solely focusing on taxing robots, it is paramount that we address the real-life struggles and economic hardships faced by many individuals. By redirecting our attention and resources towards initiatives that directly support people in need, such as providing adequate job training, affordable housing, and accessible healthcare, we can foster a more inclusive and just society. While an AI tax could potentially generate revenue for social programs, let us not lose sight of the fact that it is people, not robots, who require financial assistance and opportunities for a better future. Therefore, we should prioritize policies and measures that uplift individuals and communities, ensuring a more equitable distribution of wealth and resources.

Submission 19:

The impact on innovation in science and technology is one of the primary reasons we should not have an AI tax. If companies are dissuaded from investing in AI research and development due to the added tax burden, it could impede progress and prevent technological advancements that could significantly improve society.

Submission 20:

Implementing and enforcing an AI tax poses significant challenges for governments. History provides examples of tax systems that struggled to achieve their intended outcomes. For instance, the U.S. luxury tax in the 1990s aimed to generate revenue and address income inequality but resulted in job losses and minimal wealth redistribution. Predicting the revenue from an AI tax is uncertain, making it difficult to fund education and training programs effectively. Moreover, concerns exist that an AI tax could stifle innovation and discourage investment, as seen with the UK's sugary beverage tax, which led to product reformulation rather than healthier alternatives. The complexities of taxation require governments to carefully assess the viability of an AI tax, considering real-world examples and potential drawbacks. Effectively implementing and enforcing such a tax is crucial to ensure the intended goals are met without unintended negative consequences.

E.0.2 Survey Instructions. Before starting the survey, participants must first read through a letter of consent describing the objectives and risks associated with the study and give their consent to participate. Once read, participants receive instructions on their task, including contextual information on the grading environment, described in the text below:

For the next few minutes, we will ask you to act as if you are a teacher grading a sequence of homework submissions.

Grading Scenario: In a fictitious class discussion, students discussed the pros and cons of companies using artificial intelligence to produce goods and services. Pretend these fictitious students had to answer the following question on their homework: ”Should we tax companies that use artificial intelligence in place of human workers?” You can reference brief class notes on the discussion and the grading rubric as you grade.

Your task is to grade eight short free-response homework submissions from these fictitious students.

NOTE: No actual students' work was used in this survey.

- You will have access to a grading rubric.

- Each submission includes the grading rubric, so you can always reference it as you grade.

- Each submission also includes the brief lecture notes you can reference as you grade.

- To grade, you can assign a numerical score between 0 and 5 on each submission, where 5 is 100%, and 0 is 0%.

E.0.3 Accuracy Incentive Payment.

The text I use to explain the accuracy incentive to participants is the following:

Explanation of bonus payment:

The points on each submission have been pre-determined by a group of university-level instructors. The average pre-determined score given to each submission is between zero and five.

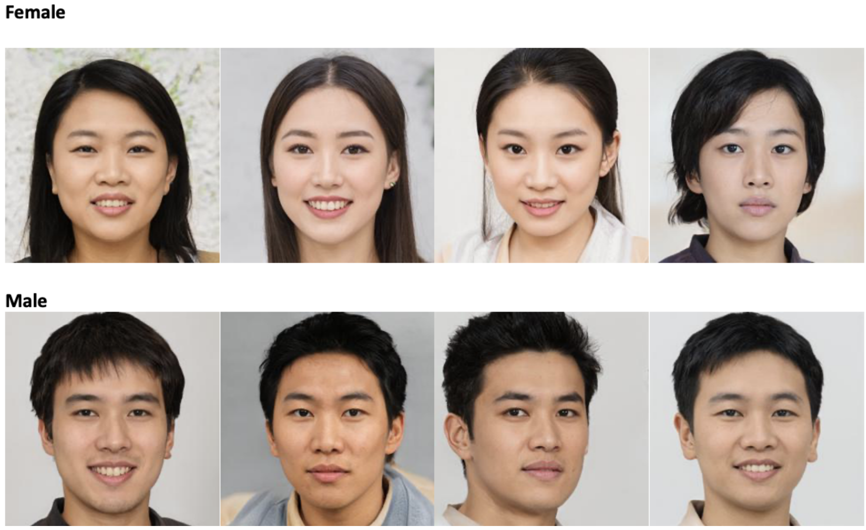
As you grade, if you assign a score within half a point of the pre-determined average score, you will earn a bonus of [random amount between 0.05 and 0.45 cents in ten-cent increments] per submission.

There are eight submissions, so you can earn up to [8 times the random amount in bonus payments] plus the $1.75 for participating in this study.

In total, you can earn up to [Total amount]

The $1.75 will be distributed to you immediately after completing the survey and registering your submission with Prolific. The bonus will be distributed within three business days.

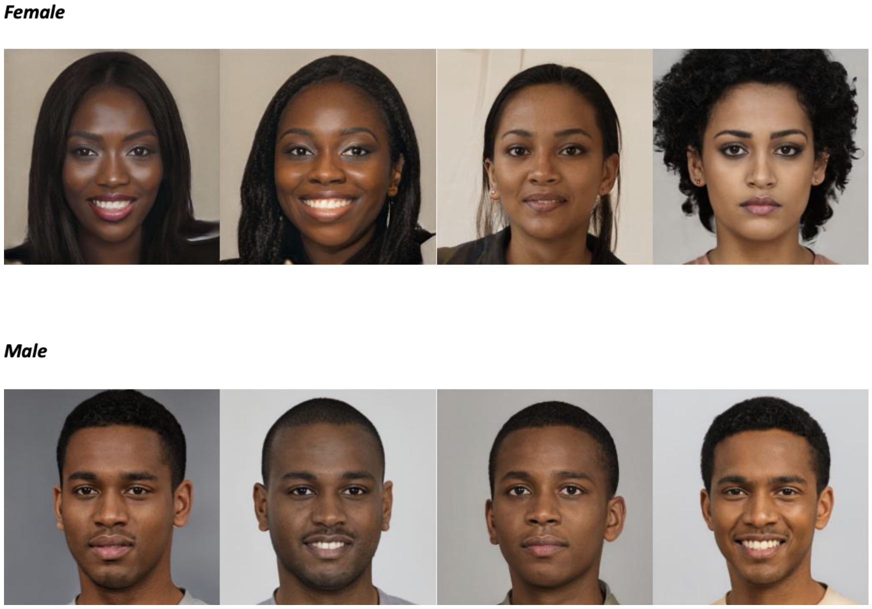
Figure E.1 Photos of the Asian Students



Female

Male

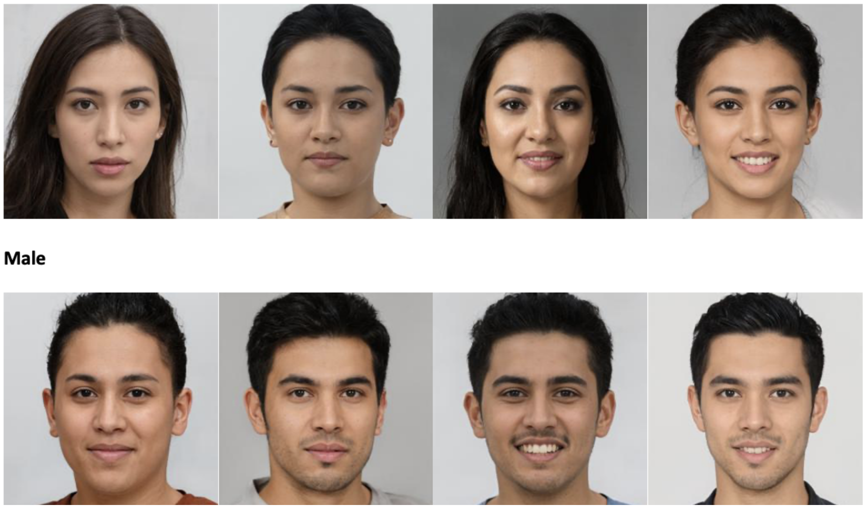
Figure E.2 Photos of the Black Students



Female

Male

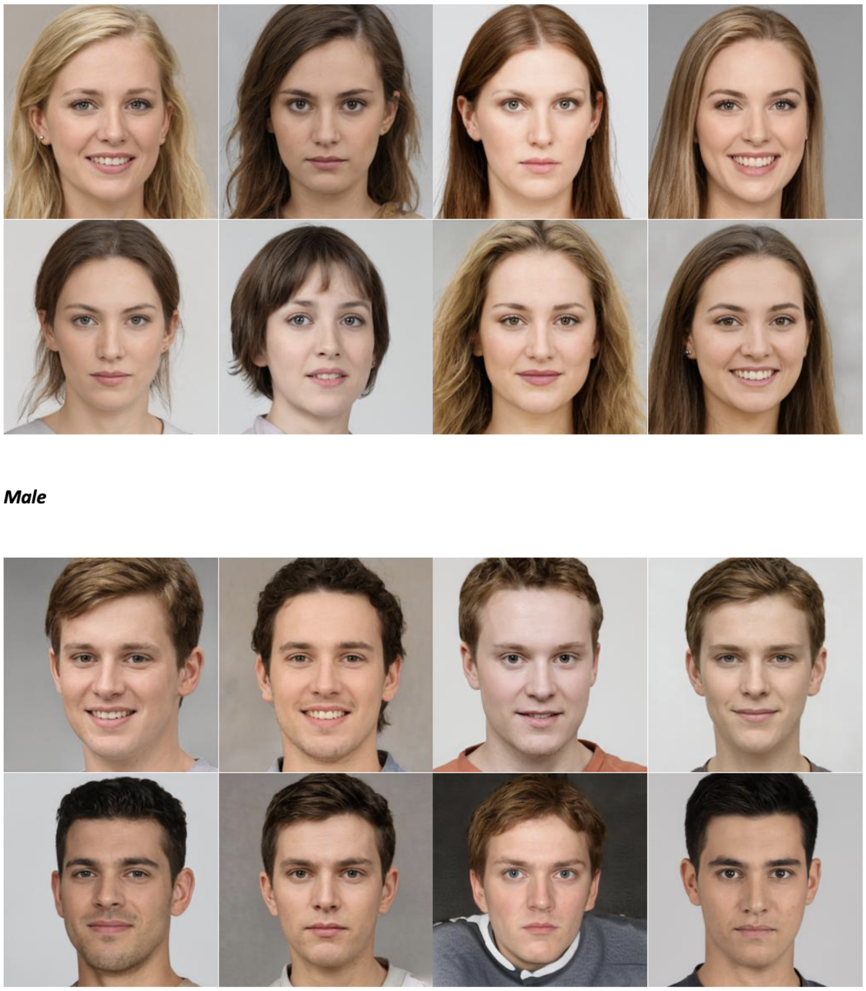
Figure E.3 Photos of the Hispanic Students



Female

Male

Figure E.4 Photos of the White Students



Female

Male

Figure E.5 Names of the Female Students

Female			
White	Black	Asian	Hispanic
Danielle Kane	Abigail Anthony	Emily Vang	Bella Cortez
Caroline Bowen	Kimora Boone	Hong Li	Maya Lugo
Lila Stanton	Arielle Person	Evelyn Huynh	Camila Cervantes
Phoebe Harding	Laila Griffen	Elizabeth Nguyen	Alison Salas
Victoria Vance			
Juliet Barrett			
Samantha Anderson			
Mary Burns			

Figure E.6 Names of the Male Students

Male			
White	Black	Asian	Hispanic
Henry Marsh	Isaiah Walker	Ethan Zhang	Erick Nunez
Charles Arnold	Christian Walton	Benjamin Huang	Joel Rubio
Oliver Brandt	Jaiden Dixon	Felix Chen	Jose Delgado
Peter Strickland	Michael Miles	Wei Xiong	Ian Alfaro
Aaron Casey			
Evan Hendricks			
Steven Carey			
Andrew Horn			

APPENDIX F

CHAPTER 3 - DATA CLEANING

In this section, we discuss agencies we had to drop from our data (email error), and how we assessed the sentiment of the responses from law enforcement.

Email Error After finishing the experiment and analyzing the responses/sent emails, we ended up having to remove 224 law enforcement agencies from our final data set for a few reasons, giving an email error rate of 5.6%.

1. The sheriff's offices with only the online forms for contact information required sending the messages manually to each website. There were four days during the eight week experiment in which those messages were not sent. Furthermore, several websites did not allow messages to go through, either they requested a phone number (which we did not have) or had a minimum number of characters we could use in the message (which was shorter than our message itself). We did not send 66 messages via forms.

2. The software program that we used for automating our emails to go out to law enforcement agencies did not send messages out on the last day of the experiment due to some technical error. We lost *149 observations* from this error.

3. While gathering sheriff's offices contact information, 27 observations included duplicated email addresses that belonged to different agencies than the one the addresses were connected to. We caught this mistake during the fourth week of the experiment. Because we caught it in the middle of the experiment, we were able to gather the appropriate email addresses for the sites that had an incorrect email address. For those sites in which we had

already contacted earlier in the experiment by mistake, we dropped them. Of the 27 duplicates, we dropped *12 agencies* from our data.

To verify that this email error rate is uncorrelated with treatment, we regressed an indicator variable for whether the agency is to be removed from our data on the indicators of treatment. Table F.1 shows the results. Column 1 shows the correlation across race and gender. We find no significant differences across treatment. Column 2 shows the correlation across email signal. We also evaluated the interaction between the putative identities and the assigned signals. The hypothesis tests are shown in Figure H.2. We find no statistically significant correlation in missingness.

