Hierarchical Audit Design: The Benefits of Increasing Transactional Density in Audit Studies[*]

James Bisbee[†]and Dan Honig[‡]

Abstract

Audit studies are a powerful tool, one that has helped advance our understanding of the world across a range of contexts. However, the typical methodological design has limited the ability of these experiments to address questions of mechanisms, heterogeneous effects, or otherwise test theoretically or policy-relevant questions. In this paper, we describe and demonstrate a design innovation we call a "hierarchical" audit. This design allows researchers to document not just the incidence of theoretically interesting and substantively consequential social science phenomena, but also to explore where, when, how, and why these phenomena appear. Hierarchical audits, where feasible, can offer more nuanced exploration of theory and mechanisms, with benefits for both scholars and policy practitioners.

Audit Studies | Methodology | Discrimination | Theory Testing

10,077 words

[*] We are grateful to John Yinger, Richard Murnane, Neal Beck, Shelby Grossman, Molly Roberts, Charles Crabtree, Kevin Munger, Steve Rashin, Sean Kates, and Cyrus Samii for feedback on this project. All mistakes are our own.

[†] Corresponding author. PhD Candidate at NYU Wilf Family Department of Politics, 19 W. 4th St., New York, NY 10012. (802) 498-5215. james.bisbee@nyu.edu

[‡] Assistant Professor at Johns Hopkins SAIS. 1717 Massachusetts Avenue N.W. Room 735A, Washington, DC 20036. (202) 587-3253. dhonig@jhu.edu

1 Introduction

An “audit design” (also referred to as a “correspondence design”) randomly exposes subjects (or “auditees”) to identical auditors who differ only in some specific dimension, such as membership in a social group. Differences in auditee behavior can therefore be attributed to differences in the dimension on interest. Audit experiments have most famously been used to document systemic discrimination in social science fields ranging from economics to psychology to sociology to political science. Our knowledge of the extent and intensity of discriminatory behavior is indebted to the audit methodology.

However, audit experiments suffer from two shortcomings. First, while they are best suited to identify the existence of discrimination, they typically aren’t able to identify the mechanisms behind discrimination, making them ill-suited to policy evaluation and theory testing. Second, and more importantly, there remain challenging inference issues that are under-appreciated in the field. This paper describes these issues in detail and presents a methodological innovation we call a “hierarchical audit design”, designed to address these issues.

This design innovation overlays additional treatment arms on top of the discrimination elicitation and subjects each auditee to each possible combination of treatment arms – effectively applying a conjoint form to audit experiments. Doing so returns the full potential outcome schedule (assuming there are no spill-overs between treated and control conditions) which can be used to address both shortcomings described above. The benefits to evaluating policy or testing theory are straightforward – the second treatment arm can be chosen to speak directly to theory or policy. In addition, the full potential outcomes schedule allows for direct tests of SUTVA violations that would otherwise undermine the identification strategy.

We acknowledge that our proposed method is not always appropriate. There are certain experimental contexts in which recording multiple observations for each auditee is infeasible, either for ethical concerns or logistical limitations. In these contexts, we hope our paper's discussion of the inference challenges arising in between-subjects designs are illuminating. Furthermore, multiple interventions on the same auditee increases the likelihood of detection. Our proposed method is most appropriate for situations in which the auditee confronts many similar or identical phenomena as a matter of course, allowing for the research to use these phenomena as a treatment vehicle without fear of detection. But in other cases where such phenomena are unavailable, traditional audit designs are preferable.

However, we also believe that the research on discrimination has found itself in an empirical holding-pattern due largely to the field's reliance on traditional audit designs. Where empirical contexts allow, we believe a hierarchical design has the potential to galvanize the literature on discrimination by extending the empirical scope to include not just the incidence of discriminatory behavior, but also the theories build to deepen our understanding of the phenomenon and the policies developed to combat it.

2 Audit Research Designs

Audits are part of a longer body of literature on discrimination that combines experimental methods with other quantitative approaches to understanding the extend of discrimination. Non-experimental studies of discrimination rely on statistical models with a rich set of controls and an indicator variable for some social group identity such as race (Murnane, Willett and Levy, 1995; O'Neill, 1990; Neal and Johnson, 1996; Kahn, 1991; Knowles, Persico and Todd, 2001). In these studies, causal identification of discrimination relies on the assumption that the observed controls soak up non-discriminatory elements of outcome differences, leaving whatever remains explained by the group identity indicator as evidence of discrimination.

The original audit experiments attempted to relax this assumption by comparing outcomes for experimentally matched social group members (Yinger, 1995; Wienk et al., 1979; Hakken, 1979). Estimates of discrimination were calculated by subtracting one group's mean outcome from the other's, using a between-subjects design. Under the assumption that effects were constant across auditees, this design provided well-identified estimates of system-wide discrimination that were easily understood by readers of all backgrounds. In the United States, growing interest in audit experiments throughout the 1970s and 1980s lead to revelations about discrimination in the areas of housing (Hakken, 1979; Wienk et al., 1979; Turner, Struyk and Yinger, 1991), employment (Gill, 1989; Kirschenman and Neckerman, 1991; Neumark, Bank and Van Nort, 1996), and consumer markets (Ayres and Siegelman, 1995; Yinger, 1998; Feagin, 1991). These findings informed policies such as affirmative action (Bergmann, 1997) and equal opportunity employer requirements (Bertrand and Mullainathan, 2004).

Current audit studies typically use a within-subjects or "paired" design in which each auditee is subjected to both social groups (for ease of exposition, we shall refer to these social groups as the treatment and control). In this design, the counter-factual is drawn not from the mean outcome in the control population but rather from the control outcome observed for the same auditee. If a hiring manager offers an interview to a white candidate but not an equally qualified African-American candidate, this is interpreted as discriminatory behavior. The paired design allows researchers to control for non-constant effects across auditees by estimating the system-wide level of discrimination as the average of auditee-specific differences.

These audit experiments typically use human confederates who pose as a real-life interlocutor (such as a job applicant, potential home or car buyer, taxi fare, etc.). These confederates are matched on a variety of physical and personality characteristics and trained to behave as similarly as possible, differing only in some social identity dimension such as race or gender. Siegelman and Heckman (1993) and Heckman (1998) criticized the use of human confederates in these audit designs, arguing that unobserved differences could confound the results. Researchers responded by adopting "correspondence" designs in which human confederates were replaced by paper or electronic treatment conditions that simulated communication with auditees. In a particularly well-known example, Bertrand and Mullainathan (2004) send resumes that differ only in the racial association of the name to hiring departments to test for racial discrimination.[1]

The audit methods summarized here have made important contributions to our understanding of discrimination, both in terms of its prevalence as well as its spread. However, both types of audit designs suffer from two limitations. First, they are best equipped to document the *incidence* of discrimination but struggle to explain the *mechanisms* behind it. A hierarchical design specifically allows for exploring mechanisms by overlaying a second treatment arm that is motivated either by theory (for example, manipulating primes for taste-based versus statistical discrimination) or by policy (for example, piloting an anti-discrimination intervention on a random subset of auditees). The richness of a hierarchical design's mechanism testing generalizes beyond discrimination to any context in which an audit design can be applied.