Bounced Emails Occasionally, emails we had sent would receive an automatic response saying the email did not go through to the law enforcement agency. In fact, we had 299 locations in which we received at least one bounced email. Additionally, while monitoring our mailboxes during the experiment, we noticed that during the second week of the experiment right after the software we used had updated, all sent emails from two email addresses associated with the Hispanic identity were automatically blocked. We decided to create a new email address for those identities and resend those emails throughout the following six weeks of the experiment.[1] We resent 32 emails, in which law enforcement responded to 13 of them.

To verify that the bounced emails are uncorrelated with the assigned race and gender, we regressed an indicator for whether the law enforcement agency had a bounced email on indicators of treatment. Table F.2 shows those results. Column

[1] We re-randomly assigned the day and week that those emails would go out.

1 reports the results for the whole dataset. We find that there is a statistically significant correlation between the Hispanic female identity and the bounced emails. This is to be expected given that two email addresses for the Hispanic identity we used had automatically blocked all emails going out from our mailbox. However, when we remove those locations from the data, the correlation disappears, as shown in column 2. In our final dataset, we do not remove any location that had a bounced email. We believe that because we were able to resend the 32 emails to law enforcement agencies and 13 of those agencies responded, we need not drop the agencies that had been assigned the Hispanic identity which had been initially blocked.

Responses from law enforcement In our analysis we (1) identify if law enforcement responded and (2) how law enforcement responded within their emails, specifically the word count and the sentiment of the text. To evaluate the word count and the sentiment of the text, we took several steps to connect every received email with the unique law enforcement agency's identifier in our data. Some agencies responded multiple times, some used different email addresses than the one we used to email them. After going through each email, we identified emails as *automatic*, *bounced*, or *response*. Emails labeled as "automatic" are emails in which law enforcement sent an automatic reply with an out of office message or a request for our identifying information. We do not treat these responses as a "response" to our inquiries. Bounced emails are those as described in the previous section. And lastly, any email that was a direct response to our inquiry, we counted as one response.

To conduct a word count and text sentiment analysis, we removed from each email the names of the person responding from the agency, any signature

Table F.1 Correlation of Email Error with Treatment

	Dependent variable:	
	Email Error (1=Remove from Data)	
	(1)	(2)
Black Female	−0.007	
	(0.012)	
Black Male	0.006	
	(0.012)	
Hispanic Female	−0.011	
	(0.012)	
Hispanic Male	0.002	
	(0.012)	
White Female	−0.004	
	(0.012)	
Positive Signal		0.016^{**}
		(0.008)
Negative Signal		0.014^{*}
		(0.008)
Observations	3,986	3,986
R^2	0.184	0.184
Omitted Variable	White Male	Neutral Signal
Note:		*p<0.1; **p<0.05; ***p<0.01

Column 1 reports the results of regressing an indicator variable of the email error (1=Remove from the data, 0 otherwise) on indicator variables of the assigned putative race by gender identities. Column 2 reports the correlation of the email error rate with the assigned signals. We find no correlation between the identity treatment and the email error rate. However, we do find that there are 1.6 pp and 1.4 pp more positive and negative signals in our final data set.

Table F.2 Bounced emails by treatment

	Dependent variable:	
	Bounced (Yes=1)	
	(1)	(2)
Black Female	−0.020	−0.020
	(0.014)	(0.014)
Black Male	−0.019	−0.019
	(0.014)	(0.014)
Hispanic Female	0.038**	0.006
	(0.017)	(0.016)
Hispanic Male	0.026*	0.008
	(0.016)	(0.015)
White Female	−0.001	−0.001
	(0.015)	(0.015)
Observations	3,762	3,762
R^2	0.026	0.022
Note:	*p<0.1; **p<0.05; ***p<0.01	

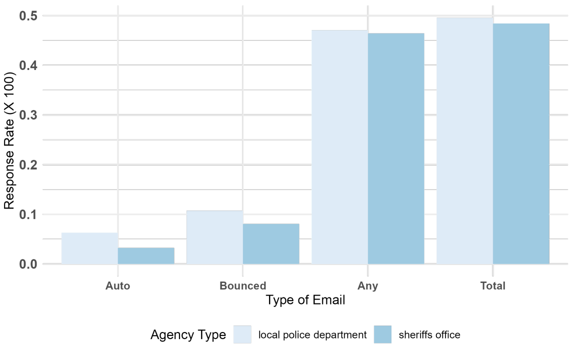
This table shows the results from regressing an indicator variable denoting whether an email bounced. Column 1 is on the full sample, after removing the missing data. Hispanic Female and Hispanic Male are correlated with bounced. However, this is due the fact that during the second week of the experiment, all emails from two email addresses associated with the Hispanic identity were automatically blocked. In looking at the error code, we came to the conclusion that the email address had been flagged as spam. We created a new email address for the same identity and randomly reassigned the week and day in which we would send law enforcement those emails. A total of 31 emails were resent. Of the 31 resent, we received 13 responses from Law Enforcement. Column 2 shows the results from the same regression after removing those 31 sites from our data set, showing that the correlation shown in Column 1 is likely due to the email address being flagged as Spam. All emails after the second week going out for the specific Hispanic identities were with the new email addresses.

information, and any text that was automatically placed within the text (like a warning that our email was not from within their organization). We also removed links. One agency in our data set responded with just a link, so that response did not have a word count. We also removed any special characters, stop words, and punctuation in the text to separately analyze the text sentiment. Stop words are words like 'the', 'is, 'she', or 'he.' We did this so we could use the Vader R package, a commonly used package by social scientists for text sentiment, to identify whether words are negative, positive, or neutral in each email. The function we used in this package also creates an index of sentiment ranging from -1 to +1, with -1 being the message that is the most negative and +1 being the message that is the most positive. For law enforcement agencies who responded multiple times, we used the median sentiment score and word count in our final data set. Figure H.4 and Figure H.5 show the densities of the word count and the text sentiment in our data.

One important thing to note is that we believe there is a weakness with using standard text sentiment analysis for this experiment. Words associated with gun or firearm received negative scores. Given that we asked law enforcement for information about how to apply for a permit (if necessary), responses that included words related to firearms may actually be incredibly helpful and positive. However, the Vader package labels those emails as very negative. Although we do analyze the text sentiment in this study, it is important to note that for this experiment the Vader package and standard text sentiment practices may not be complete in providing sentiment scores for the email responses.

F.0.1 Summary of Response Rates. In this section, we discuss the four types of summary statistics of our final data set. They include (i) automatic responses, (ii) bounced emails, (iii) any response, and (iv) total response. We show the response rates for each of those variables in Figure F.1 and Figure F.2 by law enforcement agency type, i.e., police department or sheriff’s office, and the assigned signal on the initial email we had sent, i.e., neutral, positive or negative. Each of these data are explained in further detail below.

Figure F.1 **Summary Statistics** - Response Rates by Response Type



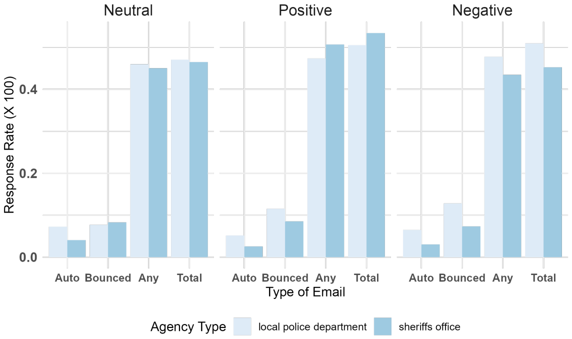
This figure shows the response rates based on whether the response from law enforcement is automatic, such as an out of office notification (**Auto**), the total number of law enforcement agencies in which our inquiries resulted in a bounced email (**Bounced**), the total number of law enforcement agencies that responded (**any**), and the total number of responses from all law enforcement agencies (**Total**).

i. Automatic responses. The first statistic is the number of automatic responses from law enforcement. These emails include responses such as ”Thank you for reaching out to [Law Enforcement Agency]. We will get back to you shortly,”[2]. or emails that require some type of authentication of our identity. After sending our initial inquiries, we never responded to any questions or responses from law enforcement (we made this decision deliberately during our pre-work with the IRB).

ii. Bounced. The second statistics is the response rate on whether we received a bounced email after sending our own emails to law enforcement. There

[2] This text is made up, but provides the essence or sentiment of many automatic responses that we had received

***Figure F.2* Summary Statistics** - Response Rates by Email Type and Assigned Email Signal



This figure shows the response rates based on whether the response from law enforcement is automatic, such as an out of office notification (**Auto**), the total number of law enforcement agencies in which our inquiries resulted in a bounced email (**Bounced**), the total number of law enforcement agencies that responded (**any**), and the total number of responses from all law enforcement agencies (**Total**). Additionally, we show the responses by the assigned signal on the emails we had sent out.

were a few instances in which we received multiple bounced emails from the email server for one email address we had tried to contact. This did not happen often. Figure H.3 in the Appendix shows the count of bounced emails by each law enforcement agency. Some of the errors received in the bounced emails include the mailbox being full or the the email address not existing in the network.

iii. Any Response. The third statistic is the response rate based on whether a law enforcement agency responded to our inquiries. This variable is the main outcome we use in our response rate analysis, as we are interested in whether law enforcement will respond, not how many times they do respond.

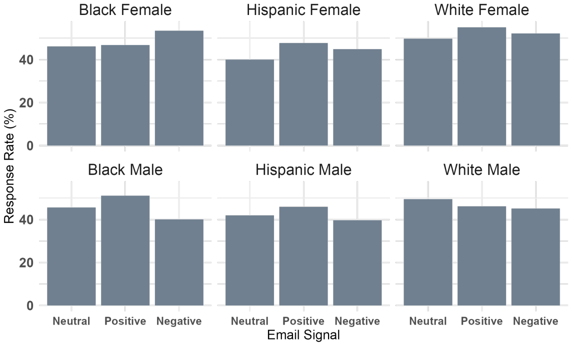
iv. Total Response. Lastly, our fourth statistic shows the response rate based on how many times law enforcement responded. There are a few instances in which a specific agency responded multiple times, asking if we had received their previous email, or that they had forwarded our message to the appropriate person, and so on. Although this is an interesting metric, we do not use this metric in our main results, primarily because the number of law enforcement agencies that responded multiple times are few (under a hundred) and because we're more interested in whether law enforcement do respond.

Response Rates. In total, the 'Any' response rate is between 46.4% for Sheriff's Offices and 47% for Local Police Departments. If we consider 'Total responses' from law enforcement, the response rates increase to 48.4 % and 49.5%, respectively. When we evaluate how the 'Any' response rates change depending on the signal assigned to the initial email we had sent, response rates do vary across signal and LEA.

Figure F.2 highlights how response rates are nearly identical between LEAs (45.9% for police departments and 45.1% for sheriff's offices) when we sent the neutral signal (i.e., the email with just a question about the permit). For emails with a positive signal or a negative signal, response rates diverge. For the positive signal (i.e., a statement about willingness to undergo a background check), sheriff's offices responded 50.6% of time time, whereas police departments responded 47.3 % of the time. For the negative signal (i.e., a question about how far back background checks go in one's history), sheriff's responded 43% of the time and police departments responded 47.7% of the time. We find no statistically significant differences when we conduct hypothesis tests that for each signal the response rates are statistically different between agency types, nor do we find any significant differences in response rates across signals for each agency type.

Lastly, we show the response rates by the assigned signal and the demographic identity in Figure F.3. The height of the bars are the response rates. The horizontal axis is denoted by the assigned signal, and the individual charts in the figure are the response rates for each demographic identity. The variation in response rates across signals seem to be more pronounced for Hispanic and Black males. For the female identity, there is an increase in responses with both the positive and negative signal. For the white male identity, response rates decrease relative to the neutral signal.

Figure F.3 **Summary Statistics** - Response Rates by Assigned Email Signal and the Putative Demographic



This figure shows the 'Any' response rates by the assigned email signal and the demographic. The raw data in the Appendix: Table G.3.

APPENDIX G

CHAPTER 3 - TABLES

Table G.1 Last names used in study

White	Black	Hispanic
Olson	Washington	Hernandez
Schmidt	Jefferson	Gonzalez
Meyer	Jackson	Rodriguez
Snyder	Joseph	Ramirez
Hansen	Williams	Martinez
Larson	Banks	Lopez

Table G.2 First names used in study

White Men	White Women	Black Men	Black Women	Hispanic Men	Hispanic Women
Hunter	Katelyn	DaShawn	Tanisha	Alejandro	Mariana
Jake	Claire	Tremayne	Lakisha	Pedro	Guadalupe
Seth	Laurie	Jamal	Janae	Santiago	Isabella
Zachary	Stephanie	DaQuan	Tamika	Luis	Esmeralda
Todd	Abigail	DeAndre	Latoya	Esteban	Jimena
Matthew	Megan	Tyrone	Tyra	Pablo	Alejandra
Logan	Kristen	Keyshawn	Ebony	Rodrigo	Valeria
Ryan	Emily	Denzel	Denisha	Felipe	Lucia
Dustin	Sarah	Latrell	Taniya	Juan	Florencia
Brett	Molly	Jayvon	Heaven	Fernando	Juanita

Table G.3 Summary of Response Rates by Email Signal and Putative Demographic

Signal	Demographic	Treated	Responses	Response Rate
Neutral	Black Female	208	96	46.15
Positive	Black Female	203	95	46.80
Negative	Black Female	202	108	53.47
Neutral	Hispanic Female	215	86	40.00
Positive	Hispanic Female	199	95	47.74
Negative	Hispanic Female	225	101	44.89
Neutral	White Female	215	107	49.77
Positive	White Female	218	120	55.05
Negative	White Female	205	107	52.20
Neutral	Black Male	206	94	45.63
Positive	Black Male	219	112	51.14
Negative	Black Male	212	85	40.09
Neutral	Hispanic Male	205	86	41.95
Positive	Hispanic Male	209	96	45.93
Negative	Hispanic Male	204	81	39.71
Neutral	White Male	214	106	49.53
Positive	White Male	197	91	46.19
Negative	White Male	206	93	45.15

Table G.4 Heterogeneity of the Results by the Agency Type

	Response Rate		Email Length		Email Sentiment	
	(1)	(2)	(3)	(4)	(5)	(6)
Black	-0.01		$0.12^{\cdot}$		0.01	
	(0.03)		(0.07)		(0.04)	
Hispanic	$-0.05^{\cdot}$		0.04		-0.03	
	(0.03)		(0.07)		(0.04)	
Sheriff's Office	$0.05^{\cdot}$	0.02	0.06	-0.08	-0.09^{*}	-0.09^{*}
	(0.03)	(0.02)	(0.07)	(0.06)	(0.04)	(0.04)
Black × Sheriff's Office	-0.04		-0.20^{*}		0.03	
	(0.04)		(0.10)		(0.06)	
Hispanic × Sheriff's Office	-0.04		-0.20^{*}		0.03	
	(0.04)		(0.10)		(0.06)	
Female		0.04		0.15^{**}		0.01
		(0.02)		(0.05)		(0.03)
Female × Sheriff's Office		0.00		0.01		0.03
		(0.03)		(0.08)		(0.05)
Num. obs.	3762	3762	1759	1759	1758	1758
R^2 (full model)	0.06	0.06	0.05	0.06	0.08	0.08
Omitted Demographic Variable	White	Male	White	Male	White	Male

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $\cdot p < 0.1$

All regressions include week, US state, and agency type fixed effects. Cluster standard

Table G.5 Heterogeneity of the Results by the Background Checks

	Response Rate		Email Length		Email Sentiment	
	(1)	(2)	(3)	(4)	(5)	(6)
Black	−0.00		−0.02		−0.00	
	(0.03)		(0.08)		(0.05)	
Hispanic	-0.06^{*}		−0.06		0.00	
	(0.03)		(0.08)		(0.05)	
Background Check	0.04	0.03	−0.02	0.05	−0.04	$-0.07^{\cdot}$
	(0.03)	(0.03)	(0.07)	(0.07)	(0.04)	(0.04)
Black × Background Check	−0.03		0.08		0.02	
	(0.04)		(0.10)		(0.06)	
Hispanic × Background Check	−0.01		0.01		−0.02	
	(0.04)		(0.10)		(0.06)	
Female		0.05^{*}		0.20^{**}		−0.01
		(0.02)		(0.06)		(0.04)
Female × Background Check		−0.02		−0.10		0.05
		(0.03)		(0.08)		(0.05)
Num. obs.	3621	3621	1700	1700	1699	1699
R^2 (full model)	0.06	0.06	0.05	0.06	0.08	0.08
Omitted Demographic Variable	White	Male	White	Male	White	Male

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $\cdot p < 0.1$

All regressions include week, US state, and agency type fixed effects. Cluster standard errors at the county level. In columns 1 and 2, we regress a dummy variable on whether law enforcement responded to the sent emails that are assigned a putative race or gender interacted with the assigned signal in the sent email. Columns 3 through 6 are analysis of the responses from law enforcement. In columns 3 and 4, we log linearized the word count of the responses and regressed it on indicators of race or gender and signal on the assigned sent emails. In columns 5 and 6, we use the Vader R package to create a sentiment score of the responses from law enforcement that range between -1 (most negative) to +1 (most positive) and regressed those values on the assigned indicators.

Table G.6 Hypothesis Tests based on the Assigned Signals and Gender

Statement	**Equation**
Relative to the Male Identity:	
1. No difference in the neutral signal between the female and male identity	$\delta_1 = 0$
2. No difference in the positive signal between the female and male identity	$\delta_1 + \delta_4 = 0$
3. No difference in the negative signal between the female and male identity	$\delta_1 + \delta_5 = 0$
Relative to the OWN identity (i.e., female to female, male to male):	
1. No difference in the positive signal relative to the neutral signal for the female identity	$\delta_2 + \delta_4 = 0$
2. No difference in the negative signal relative to the neutral signal for the female identity	$\delta_3 + \delta_5 = 0$
3. No difference in the negative signal relative to the positive signal for the female identity	$\delta_3 + \delta_5 = \delta_2 + \delta_4$
4. No difference in the positive signal relative to the neutral signal for the male identity	$\delta_2 = 0$
5. No difference in the negative signal relative to the neutral signal for the male identity	$\delta_3 = 0$
6. No difference in the negative signal relative to the positive signal for the male identity	$\delta_3 = \delta_2$

Note: These hypothesis tests are based on the following regression:

$$(1) \quad Y = \beta_0 + \delta_1 \text{Female} + \delta_2 \text{Pos} + \delta_3 \text{Neg} + \delta_4 \text{Female} \times \text{Pos} + \delta_5 \text{Female} \times \text{Neg} + \gamma \text{FE} + \epsilon$$

Depending on the number of factors (such as the racial identity) in our dependent variables, the number of hypotheses we test do increase.

APPENDIX H

CHAPTER 3 - FIGURES

Figure H.1 Number of Law Enforcement Agencies in Each County

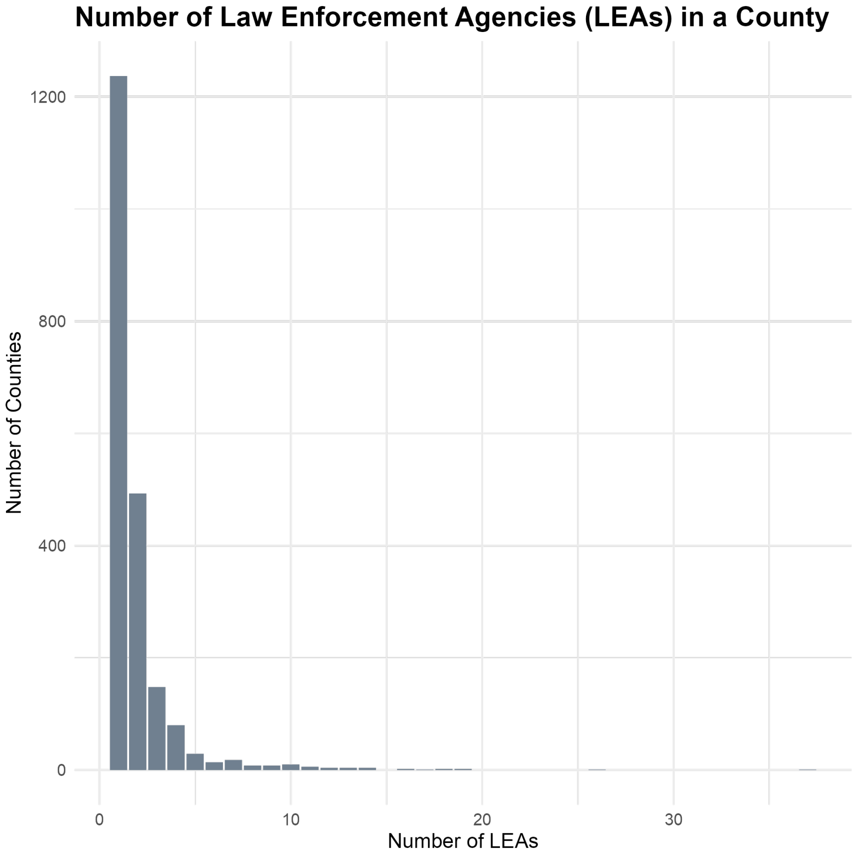
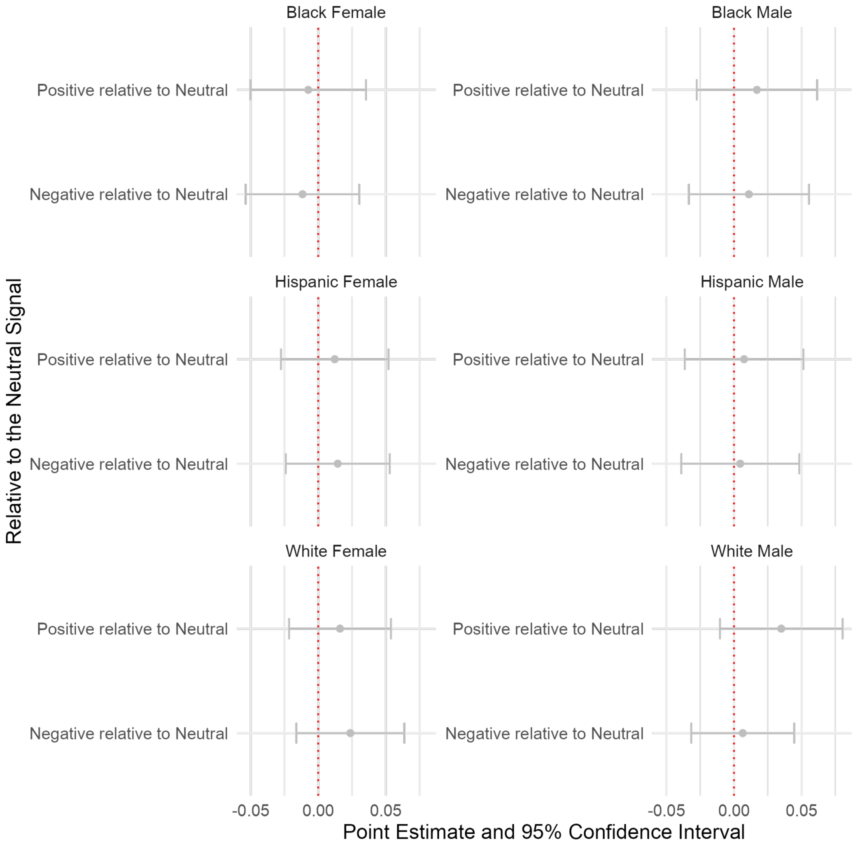
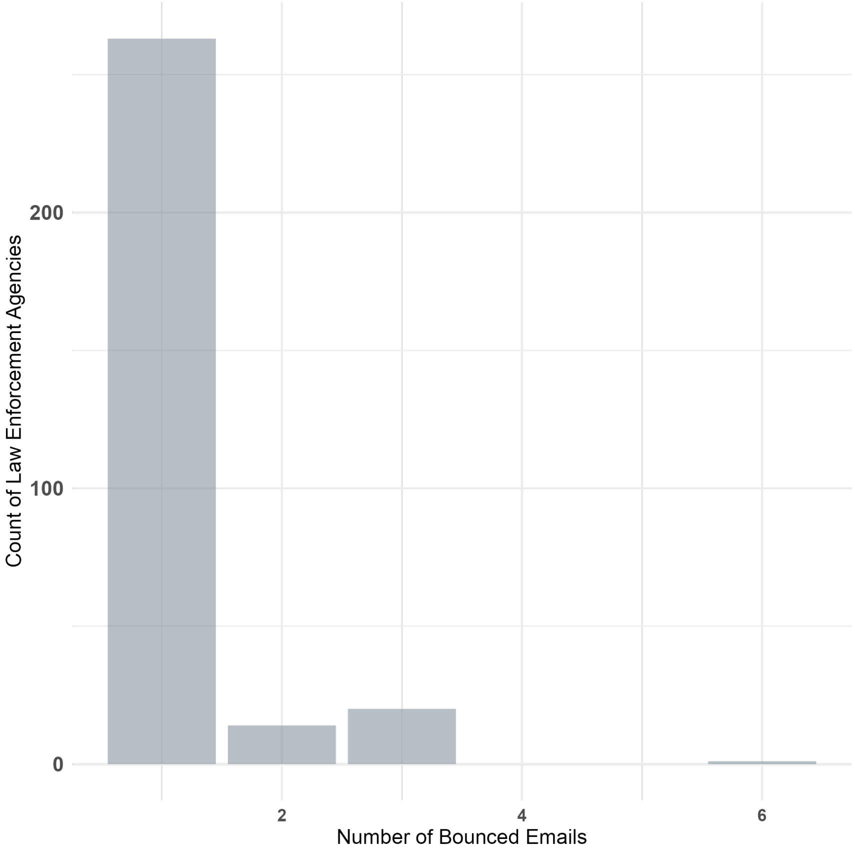


Figure H.2 Hypothesis Tests of Email Error by Signal, Race, and Gender



Note: This shows the results of conducting hypotheses tests based on the outcomes from Table F.1. It shows, relative to the neutral signal, whether email error is correlated with the positive or negative signal for each race by gender group.

Figure H.3 Number of Bounced emails by the LEA



This shows the number of law enforcement agencies that are associated with at least one bounced email. The horizontal axis shows the number of bounced email notifications we received. The vertical axis shows the count of law enforcement agencies. Of the 298 law enforcement agencies that had a bounced email in our final data set, only 35 had more than one bounced email.