Second, both audit and correspondence studies are vulnerable to non-constant effects across treatment arms which may undermine the accuracy of even system-wide estimates. This is an acknowledged yet under-appreciated threat to causal inference in the field, particularly in cases where researchers must rely on human confederates. Careful preparation is not always enough to ensure that treatment delivery is identical across confederates, potentially leading to spurious conclusions. In theory, auditor-fixed effects protect against heterogeneous effects concerns. In practice, traditional audit designs for the research to choose either auditor or auditee fixed effects, but not both – a limitation relaxed in a hierarchical design.

[1] While the literature accepted this solution, even correspondence studies can suffer from the critiques presented in Siegelman and Heckman (1993) and Heckman (1998) when using binary outcome variables, as detailed by Neumark (2012).

2.1 Audit Designs

We consider audit designs as existing on a spectrum from between-subject designs (representing one interaction per auditee) to hierarchical designs (representing multiple interactions per auditee), illustrated in Figure 1. We separate the goal of these experiments into two scope categories: documenting the incidence of system-wide discrimination on the one hand, and exploring the mechanisms that drive this discrimination on the other. The former scope condition comprises the vast majority of existing audit experiments and is best equipped to answer the twin questions of "is there discrimination?" and "how much is there?" The latter scope condition seeks to answer questions such as "why is there discrimination?", "among whom is it most prevalent?", and "what policy responses can minimize it?"

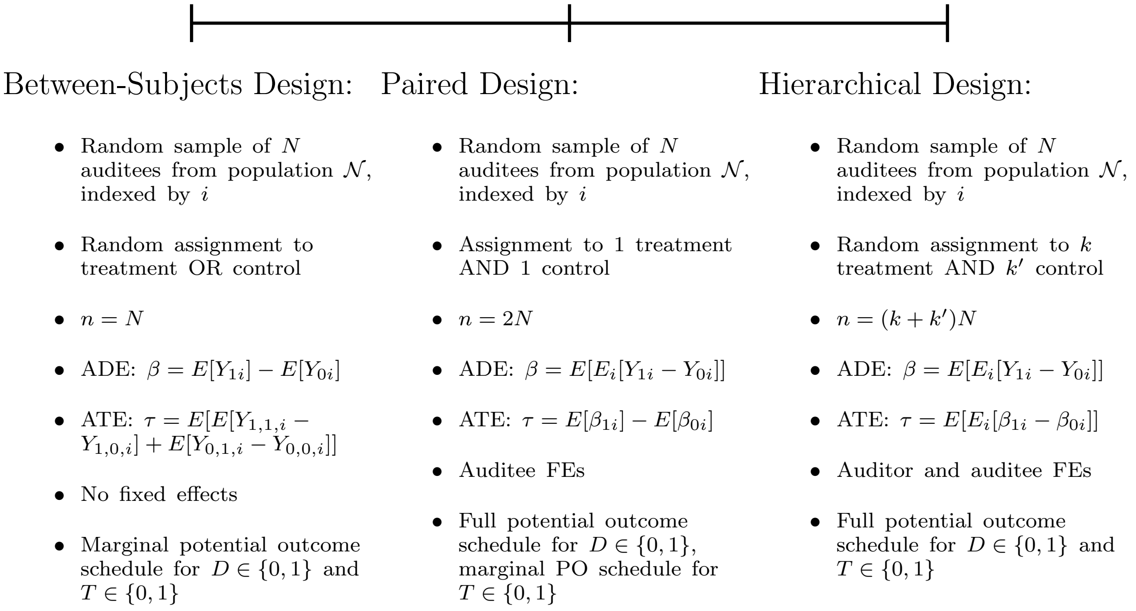
In the discussions that follow, we refer to estimates of the former scope category as the Average Discrimination Effect (ADE, denoted by β) and estimates of the latter scope category as the Average Treatment Effect (ATE, denoted by τ). We refer to treatment conditions designed to estimate the ADE as "elicitation treatments" denoted by $D \in 0, 1$. We refer to treatment conditions designed to estimate the ATE as "evaluation treatments" denoted by $T \in 0, 1$.[2]

The difference between the ADE and the ATE is a matter of degree. The more heterogeneity a researcher accounts for in describing the incidence of discrimination, the closer she gets to the underlying mechanisms. In this paper, we treat the ADE and ATE as extreme poles for two reasons. First, doing so helps clarify the benefits associated with our proposed hierarchical design. Second, muddying the delineation between effect heterogeneity and mechanisms is counter-productive to rigorous identification. Throughout, we emphasize the importance of pre-registration when conducting hierarchical.

On the left are between-subjects designs in which N auditees (indexed by i) are randomly

[2] Our discussion generalizes to continuous treatments.

Figure 1: Types of Audit Experimental Designs



Between-Subjects Design:

- Random sample of N auditees from population $\mathcal{N}$, indexed by i
- Random assignment to treatment OR control
- $n = N$
- ADE: $\beta = E[Y_{1i}] - E[Y_{0i}]$
- ATE: $\tau = E[E[Y_{1,1,i} - Y_{1,0,i}] + E[Y_{0,1,i} - Y_{0,0,i}]]$
- No fixed effects
- Marginal potential outcome schedule for $D \in \{0,1\}$ and $T \in \{0,1\}$

Paired Design:

- Random sample of N auditees from population $\mathcal{N}$, indexed by i
- Assignment to 1 treatment AND 1 control
- $n = 2N$
- ADE: $\beta = E[E_i[Y_{1i} - Y_{0i}]]$
- ATE: $\tau = E[\beta_{1i}] - E[\beta_{0i}]$
- Auditee FEs
- Full potential outcome schedule for $D \in \{0,1\}$, marginal PO schedule for $T \in \{0,1\}$

Hierarchical Design:

- Random sample of N auditees from population $\mathcal{N}$, indexed by i
- Random assignment to k treatment AND k' control
- $n = (k + k')N$
- ADE: $\beta = E[E_i[Y_{1i} - Y_{0i}]]$
- ATE: $\tau = E[E_i[\beta_{1i} - \beta_{0i}]]$
- Auditor and auditee FEs
- Full potential outcome schedule for $D \in \{0,1\}$ and $T \in \{0,1\}$

sampled from a population $\mathcal{N}$ and assigned to either elicitation treatment ($D = 1$) or control ($D = 0$) conditions, producing a total sample size $n = N$. Potential outcomes Y are recorded and the average discrimination effect (ADE) is calculated by subtracting the average outcomes of the control group from those of the treated ($\beta = E[Y_{1i}] - E[Y_{0i}]$). Random assignment to treatment condition allows us to assume that the treated and control units are valid counter-factuals for each other.

In the middle are paired audit designs which differ only in their treatment assignment. Here, N auditees are assigned to *both* elicitation treatment and control conditions, yielding recorded potential outcomes Y_{1i} and Y_{0i} for the sample individual i. Under this design with a binary treatment condition, the total sample size $n = 2N$ although with categorical or continuous treatments, n can be even larger. In a paired design, the crucial identifying assumption is that an individual's experience of the treatment condition does not influence their potential outcome in the control condition, and vice versa. This assumption is often referred to as the Stable Unit Treatment Value Assumption or SUTVA. Under this assumption, the researcher can estimate an auditee-specific ADE $\beta_i = E_i[Y_{1i} - Y_{0i}]$ and then average over this vector of estimates, effectively implementing auditee fixed effects. The final ADE can be expressed as $\beta = E[\beta_i] = E[E_i[Y_{1i} - Y_{0i}]]$.

Between-subjects and paired designs constitute the vast majority of existing audit studies in the literature on discrimination and have been used to persuasively document the existence and extent of discrimination across a variety of contexts. Of course, it is only natural that sociologists of all disciplines would want to understand the mechanisms by which discriminatory behavior is realized, either to speak to theory or maximize policy interventions. Identifying discriminatory behavior's existence and extent is a crucial first step. But richer questions concerning why people discriminate and how discrimination can be influenced require an additional estimator that we refer to as the ATE.

It is helpful to consider estimation of the ADE and ATE sequentially although in application, a hierarchical experiment identifies both simultaneously. First, note that estimation of the ADE is fundamentally a latent analysis in which a proxy behavioral measure Y is theorized to capture the underlying discrimination. In practice, estimation can be obtained via linear regression of the form:[3]

$$Y = \alpha + \beta D + \epsilon \tag{1}$$

With a paired design, one obtains a vector of β_i coefficients for $i \in N$ that represent the auditee-specific discrimination measures. In a hierarchical design, we propose adding an additional evaluation treatment arm T and treating the vector of β's as the outcome of interest. Thus the ATE is equal to $E[\beta_{1i}] - E[\beta_{0i}]$.

[3] Depending on the distribution of the outcome variable, other estimators may be more appropriate.

To make concepts concrete, consider a variation of the famous audit experiment described in Bertrand and Mullainathan (2004). In this experiment, a sample of N job postings (corresponding to N hiring departments who are the auditees of interest) are sent resumes containing different names. The names have strong racial associations with the elicitation treatment condition $D = 1$ corresponding to a white-sounding name and the control condition $D = 0$ corresponding to a black-sounding name.

In addition, the researcher is interested in adjudicating between two popular theories of discrimination: taste-based and statistical. Taste-based discrimination captures the intuition that people simply have an innate preference for one race over another. Conversely, statistical discrimination refers to a skill-based stereotype that members of a certain race are better at certain tasks. Under this theoretical framework, discrimination is an informational shortcut that is activated when other sources of information are not available. If we assume that socioeconomic status (SES) fills in any uncertainty about skill that would otherwise be soaked up by statistical discrimination, our second evaluation treatment arm will allow us to test which mechanism of discrimination is at play. For the sake of simplicity, we refer to $T \in \{0, 1\}$ as capturing a randomized socioeconomic status indicator on the resume.

In a between-subjects design, N auditees are randomly assigned to receive both elicitation and evaluation treatments. One quarter receives white names with high SES, one quarter receives white names with low SES, one quarter receives black names with high SES, and one quarter receives black names with low SES. Average discrimination effects are calculated as the average of the difference between white and black names with high SES and the difference of white and black names with low SES. The random assignment to each treatment arm allows each discrete group of auditees to serve as valid counter-factuals for the others. However, estimation relies on the marginal potential outcome distribution for each group. If there are non-constant treatment effects, a between-subjects design risks missing potentially interesting variation. Furthermore, a between subjects design is unable to account for either auditee or auditor fixed effects.

In a paired audit design, all auditees receive both elicitation discrimination treatment conditions $D \in \{0,1\}$, seeing resumes with both white and black names. In addition, the N hiring departments are randomly divided into evaluation treatment and control groups $T \in \{0,1\}$ for the socioeconomic status arm. The researcher can thus calculate β_i for each hiring department and then run a second regression of this vector of discrimination estimates on the SES treatment to capture the ATE although in practice the preferred method is via an interacted specification.

The paired audit design can, in fact, be understood as a paired-subjects design for the ADE and a between-subjects design for the ATE. Random assignment ensures that those auditees in $T = 1$ serve as a valid counter-factual group for those in $T = 0$. Even so, the ATE is still obtainable only via the marginal potential outcome distribution of β. Furthermore, the researcher can only implement auditee *or* auditor fixed effects in the first stage, making estimates of the ATE vulnerable to non-constant effects.

By contrast, our hierarchical design exposes auditees to both elicitation and evaluation treatment arms as illustrated in the 2-by-2 design in Table 1. Specifically, each auditee sees a white resume with high SES ($Y_{1,1}$), a white resume with low SES ($Y_{1,0}$), a black resume with high SES ($Y_{\{}0,1\}$), and a black resume with low SES ($Y_{0,0}$). This rich design returns the full potential outcome schedule for each auditee, relaxing the reliance on marginal distributions and removing the need to assume constant effects.[4]

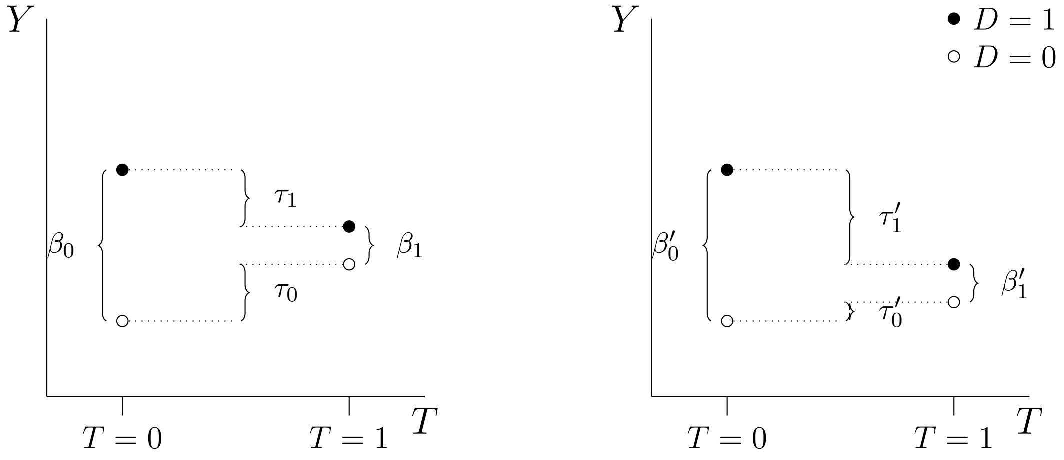
To illustrate these benefits, consider a hypothetical hierarchical experiment visualized in Figure 2 where D represents the elicitation treatment (i.e., randomly assigning an auditor to act out one ethnicity or another), T represents the evaluation treatment (i.e., exposing the auditee to a policy intervention designed to increase empathy for an out-group member),

[4] However, the researcher must therefore be confident that SUTVA is not violated in order for the observed potential outcomes to be interpreted as the full schedule.