Figure H.4 Distribution of the Word Length in the Email Responses

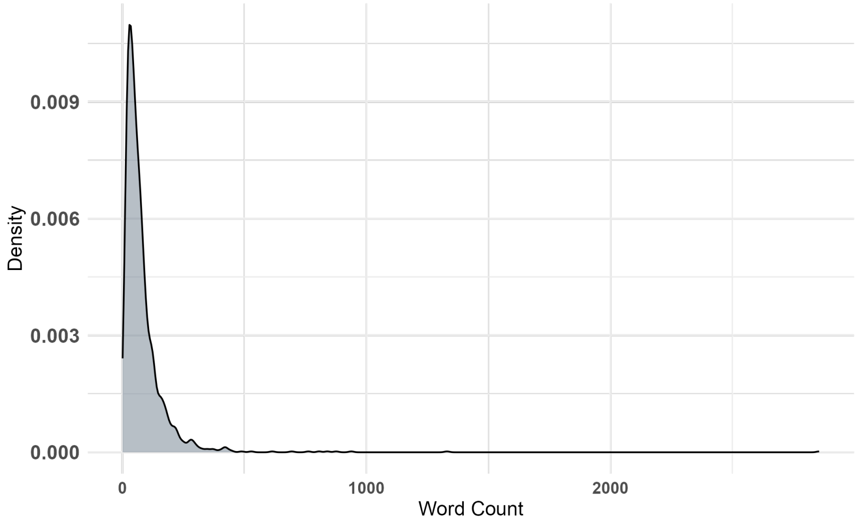
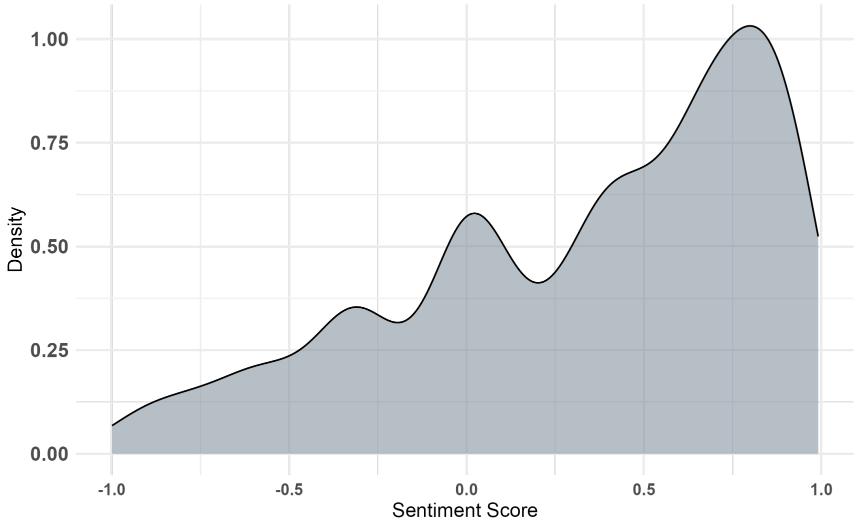
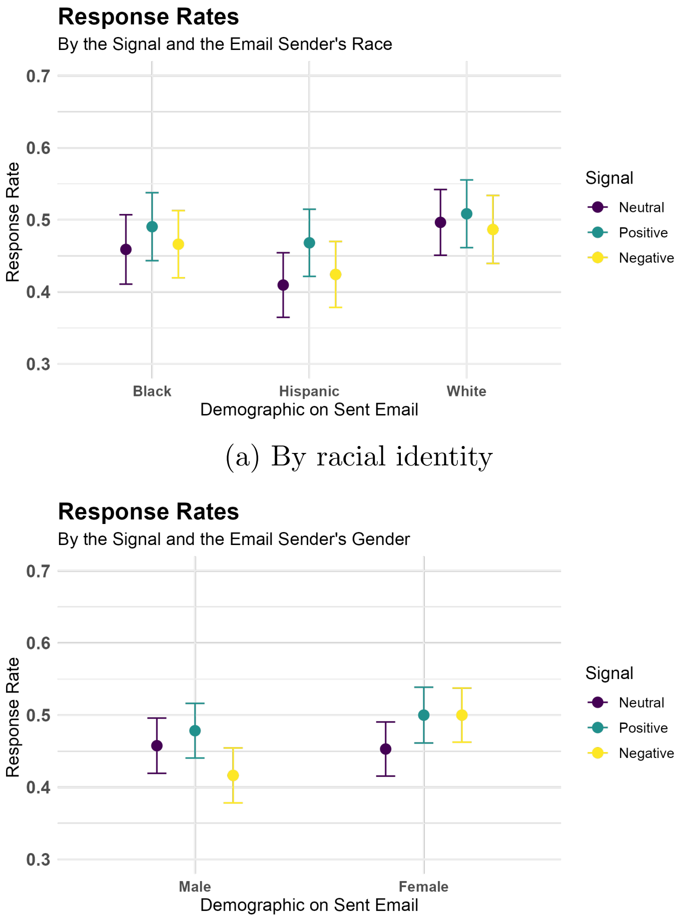


Figure H.5 Distribution of the Sentiment Score in the Email Responses



Note: Sentiment is a score between -1 to +1. We use the Vader package in R to assess whether each word in the email is negative, neutral, or positive. The functionality of the Vader package allows us to create an index of messages that are the most negative or most positive.

Figure H.6 Predicted **response rates (RR)** by the assigned signal, race OR gender

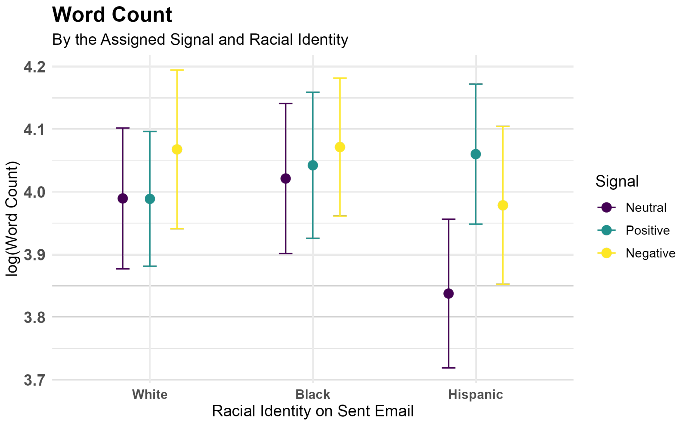


(a) By racial identity

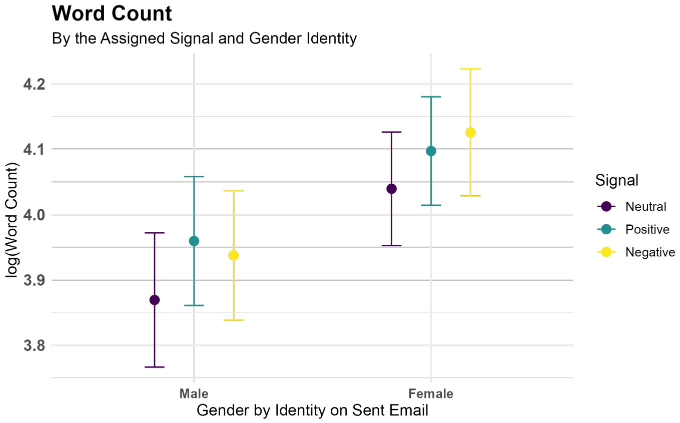
(b) By gender identity

Note: We use different regression to estimate the predictions in panels (a), (b), and (c). However, the estimates from each panel are derived from the same regression. In all three panels, we use state, week, and law enforcement agency type fixed effects and cluster the standard errors at the county level.

Figure H.7 Predicted **log(word count)** by the assigned signal, race or gender



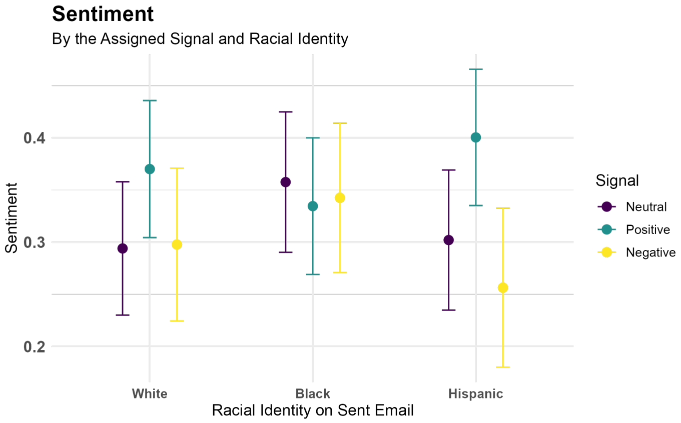
(a) By racial identity



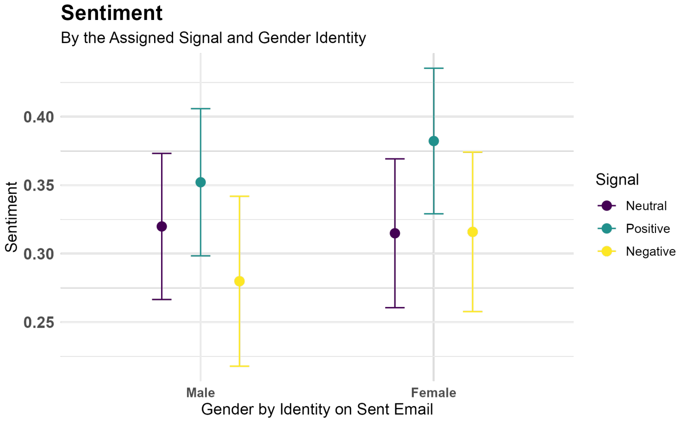
(b) By gender identity

Note: We use different regression to estimate the predictions in panels (a), (b), and (c). However, the estimates in each panel are derived from the same regression. In all three panels, we use state, week, and law enforcement agency type fixed effects and cluster the standard errors at the county level.

Figure H.8 Predicted **text sentiment** by the assigned signal, race or gender



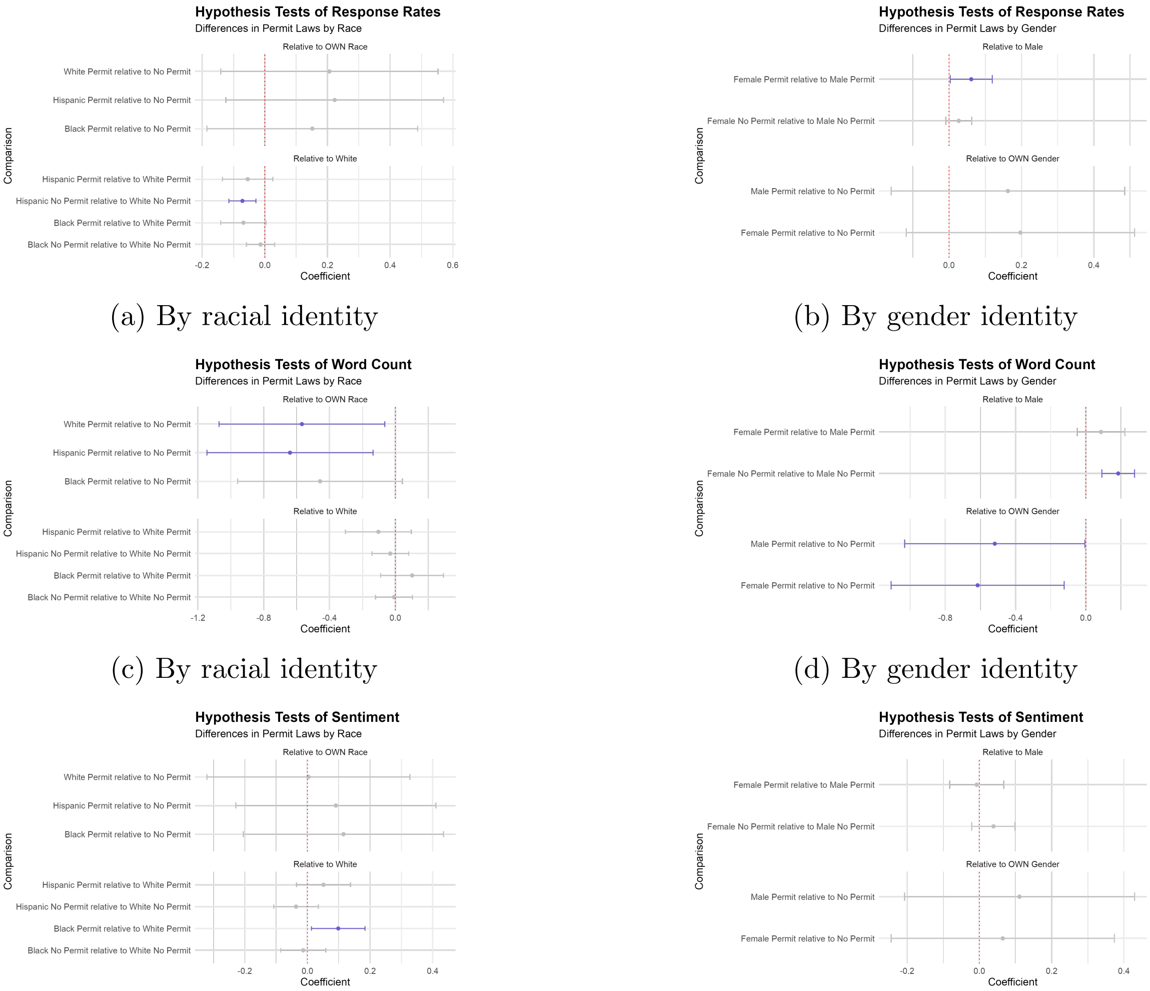
(a) By racial identity



(b) By gender identity

Note: The sentiment score is an index that ranges from -1 to 1, where -1 is the most negative and 1 is the most positive. We use different regression to estimate the predictions in panels (a), (b), and (c). However, the estimates in each panel are derived from the same regression. In all three panels, we use state, week, and law enforcement agency type fixed effects and cluster the standard errors at the county level.

Figure H.9 Heterogeneity of the results: US states with firearm permit laws



(a) By racial identity

(b) By gender identity

(c) By racial identity

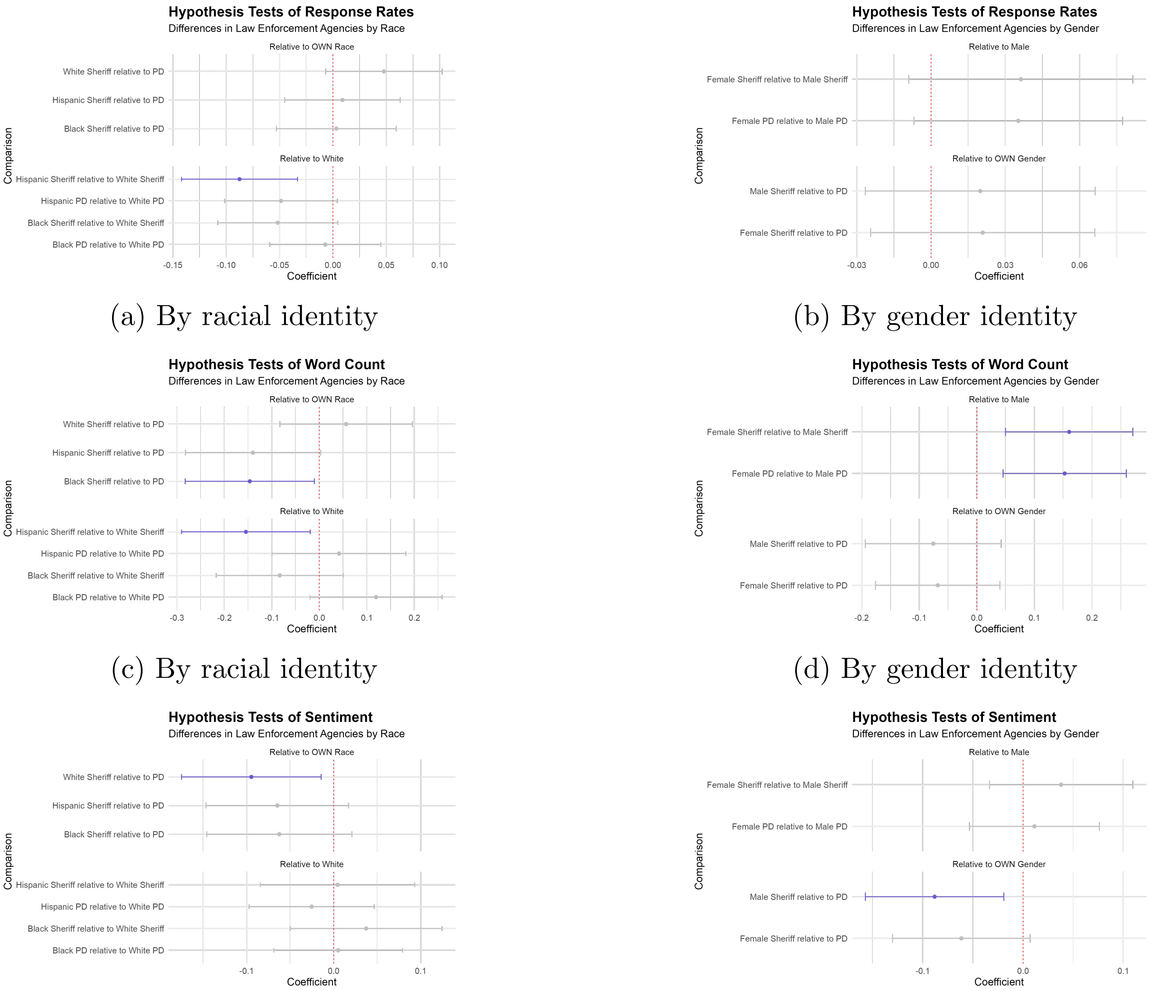
(d) By gender identity

(e) By racial identity

(f) By gender identity

Note: We use different regression to estimate the predictions in panels (a), (b), and (c). However, the estimates from each panel are derived from the same regression. In all three panels, we use state, week, and law enforcement agency type fixed effects and cluster the standard errors at the county level.

Figure H.10 Heterogeneity of the results: sheriff and police departments (PD)



(a) By racial identity

(b) By gender identity

(c) By racial identity

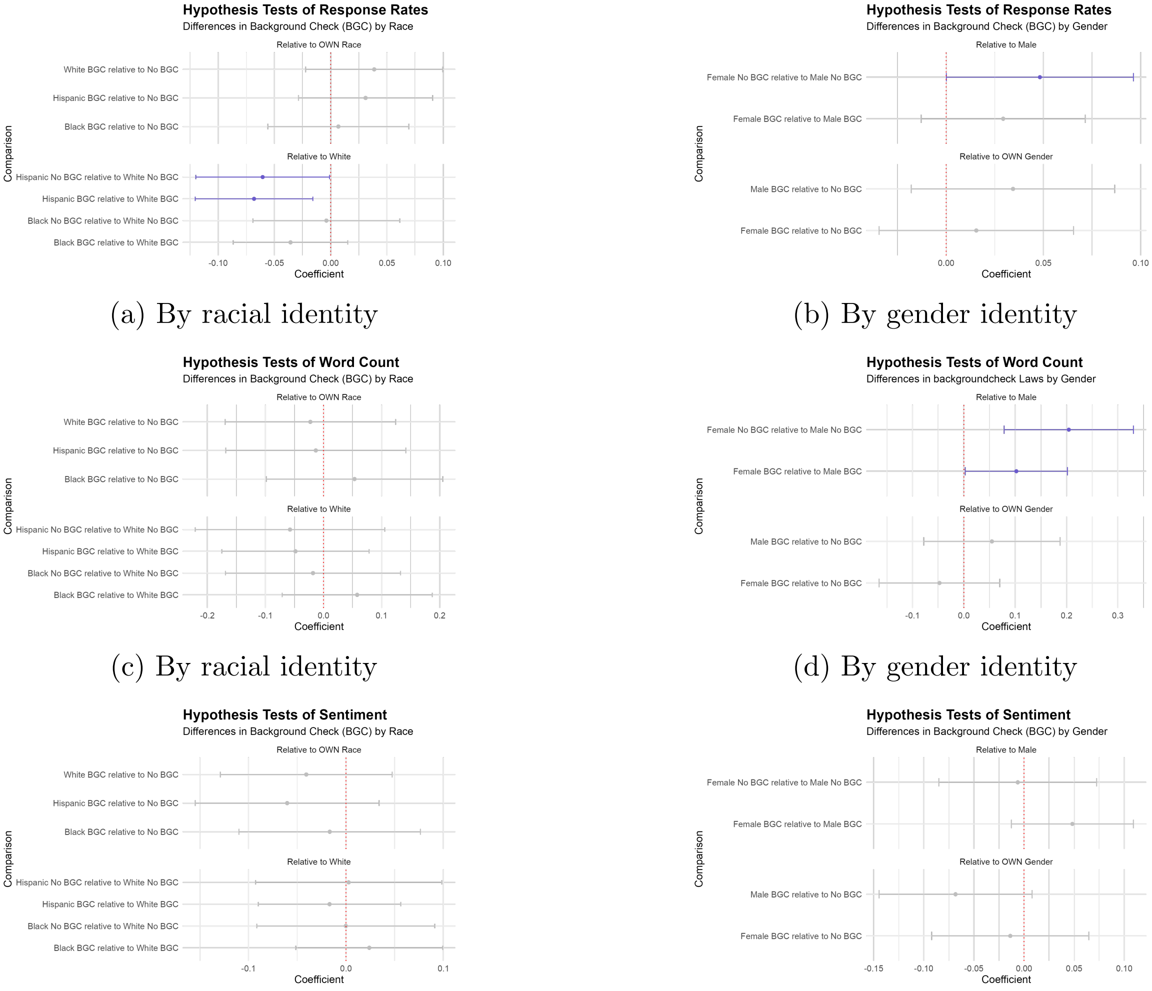
(d) By gender identity

(e) By racial identity

(f) By gender identity

Note: We use different regression to estimate the predictions in panels (a), (b), and (c). However, the estimates from each panel are derived from the same regression. In all three panels, we use state, week, and law enforcement agency type fixed effects and cluster the standard errors at the county level.

Figure H.11 Heterogeneity of Results: Law Enforcement with Background Checks



(a) By racial identity

(b) By gender identity

(c) By racial identity

(d) By gender identity

(e) By racial identity

(f) By gender identity

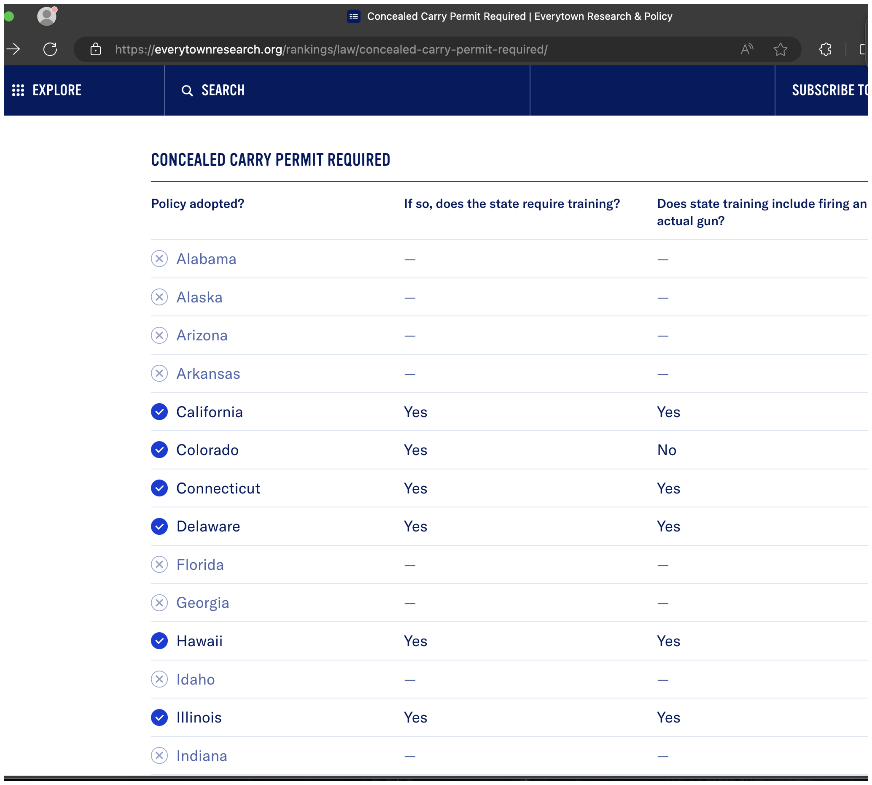
Note: We use different regression to estimate the predictions in panels (a), (b), and (c). However, the estimates from each panel are derived from the same regression. In all three panels, we use state, week, and law enforcement agency type fixed effects and cluster the standard errors at the county level.

APPENDIX I

CHAPTER 3 - US STATES THE REQUIRE CONCEALED FIREARM PERMITS

We use the website Every Town Research to identify states that require a concealed firearm permit. This data was pulled in March, 2024. Given that experiment ran at the end of 2023, we are comfortable using the US states listed in the figures below as the US states that require a firearm permit.

Figure I.1 US States with Concealed Firearm Permits

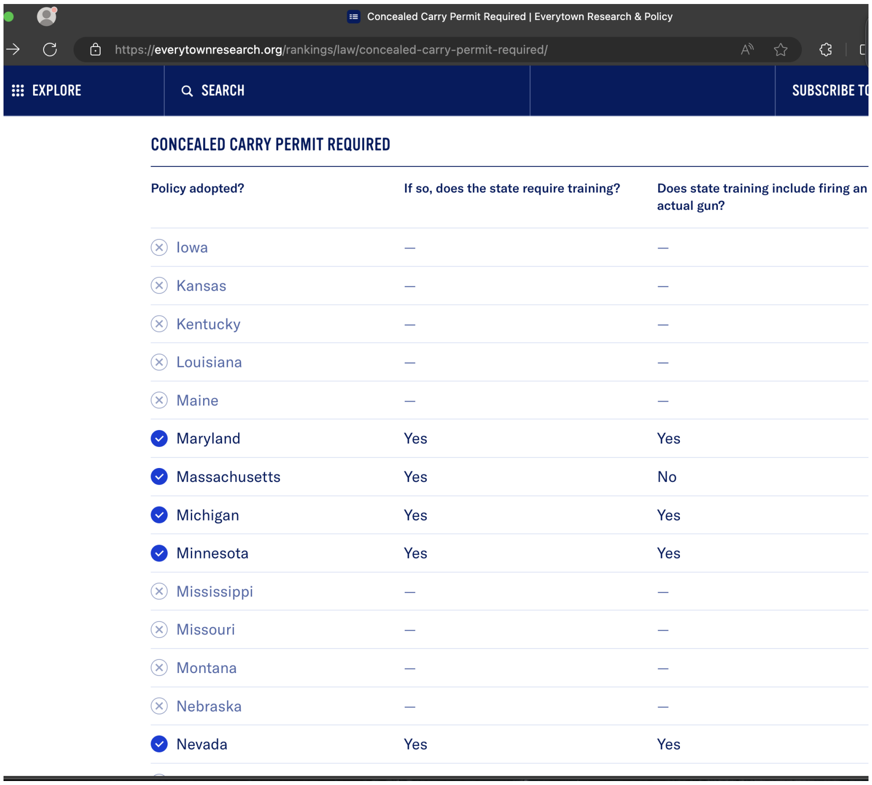


Concealed Carry Permit Required | Everytown Research & Policy
https://everytownresearch.org/rankings/law/concealed-carry-permit-required/
EXPLORE
SEARCH
SUBSCRIBE TO

CONCEALED CARRY PERMIT REQUIRED

Policy adopted?	If so, does the state require training?	Does state training include firing an actual gun?
⊗ Alabama	—	—
⊗ Alaska	—	—
⊗ Arizona	—	—
⊗ Arkansas	—	—
✓ California	Yes	Yes
✓ Colorado	Yes	No
✓ Connecticut	Yes	Yes
✓ Delaware	Yes	Yes
⊗ Florida	—	—
⊗ Georgia	—	—
✓ Hawaii	Yes	Yes
⊗ Idaho	—	—
✓ Illinois	Yes	Yes
⊗ Indiana	—	—