Table 1: 2-by-2 Example of ADE and ATE in a Hierarchical Design

	$D = 0$	$D = 1$	β
$T = 0$	$Y_{0,0}$	$Y_{1,0}$	$\beta_0 = E[Y_{1,0} - Y_{0,0}]$
$T = 1$	$Y_{\{}0, 1\}$	$Y_{1,1}$	$\beta_1 = E[Y_{1,1} - Y_{\{}0, 1\}]$
τ	$\tau_0 = E[Y_{\{}0, 1\} - Y_{0,0}]$	$\tau_1 = E[Y_{1,1} - Y_{1,0}]$	$\tau = \beta_1 - \beta_0$ $\beta = \tau_1 - \tau_0$

and potential outcomes Y are subscripted $Y_{d,t}$.



(a): $\tau = \beta_1 - \beta_0$

(b): $\tau' = \beta'_1 - \beta'_0$

Figure 2: Elicitation treatment ($D = 1$) and control ($D = 0$) conditions illustrated by black and white circles, respectively. Evaluation treatment ($T \in \{0, 1\}$) indicated along x-axis. Extent of discrimination (β) and effect of evaluation treatment (τ) are identical in both panels. However, the process by which identical ATEs manifest differ ($\tau'_1 \neq \tau_1$ and $\tau'_0 \neq \tau_0$).

In the pretreatment period, the comparison between Y_{10} and Y_{00} yields an estimate of the auditee's baseline level of discrimination, represented by β_0 while the analogous measure in the post-treatment period is β_1. The estimate of the ATE, τ, is simply the difference between β_1 and β_0, analogous to the case scenario presented in Table 1.

However, as illustrated in panel (b) of Figure 2, identical ATEs may mask different outcome behaviors. If Y is the call-back rate for resumes, panel (a) illustrates a post-treatment outcome in which the average call-back rate has not changed due to the decline in discrimination against $D = 1$ being offset by the increased call-back rate enjoyed by $D = 0$. Conversely, in panel (b) the brunt of the ATE manifests as a reduction in discrimination against $D = 1$, resulting in a decline in the overall call back rate.

By observing all counter-factual outcomes for each auditee, the researcher can assess not just the average discrimination and treatment effects (ADE and ATE) but also the equilibrium behaviors that produce them. Furthermore, the full potential outcomes schedule can also be leveraged to characterize non-constant effects across pre-treatment characteristics which can further inform both theory and policy interventions. These types of analyses would be impossible in conventional between-subjects or paired audit designs without heroic assumptions.

In the application that follows, we describe three methods that leverage the full potential outcomes schedule produced by a hierarchical audit experiment. The first uses the treated and control potential outcomes to yield rich understandings of effect heterogeneity. The second method decomposes this effect heterogeneity into the components that differ due to differences in groups or differences in group-specific covariates. And the third method describes how to design optimal treatment regimes based on effect heterogeneity.

3 Empirical Context

To demonstrate the opportunities afforded by increasing the number of interactions per auditee, we use pilot data from an audit experiment fielded in Nigeria in 2012 (Grossman and Honig, 2017). This experiment used a hierarchical design to adjudicate between two competing theories of why people discriminate: taste-based and statistical discrimination.

Taste-based discrimination predicts that an individual will reward in-group members and will discriminate against out-group members purely for personal utility reasons (as modeled by Becker (1957) and documented by Adida et al. (2013), Michelitch (2015), Neumark (1996) and others). Conversely, statistical (or instrumental) discrimination assumes that individuals discriminate to maximize some objective component of welfare, such as income (as modeled by Phelps (1972) and documented by Barr and Oduro (2002) and others).

To speak to these theories, the authors chose an experimental context where neither buyer nor seller would have expectations of a repeated interaction, commensurate to a no repeated game. Grossman and Honig drew auditees from housing agents[5] and rice sellers in Lagos and applied three different treatments to test for discrimination: one class-based, one ethnicity-based, and one religion-based. In the context of no repeated interaction, the authors argue that ethnic and religious discrimination is taste-based whereas discrimination along class lines is statistical. Put simply, should sellers offer cheaper prices to co-ethnics/co-religious confederates under the assumption that they would never interact with these auditors again, they must be maximizing their personal utility function. Conversely, should sellers charge higher prices to auditors who appear to be high class, they are likely doing so under the assumption that high class individuals can afford to pay more, thereby discriminating in a statistical manner. The experiment is described in greater detail below.

[5] In Lagos State, Nigerians who wish to rent housing approach agents in the neighborhoods where they wish to reside. Landlords who have available housing will give the key to the housing to a single agent, who then searches for renters. Once an agent finds a renter, the agent manages the entire transaction, negotiating the rent and all fees, and pocketing whatever is earned above what the landlord had asked for in rent. Across Nigeria, the norm is for agents to demand two year's rent paid up front, and one year's rent in advance for each subsequent year. The equilibrium is essentially landlords never meet tenants, keep properties in poor conditions, and expect tenants to move out after the initial two years. Thus the housing market involves no expectation of a repeated game by tenants, landlords, or housing agents.

3.1 Experimental Context

Grossman and Honig recruited 11 male University of Lagos students between the ages of 20 and 27 to act as confederates and engage in housing and rice transactions, one of whom doubled as a research assistant. Among the confederates, 6 were Yoruba (5 Christian, 1 Muslim), 2 Hausa (1 Christian, 1 Muslim), and 3 Igbo (all Christian, as Igbo Muslims are very uncommon, particularly in Lagos). Critical for the purposes of this experiment, while there are stereotypes about how individuals of certain ethnic groups look, many (perhaps an overwhelming majority of) Nigerians do not fit neatly into these stereotypes. A Yoruba Nigerian can pretend to be Igbo. A Hausa Nigerian can pretend to be Yoruba. Additionally, there are of course no physical characteristics associated with a particular religion.

Before the start of the experiment, the research assistant and authors identified 10 housing agent offices and 11 rice sellers in an ethnically heterogeneous neighborhood called Bariga, about 7 miles north of Lagos Island. The authors selected the Bariga neighborhood because of its proximity to University of Lagos, which made it ordinary that students were looking for housing in the area. Further alleviating suspicion was the fact that a new semester was starting just two weeks after the experiment, making it even less likely to raise eyebrows that several students over the course of a week were looking for housing.

Treatment

The authors experimentally manipulated confederate signifiers of socioeconomic status and ethnicity and randomly assigned auditees to hierarchical treatment conditions. The class treatment was achieved by alternatively assigning confederates to dress in a higher class way (what their confederates called “radiant”) or a lower class way (“unkempt”). On days when confederates were radiant, they would wear a button-down shirt, pants, dress shoes, and often a watch. On unkempt days they would wear a t-shirt, shorts, and sandals.

To keep the identity treatment as natural as possible, assigned ethnic and religious identity were conveyed slightly differently in the agent and rice stages of the experiment as well. When visiting agents, confederates were assigned names that are clearly associated with an ethnicity and religion. For example, one assigned name was Emmanuel Abubakar, a name most Nigerians would know conveyed Christianity (Emmanuel) and Hausa identity (Abubakar).[6] When visiting rice sellers, confederates would wear a flashy cross necklace if they were going as a Christian and would use the traditional Islamic greeting "As-salamu alaykum" (peace be upon you) if going as a Muslim. Confederates conveyed their ethnicity by greeting the rice seller (after the Islamic greeting, if going as a Muslim) in Hausa, Igbo, or Yoruba. If the seller did not respond in the same language, confederates reverted immediately to Pidgin English.