Figure I.2 US States with Concealed Firearm Permits

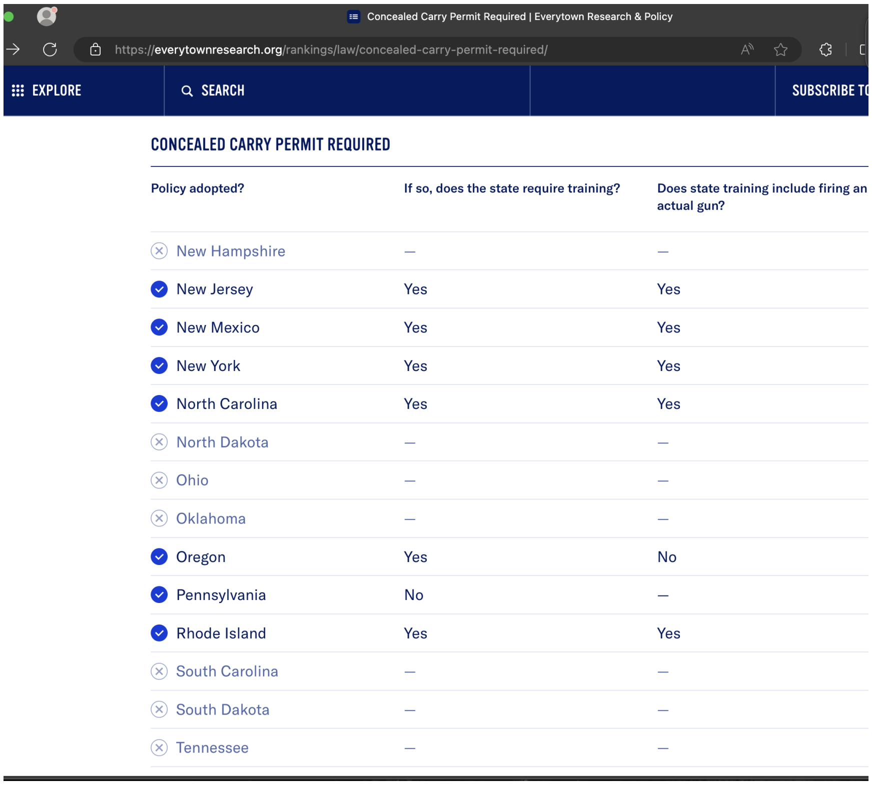


Concealed Carry Permit Required | Everytown Research & Policy
https://everytownresearch.org/rankings/law/concealed-carry-permit-required/
EXPLORE
SEARCH
SUBSCRIBE TO

CONCEALED CARRY PERMIT REQUIRED

Policy adopted?	If so, does the state require training?	Does state training include firing an actual gun?
⊗ Iowa	—	—
⊗ Kansas	—	—
⊗ Kentucky	—	—
⊗ Louisiana	—	—
⊗ Maine	—	—
✓ Maryland	Yes	Yes
✓ Massachusetts	Yes	No
✓ Michigan	Yes	Yes
✓ Minnesota	Yes	Yes
⊗ Mississippi	—	—
⊗ Missouri	—	—
⊗ Montana	—	—
⊗ Nebraska	—	—
✓ Nevada	Yes	Yes

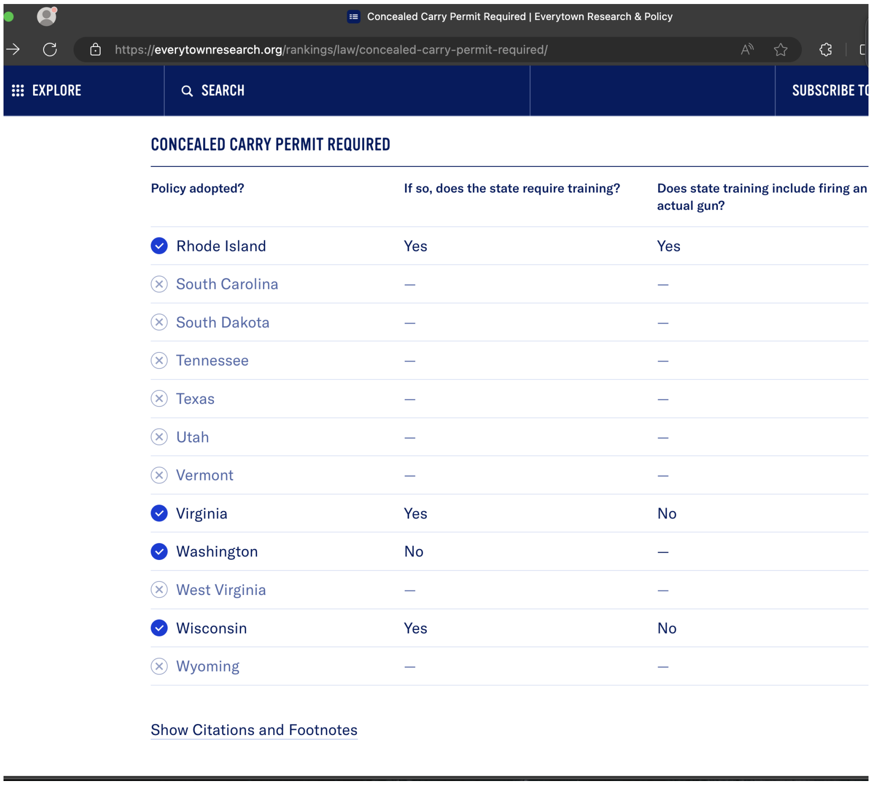
Figure I.3 US States with Concealed Firearm Permits


Concealed Carry Permit Required | Everytown Research & Policy
https://everytownresearch.org/rankings/law/concealed-carry-permit-required/
EXPLORE
SEARCH
SUBSCRIBE TO

CONCEALED CARRY PERMIT REQUIRED

Policy adopted?	If so, does the state require training?	Does state training include firing an actual gun?
✗ New Hampshire	—	—
✓ New Jersey	Yes	Yes
✓ New Mexico	Yes	Yes
✓ New York	Yes	Yes
✓ North Carolina	Yes	Yes
✗ North Dakota	—	—
✗ Ohio	—	—
✗ Oklahoma	—	—
✓ Oregon	Yes	No
✓ Pennsylvania	No	—
✓ Rhode Island	Yes	Yes
✗ South Carolina	—	—
✗ South Dakota	—	—
✗ Tennessee	—	—

Figure I.4 US States with Concealed Firearm Permits


Concealed Carry Permit Required | Everytown Research & Policy
https://everytownresearch.org/rankings/law/concealed-carry-permit-required/
EXPLORE
SEARCH
SUBSCRIBE TO

CONCEALED CARRY PERMIT REQUIRED

Policy adopted?	If so, does the state require training?	Does state training include firing an actual gun?
✓ Rhode Island	Yes	Yes
⊗ South Carolina	—	—
⊗ South Dakota	—	—
⊗ Tennessee	—	—
⊗ Texas	—	—
⊗ Utah	—	—
⊗ Vermont	—	—
✓ Virginia	Yes	No
✓ Washington	No	—
⊗ West Virginia	—	—
✓ Wisconsin	Yes	No
⊗ Wyoming	—	—

Show Citations and Footnotes

APPENDIX J

CHAPTER 4 - THE ALGORITHM

J.1 Mathematical Proof

I first outline the assumptions and the notation used in the algorithm before illustrating the mathematics of the algorithm in three different scenarios: (i) perfect information, (ii) imperfect information with non-rounded data, (iii) imperfect information with rounded data. Considering Google Trends data is rounded and limited, I first discuss how one would estimate panel data if they had access to the raw search volumes in order to layout the algorithm's framework. I delineate between rounded and non-rounded data to illustrate how there exists rounding errors and why the algorithm attempts attempts to estimate search volumes with minimal rounding errors. Finally, I discuss how to adjust the estimates to overcome any loss of temporal variation from using incomplete data with rounding errors in estimating the search volumes.

J.2 Assumptions

There are two main assumptions that I make in the algorithm:

1) The sum of the search volumes for each search term in each subregion equals the search volumes for each search term in the parent region at any time interval.

2) Given any point in time, the fraction of each subregion's total search volumes over the parent region's total search volumes are not necessarily equal.

J.3 Notation

Before I delve into the proof, I provide notation and definitions useful for reading through the proof:

Let n refer to any region (parent or subregion) such that $n \in \{p, j\}$, where p refers to the parent region and j refers to the parent region's subregion.

Search intensities are defined as z(n,t,k) $= \frac{\text{search volume for search term k }_{nt}}{\text{search volume for all search terms}_{nt}}$ in region n at time t for search term k.

- The maximum search intensity in region n across all t for search term k is: z(n,max,k) $=$ $\max_{t,k}$ $z(n, t, k)$.

- The maximum search intensity across all regions, n, at time t for search term k is: z(max,t,k) $=$ $\max_j$ $z(j, t, k)$.

- The maximum search intensity across all regions, n, time periods t, is defined as: $z(max, max, k) = \max_{n,t} z(n, t, k)$

- z(n,t,k) is unobserved for any region at any time period.

Then the relative search intensities that users retrieve from Google Trends is denoted by time series (TS), cross section (CS), and panel data (PL):

$$\textbf{TS(n,t,k)} = \frac{\text{z(n,t,k)}}{\text{z(n,max,k)}}$$
$$\textbf{CS(n,t,k)} = \frac{\text{z(n,t,k)}}{\text{z(max,t,k)}}$$
$$\textbf{PL(n,t,k)} = \frac{\text{z(n,t,k)}}{\text{z(max,max,k)}}$$

J.4 Perfect Information

section4

I first evaluate the relationship between the subregions' cross sectional values, $CS(j, t, k)$, and the parent time series $TS(p, t, k)$ given perfection information.

Suppose you have the relative cross sectional search intensities across all time periods, search terms and subregions, $CS(j, t, k)$; the parent region's relative time series search intensities, $TS(p, t, k)$, and the total number of searches in each region. You can then construct the proportion of each subregion's total searches to the parent country's total searches, with the variable $w(j, t)$, defined as:

$$w(j, t) = \frac{\text{Total searches of subregion}_{jt}}{\text{Total searches of parent region}_t}$$

w(j,t) does not depend on the search term, because it's the fraction of all searches made in the subregion relative to the parent region. Assuming that the sum of the subregions' search volumes must equal the parent region's search volumes, one can show that sum of the $w(t, j)$s, multiplied by each subregion's search intensity, $z(j, t, k)$, is equal to the parent region's search intensity, $z(p, t, k)$ at time t:

$$\textbf{(1)} \quad \Sigma_j w(t, j) \times z(j, t, k) = z(p, t, k)$$

I reverse engineer the indexing process to transform the search intensities in (1) into the indexed search volumes from Google Trends in terms of C(j,t,k), the subregions' cross-sectional search volumes, and TS(p,t,k), the parent region's time series:

$$\textbf{(2)} \quad \Sigma_j w(t, j) \times CS(j, t, k) \times z(max, t, k) = TS(p, t, k) \times z(p, max, k)$$

Dividing (2) by z(p,max,k) allows the construction of the subregions' search intensities as a proportion to the parent region's maximum search intensity in the given time window, which I define as $x(j, t, k)$:

$$\textbf{(3)} \quad x(j, t, k) = \frac{CS(j, t, k) \times TS(p, t, k)}{\Sigma_j w(j, t, k) \times CS(j, t, k)}$$

$x(j, t, k)$ is the main value of interest as it is proportional to the subregions' actual search intensities: $x(j, t, k) = \frac{z(j,t,k)}{z(p,max,k)}$. However, $x(j, t, k)$ requires

$w(j,t)$, which values are not publicly available. Estimates of $x(j,t,k)$ would suffer from measurement error in how much $x(j,t,k)$ changes over time. I show how to overcome this by adjusting the estimates of $x(j,t,k)$ with temporal variation from each individual subregion's time series, TS(j,t,k).

Imperfect information with non-rounded data: identifying the adjusting factor γ_t

I first discuss how the algorithm overcomes the measurement error from not having access to $w(j,t)$ with non-rounded Google Trends search volumes[1]

Let $\widehat{x(j,t,k)}$ be a naive estimate of $x(j,t,k)$, such that

$$\widehat{x(j,t,k)} = \frac{CS(j,t,k) \times TS(p,t,k)}{\Sigma_j CS(j,t,k)}$$

$\widehat{x(j,t,k)}$ does not equal $x(j,t,k)$, as $\frac{x(j,t,k)}{\widehat{x(j,t,k)}} = \frac{\Sigma_j CS(j,t,k)}{\Sigma_j w(j,t) \times CS(j,t,k)}$

For any time period, t, the value of $\frac{x(j,t,k)}{\widehat{x(j,t,k)}}$ is equal to some constant constant c_t, which can be though of as measurement error. Because of the lack of data, calculating c_t itself is not possible. However, using the ratios of the measurement errors at time t and some fixed time period, l is possible, which we define $\frac{c_t}{c_l} = \gamma_t$ as the adjustment factor.

γ_t is used to readjust $\widehat{x(j,t,k)}$ with respect to the $l'th$ period, so that $\widehat{x(j,t,k)} \times \gamma_t = \widehat{x(j,t,k)} \times \frac{c_t}{c_l}$ is normalized with respect to c_l, since c_l is a fixed value for any t. Further, indexing an adjusted estimates with respect to the maximum search intensity cancels out c_l, so the remaining value is respect to $\widehat{x(j,t,k)} \times c_t$.

To calculate γ_t, it only requires only the individual time series values for each subregion at time t and l: $TS(j,t,k)$ and $TS(j,l,k)$. This is because $\frac{x(j,t,k)}{x(j,l,k)} =$

[1] As a reminder, when Google indexes the search volumes, they round the values to the nearest whole number. This may seem innocent, but actually causes rounding errors in the estimates. This is discussed in the next section

$\frac{TS(j,t,k)}{TS(j,l,k)} = \frac{\widehat{x(j,t,k)}\times c_t}{\widehat{x(j,l,k)}\times c_l}$. Thus, the adjustment factor, γ_t is the ratio of the individual subregion's time series.

In fact, one can show that $\widehat{x(j,t,k)} \times \gamma_t$ is proportional to the subregion's search intensity, $z(j,t,k)$, which can be used to estimate Google Trends time series, cross-sectional, and panel data without any measurement error:

$$
\begin{aligned}
TS^{est}(j,t,k) &= \frac{\widehat{x(j,t,k)} \times \gamma_t}{\widehat{x(j,max,k)} \times \gamma_{max,k}} = \frac{z(j,t,k)}{z(j,max,k)} = TS(j,t,k)\\
CS^{est}(j,t,k) &= \frac{\widehat{x(j,t,k)} \times \gamma_t}{\widehat{x(max,t,k)} \times \gamma_t} = \frac{z(j,t,k)}{z(max,t,k)} = CS(j,t,k)\\
PL^{est}(j,t,k) &= \frac{\widehat{x(j,t,k)} \times \gamma_t}{\widehat{x(max,max,k)} \times \gamma_{max}} = \frac{z(j,t,k)}{z(max,max,k)} = PL(j,t,k)
\end{aligned}
$$

¡!–By adjusting $\widehat{x(j,t,k)}$, estimates of $TS(j,t,k)$ and $CS(j,t,k)$ are equal to the actual $TS(j,t,k)$ and $CS(j,t,k)$, since c_{lk} is constant for any period t, search term k, and across all j subregions.–¿

J.5 Imperfect information with rounded data: identifying the adjusting Factor γ_{jt}

rounded$_{p}roof$

However, Google Trends rounds the relative search intensities, so adjusting the naive estimates with the γ contains rounding errors. I denote the rounded Google Trend data as $CS^r(n,t,k)$ and $TS^r(n,t,k)$. The true non-rounded values are within one unit of $CS(n,t)$ and $TS(n,t)$.

$$
\begin{aligned}
CS(n,t,k) &\in (CS^r(n,t,k) - \frac{1}{2}, CS^r(n,t,k) + \frac{1}{2})\\
TS(n,t,k) &\in (TS^r(n,t,k) - \frac{1}{2}, TS^r(n,t,k) + \frac{1}{2})
\end{aligned}
$$

The difference between the rounded and nonrounded data can also be written with multiplicative errors so that $CS(n,t) = CS^r(n,t) \times \epsilon^{CSr}_{nt}$ and $TS(n,t) = TS^r(n,t) \times \epsilon^{TSr}_{nt}$. When $\widehat{x(j,t,k})$ is estimated with rounded data, the estimate is erred by:

$$x^r\widehat{(j,t,k})= \frac{\frac{CS(j,t,k)}{\epsilon^{CSr}_{jtk}} \times \frac{(TS(p,t,k)}{\epsilon^{rp}_{tk}}}{\Sigma_j \frac{CS(j,t,k)}{\epsilon^{CSr}_{jtk}}}$$

Given the naive estimate, $x^r(j,t,k)$, the adjusted estimate $x^r\widetilde{(j,t,k})$, is calculated by multiplying $x^r\widehat{(j,t,k})$ by γ_{jt}. However, γ_{jt} is not the same as γ_t from the previous section.

In fact, γ_{jt} is equivalent to the following ratio of the rounded time series search volumes:

$$\begin{aligned}\gamma_{jt} &= \frac{c_{jt}}{c_{jl}} \\ &= \frac{x^r(j,l,k)}{x^r(j,t,k)} \times \frac{TS^r(j,t,k) \times \epsilon^{TSr}_{jtk}}{TS^r(j,l,k) \times \epsilon^{TS^r}_{jkl}}\end{aligned}$$

Notice that if there are no rounding errors, the ϵs here would equal one and $\gamma_{jt} = \gamma_t$. However, because all of the ϵs do not necessarily equal to one, γ_{jt} is likely different for each subregion at time t.

J.6 Calculating the adjusted estimate given imperfect information and rounded data

To minimize the rounding error generated from γ_{jt}, I reconstructed the adjusted estimate, $x^r\widetilde{(j,t,k})$ assuming that there exists a value in $x^r\widetilde{(j,t,k})$ that is proportional to $\Sigma w(j,t)CS(j,t,k)$ plus some rounding error, ϵ_{jt}:

$$(1) \quad x^r\widetilde{(j,t,k}) = \frac{CS^r(j,t,k) \times TS^r(p,t,k)}{\Sigma_j w^{prop}(j,t) CS^{prop}(j,t,k) + \epsilon_{jt}}$$

Rearranging $\widetilde{x^r(j,t,k)}$ with respect to to (1) in terms of $\Sigma_j w^{prop}(j,t)CS^{prop}(j,t,k)+ \epsilon_{jt}$, I use the adjusted estimates calculated by multiplying the naive estimates with γ_{jt} to then calculate equation (2):

$$(2)\ \Sigma_j w^{prop}(j,t)CS^{prop}(j,t,k) + \epsilon_{jt} = \frac{CS^r(j,t,k) \times TS^r(p,t,k)}{\widetilde{x^r(j,t,k)}}$$

.

Assuming that ϵ_{jt} is mean zero, then the mean of $\Sigma_j w(j,t)^{prop} CS^{prp}(j,t,k)+ \epsilon_{jt}$ is equal to $\Sigma_j w^{prop}(j,t)CS^{prop}(j,t,k)$. This is because for each subregion j at time t, the value of the denominators differs only by ϵ_{jt}.

I took the value calculated for $\Sigma_j w^{prop}(j,t)CS^{prop}(j,t,k)$ and re-calculated $\widetilde{x^r(j,t,k)}$ with it:

$$(3)\ \widetilde{x^{r,new}(j,t,k)} = \frac{CS^r(j,t,k) \times TS^r(p,t,k)}{\Sigma_j w^{prop}(j,t)CS^{prop}(j,t,k)}$$

The adjusted estimates,$\widetilde{x^{r,new}(j,t,k)}$, from (3) are the ones used to construct the panel data from the subregions' cross-sectional search volumes.

REFERENCES CITED

Ahn, T., Arcidiacono, P., Hopson, A., & Thomas, J. (2019, December). *Equilibrium Grade Inflation with Implications for Female Interest in STEM Majors* (Tech. Rep. No. w26556). Cambridge, MA: National Bureau of Economic Research. Retrieved 2023-09-07, from `http://www.nber.org/papers/w26556.pdf` doi: 10.3386/w26556

Alesina, A., Carlana, M., Ferrara, E. L., & Pinotti, P. (2018, December). *Revealing Stereotypes: Evidence from Immigrants in Schools* (Tech. Rep. No. w25333). Cambridge, MA: National Bureau of Economic Research. Retrieved 2023-03-06, from `http://www.nber.org/papers/w25333.pdf` doi: 10.3386/w25333

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2016, June). Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition. *American Economic Review*, *106*(6), 1437–1475. Retrieved 2023-02-22, from `https://pubs.aeaweb.org/doi/10.1257/aer.20140571` doi: 10.1257/aer.20140571

Becker, G. S. (1971). *The Economics of Discrimination.* University of Chicago Press. Retrieved 2023-09-07, from `http://www.bibliovault.org/BV.landing.epl?ISBN=9780226041162` doi: 10.7208/chicago/9780226041049.001.0001

Bertrand, M., & Duflo, E. (2017, January). Chapter 8 - Field Experiments on Discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (Vol. 1, pp. 309–393). North-Holland. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2214658X1630006X` doi: 10.1016/bs.hefe.2016.08.004

Bertrand, M., & 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. doi: 10.4324/9781003071709-20

Bettinger, E. P., Fox, L., Loeb, S., & Taylor, E. S. (2017, September). Virtual Classrooms: How Online College Courses Affect Student Success. *American Economic Review*, *107*(9), 2855–2875. Retrieved 2023-10-17, from `https://pubs.aeaweb.org/doi/10.1257/aer.20151193` doi: 10.1257/aer.20151193

Biasi, B. (2021, August). The Labor Market for Teachers under Different Pay Schemes. *American Economic Journal: Economic Policy*, *13*(3), 63–102. Retrieved 2023-11-16, from https://pubs.aeaweb.org/doi/10.1257/pol.20200295 doi: 10.1257/pol.20200295

Bohren, J. A., Imas, A., & Rosenberg, M. (2019, October). The Dynamics of Discrimination: Theory and Evidence. *American Economic Review*, *109*(10), 3395–3436. Retrieved 2023-09-07, from https://pubs.aeaweb.org/doi/10.1257/aer.20171829 doi: 10.1257/aer.20171829

Breda, T., & Ly, S. T. (2015, October). Professors in Core Science Fields Are Not Always Biased against Women: Evidence from France. *American Economic Journal: Applied Economics*, *7*(4), 53–75. Retrieved 2023-09-07, from https://pubs.aeaweb.org/doi/10.1257/app.20140022 doi: 10.1257/app.20140022

Bulman, G. (2019). LAW ENFORCEMENT LEADERS AND THE RACIAL COMPOSITION OF ARRESTS. *Economic Inquiry*, *57*(4), 1842–1858. (Publisher: Blackwell Publishing Inc.) doi: 10.1111/ecin.12800

Butler, D. M., & Broockman, D. E. (2011, July). Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*, *55*(3), 463–477. doi: 10.1111/j.1540-5907.2011.00515.x

Cacault, M. P., Hildebrand, C., Laurent-Lucchetti, J., & Pellizzari, M. (2021, August). Distance Learning in Higher Education: Evidence from a Randomized Experiment. *Journal of the European Economic Association*, *19*(4), 2322–2372. Retrieved 2023-10-17, from https://academic.oup.com/jeea/article/19/4/2322/6067382 doi: 10.1093/jeea/jvaa060

Carlana, M. (2019, August). Implicit Stereotypes: Evidence from Teachers' Gender Bias*. *The Quarterly Journal of Economics*, *134*(3), 1163–1224. Retrieved 2023-03-02, from https://academic.oup.com/qje/article/134/3/1163/5368349 doi: 10.1093/qje/qjz008

Carlson, J. (2020, October). Gun Studies and the Politics of Evidence. *Annual Review of Law and Social Science*, *16*(1), 183–202. Retrieved 2023-04-24, from https://www.annualreviews.org/doi/10.1146/annurev-lawsocsci-020620-111332 doi: 10.1146/annurev-lawsocsci-020620-111332

Chen, M. K., Christensen, K. L., John, E., Owens, E., & Zhuo, Y. (2021, September). Smartphone Data Reveal Neighborhood-Level Racial Disparities in Police Presence. Retrieved 2022-12-09, from `https://arxiv.org/abs/2109.12491v2` doi: 10.48550/arXiv.2109.12491

de Brey, C., Musu, L., McFarland, J., Wilkinson-Flicker, S., Diliberti, M., Zhang, A., ... Wang, X. (2019, February). *Status and Trends in the Education of Racial and Ethnic Groups 2018.* U.S Department of Education. Retrieved from `https://nces.ed.gov/pubs2019/2019038.pdf`

Doornkamp, L., Van Der Pol, L., Groeneveld, S., Mesman, J., Endendijk, J., & Groeneveld, M. (2022, October). Understanding gender bias in teachers' grading: The role of gender stereotypical beliefs. *Teaching and Teacher Education*, *118*, 103826. Retrieved 2023-09-07, from `https://linkinghub.elsevier.com/retrieve/pii/S0742051X22002001` doi: 10.1016/j.tate.2022.103826

Duflo, E., Hanna, R., & Ryan, S. P. (2012, June). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, *102*(4), 1241–1278. Retrieved 2023-11-16, from `https://pubs.aeaweb.org/doi/10.1257/aer.102.4.1241` doi: 10.1257/aer.102.4.1241

Edwards, F., Lee, H., & Esposito, M. (2019, August). Risk of being killed by police use of force in the United States by age, race–ethnicity, and sex. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(34), 16793–16798. (Publisher: National Academy of Sciences) doi: 10.1073/pnas.1821204116

Einstein, K. L., & Glick, D. M. (2017, January). Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing. *American Journal of Political Science*, *61*(1), 100–116. (Publisher: Blackwell Publishing Ltd) doi: 10.1111/ajps.12252

Ewens, M., Tomlin, B., & Wang, L. C. (2014, March). Statistical Discrimination or Prejudice? A Large Sample Field Experiment. *The Review of Economics and Statistics*, *96*(1), 119–134. Retrieved 2024-03-30, from `https://direct.mit.edu/rest/article/96/1/119/58118/Statistical-Discrimination-or-Prejudice-A-Large` doi: 10.1162/REST_a_00365

Feld, J., Salamanca, N., & Hamermesh, D. S. (2016, August). Endophilia or Exophobia: Beyond Discrimination. *The Economic Journal*, *126*(594), 1503–1527. Retrieved 2023-10-17, from `https://academic.oup.com/ej/article/126/594/1503-1527/5078052` doi: 10.1111/ecoj.12289

Fridell, L. A. (2017, September). Explaining the Disparity in Results Across Studies Assessing Racial Disparity in Police Use of Force: A Research Note. *American Journal of Criminal Justice*, *42*(3), 502–513. Retrieved 2022-07-25, from `https://link.springer.com/article/10.1007/s12103-016-9378-y` (Publisher: Springer New York LLC) doi: 10.1007/s12103-016-9378-y