In this hierarchical audit design, each seller was exposed to every possible combination of treatments, yielding $2^3 = 8$ observations in total (Yoruba vs. non-Yoruba ethnicities, Muslim vs. Christian religions, high vs. low class). Some sellers have fewer recorded observations due to confederates being unable to find their stand the following day or the office being manned by a different agent. Armed with this rich treatment design, alongside measures of pre-treatment covariates for the sellers, we turn to demonstrating the benefits of a hierarchical audit.

4 Results

To demonstrate the benefits of the hierarchical design, we move across the spectrum of audit designs summarized in Figure 1. With multiple observations per auditee, we can randomly

[6] The authors conducted a small post-audit survey where they asked Nigerians to tell them which ethnicity and religion were most commonly associated with the 49 names used in the housing audit. There was an 85% accuracy rate, commensurate to that recorded in a similar correspondence apartment rental audit conducted by Bartos et al. (2013).

sample from the data to simulate the single and paired designs. We first demonstrate the threats to system-wide inference from the assumption of homogeneous effects. We then use the data to illustrate how a hierarchical design can speak to theory or evaluate policy.

4.1 Incidence of Discrimination Estimates (ADE)

We begin by demonstrating the risks associated with the homogeneous effects assumptions necessary to estimate the prevalence of discrimination (ADE) using between-subjects or paired audit designs. To imitate a between-subjects design, we randomly sample one observation from each seller in the data. We then randomly sample two observations from each seller, one under treatment and the other under control, to imitate a paired design. We regress Grossman and Honig's discrimination index[7] on an indicator for whether the auditor shares an ethnic identity with the auditee ("ethnicity"), whether they share a religious identity with the auditee ("religion"), and whether the auditor was dressed radiantly ("class"). We save this vector of coefficients and repeat the random sample 500 times. We plot the

[7] The discrimination index combines the data on rice and real estate as follows: $\frac{P_{ij} - minP_i}{minP_i}$ where P_{ij} is the price for rice paid by confederate j to seller i or the housing quotation offered from seller i to confederate j, and $minP_i$ is the minimum price seller i quoted any confederate. This standardizes discrimination as the percentage above the minimum price offered by the seller; so were confederate 1 to receive a price 20% higher than confederate 2 from seller A, that observation would have a discrimination index of 0.2. The same is done with the weight of the rice, though in reverse, $(\frac{W_{ij} - maxW_i}{maxW_i})$ to produce an observation-specific weight discrimination index. These metrics are combined to generate two measures – one for price discrimination (pooling housing and rice transactions) and one for combined discrimination, which treats rice discrimination as the sum of weight and price discrimination and housing discrimination as only price discrimination (as the result of the housing transaction – the "weight" equivalent, actual housing quality – is unobserved in this study).

bootstrapped distribution of estimates for each explanatory variable of interest in Figure 3.

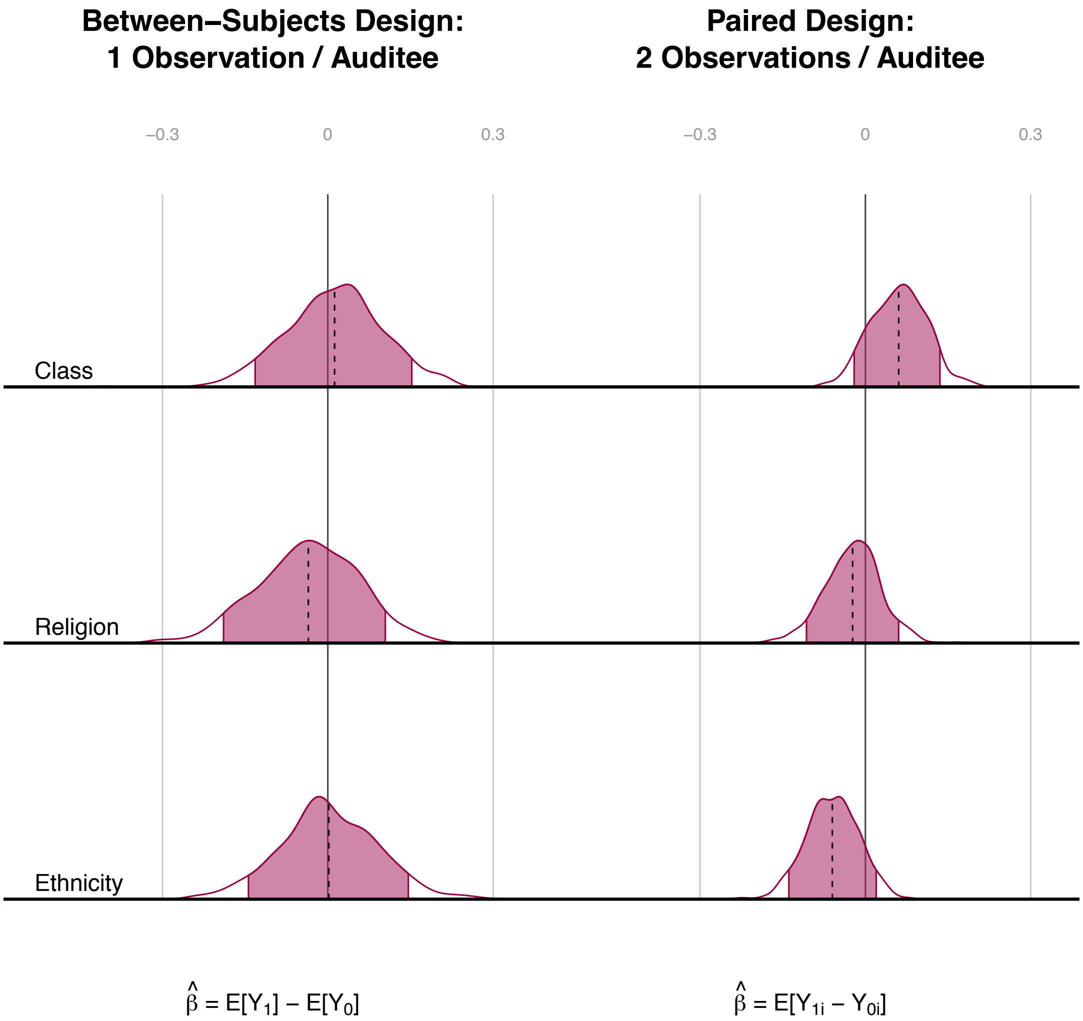


Figure 3: Between-subjects and paired design bootstrapped estimates (500 estimates presented as densities) for each explanatory variable (y-axis). 90% confidence intervals indicated by shaded regions.

Our findings start to express themselves in a paired design when we estimate the effect of treatment within each seller as $E[Y_{1i} - Y_{0i}]$ (right panel of Figure 3). There is suggestive evidence of both taste-based discrimination (charging co-ethnics cheaper prices) and statistical discrimination (charging high-class auditors more). However, the small samples for each auditee in the paired design still produce noisy estimates that fail to achieve statistical significance at conventional levels. Furthermore, we are unable to also control for unobserved auditor confounds theorized by Siegelman and Heckman (1993) and Heckman (1998) to undermine identification when using human confederates.

Turning to the full data with multiple observations per auditee allows us to compare four different specifications to illustrate how fixed effects are vital to controlling for unobserved factors in auditors and auditees. Table 2 presents the results of our full experiment with no fixed effects (column 1), auditor fixed effects (column 2), auditee fixed effects (column 3), and both auditor and auditee fixed effects (column 4).

As illustrated, implementing both auditee and auditor fixed effects yield the most substantively and statistically significant results. Both class and shared ethnicity are shown to have independent and meaningful effects on the prices sellers charge. Higher class confederates are charged more while auditors who share an ethnicity with the auditee are charged less. Meanwhile, the coefficient on religion is a precisely estimated zero. A cursory analysis of the results suggests that auditor-training was effective in standardizing treatments across confederates, with the inclusion of auditor (buyer) fixed effects producing only minimal changes in the estimates, significance, and R^2. Conversely, the inclusion of auditee (seller) fixed effects is demonstrated to be vital in identifying the prevalence of discrimination in this context, highlighting the degree of effect heterogeneity across sellers that would otherwise obfuscate the differential treatment by class and shared ethnicity.

In sum, effect heterogeneity obfuscates important results when using a between-subjects or paired audit design. A hierarchical design allows researchers to employ both auditor and auditee fixed effects, soaking up confounding heterogeneous intercepts. Doing so highlights the prevalence of discrimination favoring co-ethnics and penalizing high-class buyers: a system-wide result that would not be apparent using traditional audit designs. Even in the case where auditor fixed effects are unnecessary (as in the example above, due to robust confederate training, or in certain types of correspondence designs), hierarchical experiments

Table 2: Type II errors and the importance of fixed effects

	No FEs	Auditor FEs	Auditee FEs	Both FEs
	(1)	(2)	(3)	(4)
Same Ethnicity	−0.041	−0.045	−0.061*	−0.075*
	(0.034)	(0.039)	(0.030)	(0.035)
Same Religion	−0.038	−0.031	−0.017	−0.007
	(0.033)	(0.035)	(0.030)	(0.032)
High Class	0.046	0.049	0.066*	0.069*
	(0.033)	(0.035)	(0.030)	(0.032)
Auditor FE	No	Yes	No	Yes
Auditee FE	No	No	Yes	Yes
Observations	173	173	173	173
R^2	0.030	0.054	0.373	0.402

Notes: Linear regressions of price index regressed on indicators for whether the buyer and the seller share an ethnicity (top row), share a religion (middle row), or if the buyer is dressed as a high class individual. Different levels of fixed effects are indicated by columns. Standard errors presented in parentheses. * $p < 0.05$; ** $p < 0.01$.

provide better power for estimating auditee fixed effects by recording more observations to allow for more precise measurement of the intercepts.

4.2 Theory Testing and Policy Evaluation (ATE)

The results summarized above find that discrimination is most prevalent along two dimensions: class and ethnicity. These are fundamentally ADE estimands that capture the extent to which sellers differentially charge different types of buyers for the same goods. However, these results carry implications for two competing theories of why individuals discriminate. The higher prices charged to high class confederates may reflect the expectation that these buyers can afford more and the seller can maximize profits (a theory of statistical discrimination). Conversely, discounts given to coethnics in an environment where repeated interactions are unlikely is easiest to explain with a theory of taste-based discrimination. How can we adjudicate between the two?

It is important to acknowledge that the original experimental design was not built for recovering causally identified estimands for the analysis we conduct here. We do not use these data to make causal claims but rather to demonstrate the increased flexibility of recording multiple observations per seller when it comes to assessing policy or speaking to theory, and the ability of hierarchical analysis to identify unanticipated effects that can then be subjected to rigorous out-of-sample testing. In practice, hierarchical experiments (like all experiments) should be pre-registered.

That being said, the simplest approach to adjudicating between taste-based and statistical discrimination is to include an interaction term between the two treatment manipulations. As illustrated in Figure 4, the story gets more nuanced when comparing how ethnicity and class interact in cases of discrimination. Specifically, we find that the effect of class on discrimination obtains only for non-coethnic buyers.[8] The penalty to high class buyers disappears when the buyer and the seller share an ethnic identity, suggesting that co-ethnic status acts as a "get out of jail free card" for high class buyers.

However, with multiple observations per seller we can dig deeper into the relationship. The data we use was collected over the course of a week. We leverage this temporal variation to explore whether day of the week has a differential effect on Muslims compared to Christians, suggesting a mechanism involving identity activation. Fridays are a holy day in Islam when many Muslims attend sermons. We test whether Muslim buyers exhibit greater evidence of taste-based discrimination on Fridays when their religious identity has been activated.

To do so, we estimate the interaction effect between Muslim sellers and Muslim buyers

[8] This stylized finding echoes working paper versions of what became Grossman and Honig (2017) but does not appear in the published version.

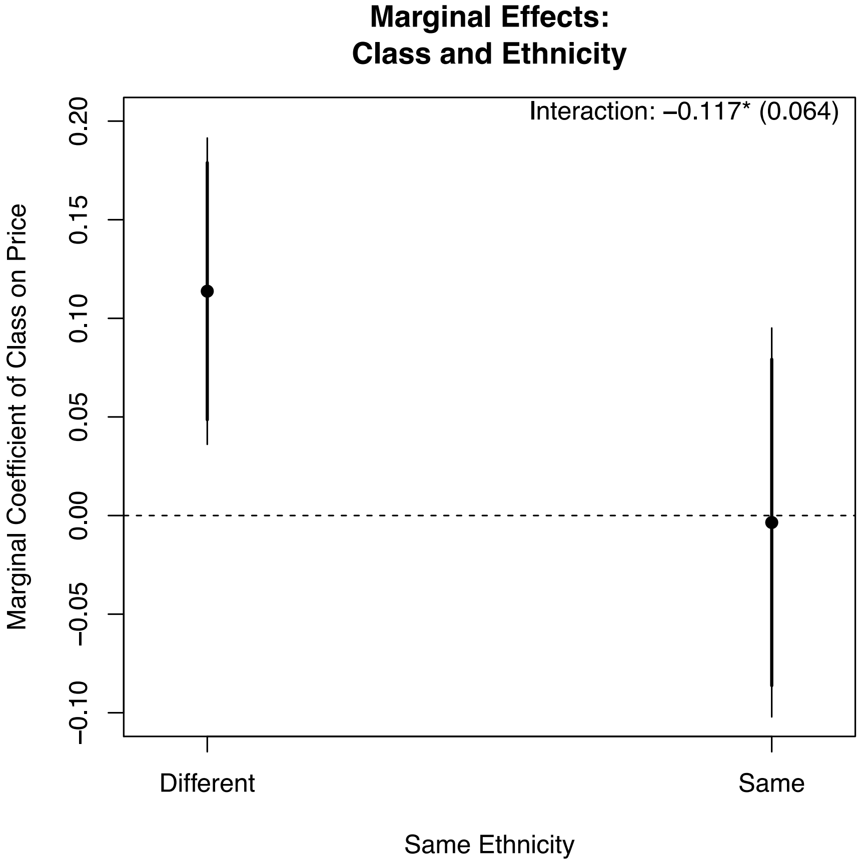


Figure 4: Marginal effects of high class treatment on discrimination (y-axis) for co-ethnics and non-co-ethnic interactions (x-axis). 90% and 95% confidence intervals indicated by thick and thin bars respectively.