Fryer, R. G. (2020). An empirical analysis of racial differences in police use of force: a response. *Journal of Political Economy*, *128*(10), 4003–4008. Retrieved 2022-07-25, from `http://www.fatalencounters.org` doi: 10.1086/710977

Gabaix, X. (2019). Behavioral inattention. In *Handbook of Behavioral Economics: Applications and Foundations 1* (Vol. 2, pp. 261–343). Elsevier. Retrieved 2023-09-07, from `https://linkinghub.elsevier.com/retrieve/pii/S2352239918300216` doi: 10.1016/bs.hesbe.2018.11.001

Gaddis, S. M. (2017a, September). How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science*, *4*, 469–489. (Publisher: Society for Sociological Science) doi: 10.15195/v4.a19

Gaddis, S. M. (2017b, October). Racial/Ethnic Perceptions from Hispanic Names: Selecting Names to Test for Discrimination. *Socius: Sociological Research for a Dynamic World*, *3*, 237-267. Retrieved 2022-03-09, from `https://journals.sagepub.com/doi/10.1177/2378023117737193` (Publisher: SAGE PublicationsSage CA: Los Angeles, CA) doi: 10.1177/2378023117737193

Gelman, A., Fagan, J., & Kiss, A. (2007). An analysis of the New York City police department's "stop-and- frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, *102*(479), 813–823. doi: 10.1198/016214506000001040

Generated Photos. (2023). *Academic dataset by Generated Photos.* Retrieved from `https://generated.photos/datasets`

Giulietti, C., Tonin, M., & Vlassopoulos, M. (2019). Racial discrimination in local public services: A field experiment in the United States. *Journal of the European Economic Association*, *17*(1), 165–204. doi: 10.1093/jeea/jvx045

Goncalves, F., & Mello, S. (2021, May). A few bad apples? racial bias in policing. *American Economic Review*, *111*(5), 1406–1441. (Publisher: American Economic Association) doi: 10.1257/AER.20181607

Hanna, R. N., & Linden, L. L. (2012, November). Discrimination in Grading. *American Economic Journal: Economic Policy*, *4*(4), 146–168. Retrieved 2023-03-17, from `https://pubs.aeaweb.org/doi/10.1257/pol.4.4.146` doi: 10.1257/pol.4.4.146

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2011, August). Are boys discriminated in Swedish high schools? *Economics of Education Review*, *30*(4), 682–690. Retrieved 2023-08-30, from `https://linkinghub.elsevier.com/retrieve/pii/S0272775711000227` doi: 10.1016/j.econedurev.2011.02.007

Huang, B., Li, J., Lin, T.-C., Tai, M., & Zhou, Y. (2021). Attention Discrimination under Time Constraints: Evidence from Retail Lending. *SSRN Electronic Journal*. Retrieved 2023-09-07, from `https://www.ssrn.com/abstract=3865478` doi: 10.2139/ssrn.3865478

Jansson, J., & Tyrefors, B. (2022, October). Grading bias and the leaky pipeline in economics: Evidence from Stockholm University. *Labour Economics*, *78*, 102212. Retrieved 2023-08-30, from `https://linkinghub.elsevier.com/retrieve/pii/S0927537122001038` doi: 10.1016/j.labeco.2022.102212

Kaplan, J. (2023a). *Jacob Kaplan's Concatenated Files: National Incident-Based Reporting System (NIBRS) Data, 1991-2021.* ICPSR - Interuniversity Consortium for Political and Social Research. Retrieved 2023-10-03, from `https://www.openicpsr.org/openicpsr/project/118281/version/V8/view` doi: 10.3886/E118281V8

Kaplan, J. (2023b). *Jacob Kaplan's Concatenated Files: Uniform Crime Reporting Program Data: Law Enforcement Officers Killed and Assaulted (LEOKA) 1960-2021.* ICPSR - Interuniversity Consortium for Political and Social Research. Retrieved 2023-10-03, from `https://www.openicpsr.org/openicpsr/project/102180/version/V12/view` doi: 10.3886/E102180V12

Knox, D., Lowe, W., & Mummolo, J. (2020). Administrative Records Mask Racially Biased Policing. *American Political Science Review*, *114*(3), 619–637. doi: 10.1017/S0003055420000039

Kofoed, M., Gebhart, L., Gilmore, D., & Moschitto, R. (2021). Zooming to Class?: Experimental Evidence on College Students&Apos; Online Learning During Covid-19. *SSRN Electronic Journal*. Retrieved 2023-09-07, from `https://www.ssrn.com/abstract=3846700` doi: 10.2139/ssrn.3846700

Lavy, V. (2008, October). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, *92*(10-11), 2083–2105. Retrieved 2023-09-08, from `https://linkinghub.elsevier.com/retrieve/pii/S0047272708000418` doi: 10.1016/j.jpubeco.2008.02.009

Lavy, V., & Megalokonomou, R. (2019, June). *Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study* (Tech. Rep. No. w26021). Cambridge, MA: National Bureau of Economic Research. Retrieved 2023-03-06, from `http://www.nber.org/papers/w26021.pdf` doi: 10.3386/w26021

Lavy, V., & Sand, E. (2018, November). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, *167*, 263–279. Retrieved 2023-09-08, from `https://linkinghub.elsevier.com/retrieve/pii/S0047272718301750` doi: 10.1016/j.jpubeco.2018.09.007

Leaver, C., Ozier, O., Serneels, P., & Zeitlin, A. (2021, July). Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools. *American Economic Review*, *111*(7), 2213–2246. Retrieved 2023-11-17, from `https://pubs.aeaweb.org/doi/10.1257/aer.20191972` doi: 10.1257/aer.20191972

Luh, E. (2020). Not So Black and White: Uncovering Racial Bias from Systematically Masked Police Reports. *SSRN Electronic Journal*. Retrieved 2022-07-25, from `https://cdn1.sph.harvard.edu/wp-content/uploads/sites/94/2018/01/NPR-RWJF-HSPH-` doi: 10.2139/ssrn.3357063

Makowsky, M. D., Stratmann, T., & Tabarrok, A. (2019, January). To serve and collect: The fiscal and racial determinants of law enforcement. *Journal of Legal Studies*, *48*(1), 189–216. (Publisher: University of Chicago Press) doi: 10.1086/700589

Maćkowiak, B., Matějka, F., & Wiederholt, M. (2023, March). Rational Inattention: A Review. *Journal of Economic Literature*, *61*(1), 226–273. Retrieved 2023-03-05, from `https://pubs.aeaweb.org/doi/10.1257/jel.20211524` doi: 10.1257/jel.20211524

Mehic, A. (2022, October). Student beauty and grades under in-person and remote teaching. *Economics Letters*, *219*, 110782. Retrieved 2023-09-07, from `https://linkinghub.elsevier.com/retrieve/pii/S016517652200283X` doi: 10.1016/j.econlet.2022.110782

Nguyen, Nhu, Ost, Ben, & Qureshi, Javaeria. (2023). OK Boomer: Generational Differences in Teacher Quality. Retrieved 2023-09-07, from `https://edworkingpapers.com/ai23-831` (Publisher: edworkingpapers.com) doi: 10.26300/XFEB-CA25

Nix, J., Campbell, B. A., Byers, E. H., & Alpert, G. P. (2017). A Bird's Eye View of Civilians Killed by Police in 2015: Further Evidence of Implicit Bias. *Criminology and Public Policy*, *16*(1), 309–340. doi: 10.1111/1745-9133.12269

Oberfield, Z. W., & Incantalupo, M. B. (2021, November). Racial Discrimination and Street-Level Managers: Performance, Publicness, and Group Bias. *Public Administration Review*, *81*(6), 1055–1070. (Publisher: John Wiley and Sons Inc) doi: 10.1111/puar.13376

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021, September). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. Retrieved 2023-09-07, from `https://link.springer.com/10.3758/s13428-021-01694-3` doi: 10.3758/s13428-021-01694-3

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

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., . . . Goel, S. (2020, May). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, *4*(7), 736–745. Retrieved 2022-02-02, from `https://www.nature.com/articles/s41562-020-0858-1` (arXiv: 1706.05678 Publisher: Nature Publishing Group) doi: 10.1038/s41562-020-0858-1

Protivínský, T., & Münich, D. (2018, December). Gender Bias in teachers' grading: What is in the grade. *Studies in Educational Evaluation*, *59*, 141–149. Retrieved 2023-09-08, from `https://linkinghub.elsevier.com/retrieve/pii/S0191491X17302584` doi: 10.1016/j.stueduc.2018.07.006

Ross, C. T. (2015, November). A multi-level Bayesian analysis of racial Bias in police shootings at the county-level in the United States, 2011-2014. *PLoS ONE*, *10*(11), e0141854. Retrieved 2022-07-25, from `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141854` (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0141854

Ross, C. T., Winterhalder, B., & McElreath, R. (2018, June). Resolution of apparent paradoxes in the race-specific frequency of use-of-force by police. *Palgrave Communications*, *4*(1), 1–9. Retrieved 2022-07-25, from `https://www.nature.com/articles/s41599-018-0110-z` (Publisher: Palgrave) doi: 10.1057/s41599-018-0110-z

Rudolph, K. E., Stuart, E. A., Vernick, J. S., & Webster, D. W. (2015). Association between connecticut's permit-to-purchase handgun law and homicides. *American Journal of Public Health*, *105*(8), e49–e54.

Sances, M. W., & You, H. Y. (2017). Who pays for government? descriptive representationb and exploitative revenue sources. *Journal of Politics*, *79*(3), 1090–1094. doi: 10.1086/691354

Schwartzstein, J. (2014, December). SELECTIVE ATTENTION AND LEARNING: Selective Attention and Learning. *Journal of the European Economic Association*, *12*(6), 1423–1452. Retrieved 2023-10-21, from `https://academic.oup.com/jeea/article-lookup/doi/10.1111/jeea.12104` doi: 10.1111/jeea.12104

Shi, Y., & Zhu, M. (2023, April). "Model minorities" in the classroom? Positive evaluation bias towards Asian students and its consequences. *Journal of Public Economics*, *220*, 104838. Retrieved 2023-08-30, from `https://linkinghub.elsevier.com/retrieve/pii/S0047272723000208` doi: 10.1016/j.jpubeco.2023.104838

Sims, C. A. (2003, April). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690. Retrieved 2023-09-07, from `https://linkinghub.elsevier.com/retrieve/pii/S0304393203000291` doi: 10.1016/S0304-3932(03)00029-1

Smith, M. R., Rojek, J. J., Petrocelli, M., & Withrow, B. (2017). Measuring disparities in police activities: a state of the art review. *Policing*, *40*(2), 166–183. Retrieved from `https://heinonline.org/HOL/License` doi: 10.1108/PIJPSM-06-2016-0074

Stanford, G. (2023). Police are less likely to respond to requests for help from minorities: Field experiment evidence of police discrimination. *Working Paper*. Retrieved from `https://www.gstanford.org/papers/Stanford_police_1-25-23.pdf`

Steidley, T., & Yamane, D. (2022, February). Special Issue Editors' Introduction: A Sociology of Firearms for the Twenty-First Century. *Sociological Perspectives*, *65*(1), 5–11. Retrieved 2023-04-12, from `https://doi.org/10.1177/07311214211040933` (Publisher: SAGE Publications Inc) doi: 10.1177/07311214211040933

Stephens-Davidowitz, S. (2014, October). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, *118*, 26–40. Retrieved 2024-04-15, from https://linkinghub.elsevier.com/retrieve/pii/S0047272714000929 doi: 10.1016/j.jpubeco.2014.04.010

Stroube, B. K. (2021, September). Using allegations to understand selection bias in organizations: Misconduct in the Chicago Police Department. *Organizational Behavior and Human Decision Processes*, *166*, 149–165. Retrieved 2022-12-09, from https://www.sciencedirect.com/science/article/pii/S074959781830339X doi: 10.1016/j.obhdp.2020.03.003

Terrier, C. (2020, August). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, *77*, 101981. Retrieved 2023-03-06, from https://linkinghub.elsevier.com/retrieve/pii/S0272775718307714 doi: 10.1016/j.econedurev.2020.101981

Tversky, A., & Kahneman, D. (1974, September). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. Retrieved 2023-09-08, from https://www.science.org/doi/10.1126/science.185.4157.1124 doi: 10.1126/science.185.4157.1124

U.S. Census Bureau. (2021). *Selected Housing Characteristics 2016-2021 American Community Survey 5-year estimates.* Retrieved from https://www.census.gov/data/developers/data-sets/acs-5year/2021.html

Vizzard, W. J. (2015). The Current and Future State of Gun Policy in the United States. *Journal of Criminal Law & Criminology*, *104*(4), 879–904. Retrieved 2023-04-12, from http://libproxy.uoregon.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=110668173&site=ehost-live&scope=site (Publisher: Northwestern University School of Law)

Webster, D., Crifasi, C. K., & Vernick, J. S. (2014). Effects of the repeal of missouri's handgun purchaser licensing law on homicides. *Journal of Urban Health*, *91*, 293–302.

Weisburst, E. K. (2019). Police Use of Force as an Extension of Arrests: Examining Disparities across Civilian and Officer Race. *AEA Papers and Proceedings*, *109*, 152–156. doi: 10.1257/pandp.20191028

West, J. (2018). Racial Bias in Police Investigations. (October), 1–36.

White, A. R., Nathan, N. L., & Faller, J. K. (2015). What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials. *American Political Science Review*, *109*(1), 129–142. Retrieved 2022-07-22, from `https://doi.org/10.1017/S0003055414000562` doi: 10.1017/S0003055414000562

Williams Jr, M. C. (2020). Gun violence in black and white: Evidence from policy reform in missouri. *Unpublished Manuscript, NYU*.