on each day of the week separately, including auditor fixed effects. We plot the coefficients for each day in Figure 5 where each coefficient represents the additional change in price associated with a Muslim seller interacting with a Muslim versus non-Muslim buyer. As illustrated, Muslim sellers give significantly more preferential treatment to Muslim buyers on Fridays but on no other days of the week, consistent with an identity activation mechanism. Crucially, our hierarchical design enables us to compare the same seller's behavior on different days of the week.

Our analysis is able to not only document the existence of inflated prices for high-class

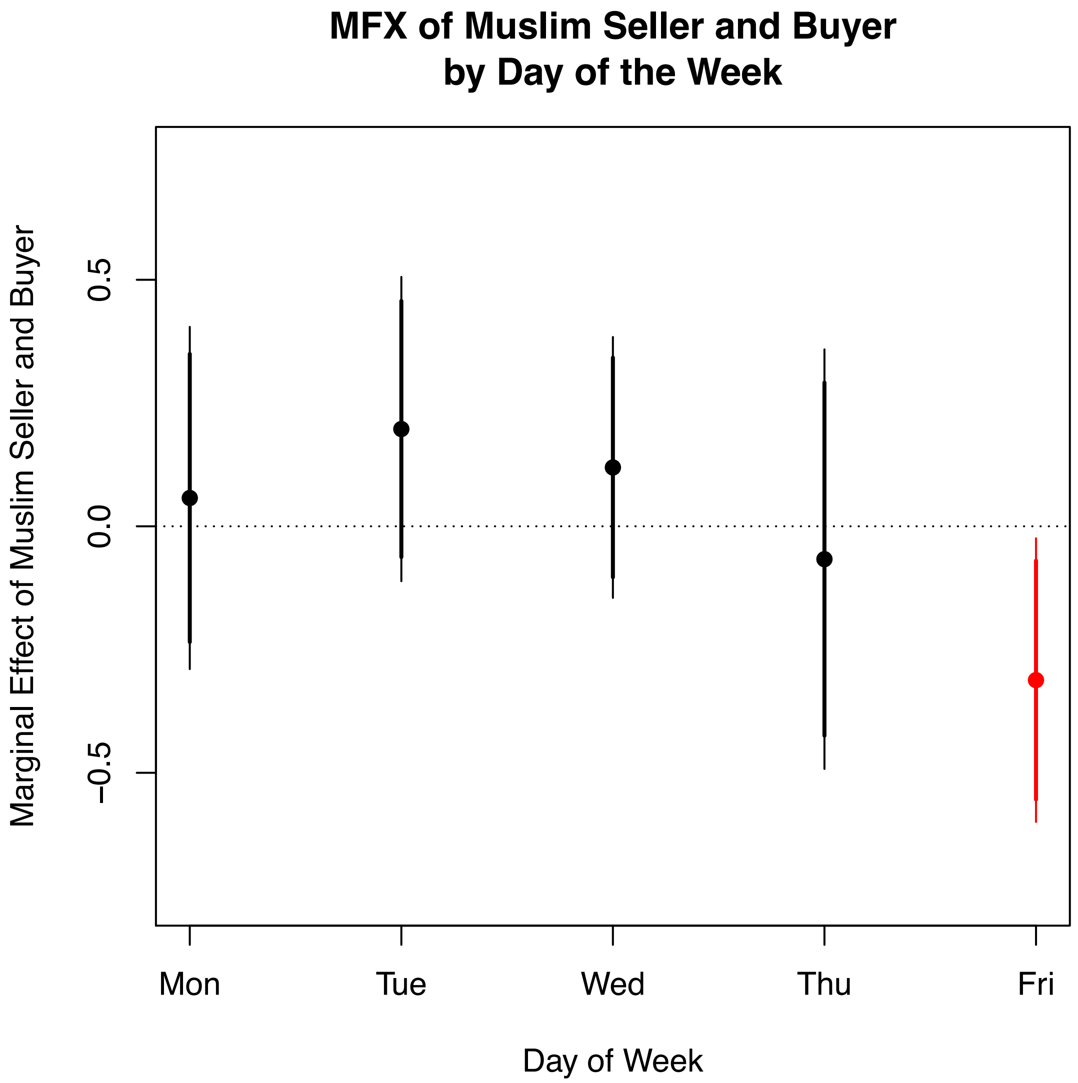


Figure 5: The marginal effect of a Muslim seller interacted with a Muslim buyer (y-axis) by day of the week (x-axis). The significant negative estimate on Fridays is consistent with taste-based discrimination increasing as religious identity is activated.

buyers, but also can test whether co-ethnicity reduces this effect. This analysis speaks to the relative strength of taste-based versus instrumental discrimination if we assume that class-based discrimination in a non-repeat interaction environment is statistical, while ethnic discrimination is taste-based. In addition, we find evidence consistent with a role played by identity activation wherein Muslim sellers quote more favorable prices to Muslim buyers when their Muslim identity has been activated. These findings from a pilot study using a hierarchical audit design demonstrate the opportunities available to researchers and highlight a potential policy intervention in the form of activating different types of identity.

4.3 Leveraging the Full Potential Outcomes Schedule

The benefits of a hierarchical design in the context of system-wide results are straightforward: more observations per seller allows for greater precision in estimating the fixed effects, thereby accounting for variation across sellers that would otherwise obfuscate the systemic discrimination we observe.

However, for researchers interested in documenting not just system-wide results but exploring the mechanisms underlying discrimination, this variation is of substantive interest. With access to the full potential outcomes schedule, we can characterize effect heterogeneity to a degree that would be impossible in conventional between-subjects or paired designs.

In the sections that follow, we demonstrate the benefits of a hierarchical design. We start with a simple analysis of heterogeneous effects across pre-treatment characteristics (focusing specifically on the religious and ethnic identity of our sellers). We then expand this analysis to examine differences in potential outcome distributions across groups and parameterize this heterogeneity via either effect decomposition or meta regression, as in Angrist, Pathak and Walters (2013).

As acknowledged earlier, the data are from a pilot study in which the researchers did not explicitly overlay a second randomized treatment arm. In the analyses that follow, we use pre-treatment covariates as though they were a randomly assigned second treatment arm. However, our results are purely descriptive characterizations of effect heterogeneity. Researchers who use a hierarchical design that randomizes the second treatment arm can use similar approaches to address causal questions relating to the ATE.

4.3.1 Interacted Regressions

The simplest way to explore heterogeneous effects across pre-treatment covariates is via interacted regressions. In Figure 6, we present our results from the following specification:

$$Y_{ij} = \alpha + \lambda + \beta_1 D_{ij} + \beta_2 ID_i + \beta_3 D_{ij} \times ID_i + \epsilon \tag{2}$$

where Y_{ij} is the indexed price for seller i to confederate j, α and λ are seller and buyer fixed effects, D_{ij} is the treatment indicator for whether the seller and buyer share the same ethnic identity, and ID_i is an indicator for the seller's (religious or ethnic) identity. Heterogeneous effects are captured by β_3.

As illustrated, there is evidence of effect-heterogeneity by pre-treatment group identities. Christians are slightly more preferential to co-ethnics than Muslims (top-right panel of Figure 6) while Muslims charge slightly more to high-class customers (bottom-right panel). However, there is more striking evidence of effect heterogeneity by seller ethnicity. For both the co-ethnic and high-class types of discrimination, there is evidence that non-Yoruba sellers discriminate more. This group-based difference is particularly striking for the prices charged to high-class confederates (coefficient -0.202; S.E. 0.086).

4.3.2 Potential Outcome Distributions

However, these marginal effects obscure interesting variation in potential outcomes that could be of interest when testing theory or evaluating policy. Following Angrist, Pathak and Walters (2013), we might suspect that differences in treatment effects may be due to differences in the potential outcome distributions of groups of sellers. Using hypothetical groups $G \in \{a, b\}$:

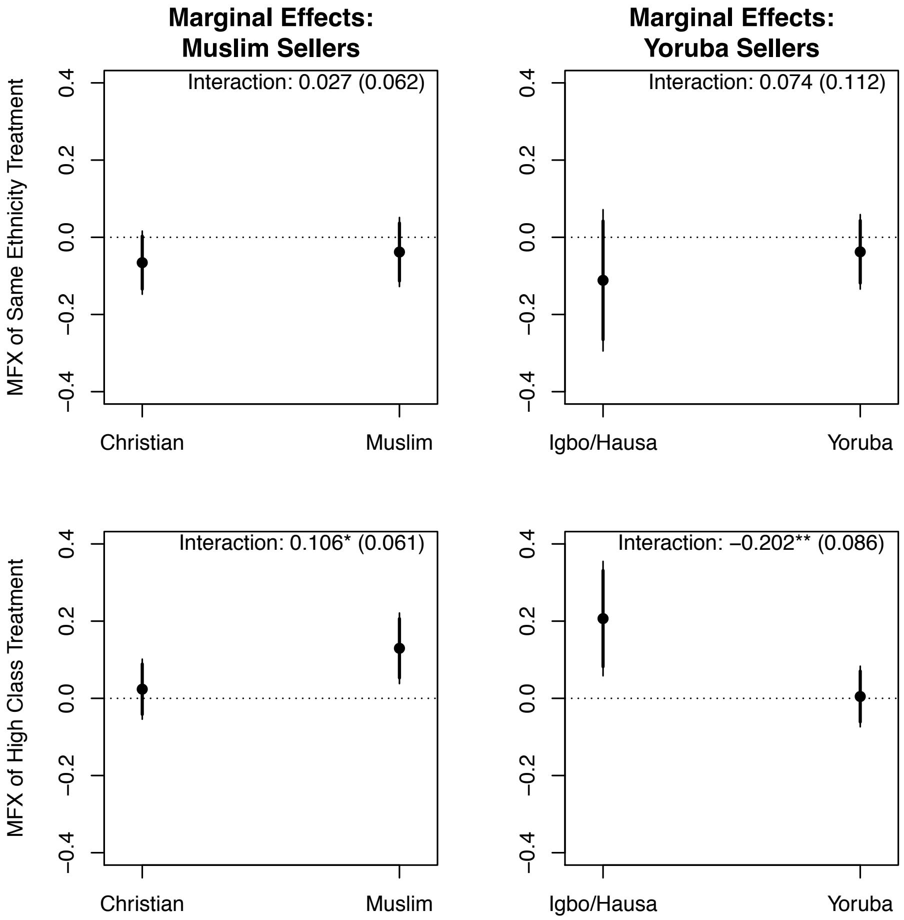


Figure 6: Marginal effects of treatment effects (same ethnicity treatment in top row, high class treatment in bottom row) against seller identity (religious identity in left column, ethnic identity in right column). Interaction term β_3 indicated in text in upper-right of each plot.

$$\beta_a - \beta_b = \underbrace{E_a[Y_{1i}] - E_b[Y_{1i}]}_{\delta_1} - \underbrace{(E_a[Y_{0i}] - E_b[Y_{0i}])}_{\delta_0}$$

where Y_{1i} is the outcome under treatment D_{1i} for seller i. Differences in treated (δ_1) and control (δ_0) potential outcomes between groups can shed light on the sources of effect heterogeneity. For example, if one group charges consistently higher prices in the control condition, treatment may have little effect due to market-determined price ceilings (this hypothetical scenario is illustrated in Figure 7).

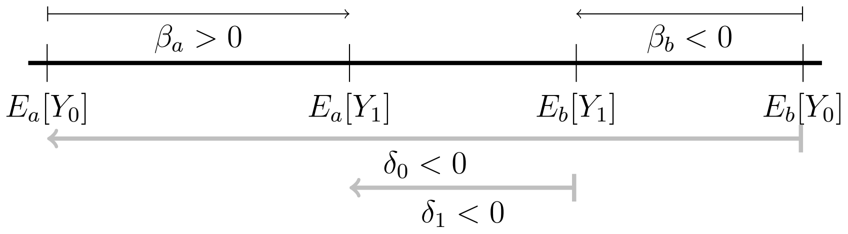


Figure 7: Differences in treatment (1) and non-treatment (0) counterfactuals between two groups a and b. The potential outcome distributions for group b are uniformly higher than for group a.

Our hierarchical design allows us to estimate δ_0 and δ_1 by using control and treated outcomes *within auditees*. Unlike in Angrist, Pathak and Walters (2013) who must rely on marginal means, our estimates of potential outcome distributions are generated from the full joint distribution. In Figure 8, we plot the potential outcome distributions for the high class and same ethnicity treatments by religious ($G \in \{M, C\}$) and ethnic identity ($G \in \{Y, N\}$) of the seller.

We now can make assessments not just of the direction and magnitude of change across sub-groups (represented by β_G) but also of the outcomes themselves. In the case of the class treatment effect (bottom two plots), it is apparent that substantial effect heterogeneity derives from the treatment itself, as illustrated by the almost identical control potential outcomes by both religious and ethnic identity. Such information might be useful to a policy-maker trying to maximize the effect of an intervention.

Conversely, the lack of effect heterogeneity in the ethnicity treatment (top two plots) masks substantial potential outcome differences in both the control and treatment conditions. In particular, while Yoruba and non-Yoruba both give similar discounts to co-ethnics (as illustrated by comparing β_N and β_Y in the top-right plot), analysis of the potential outcome

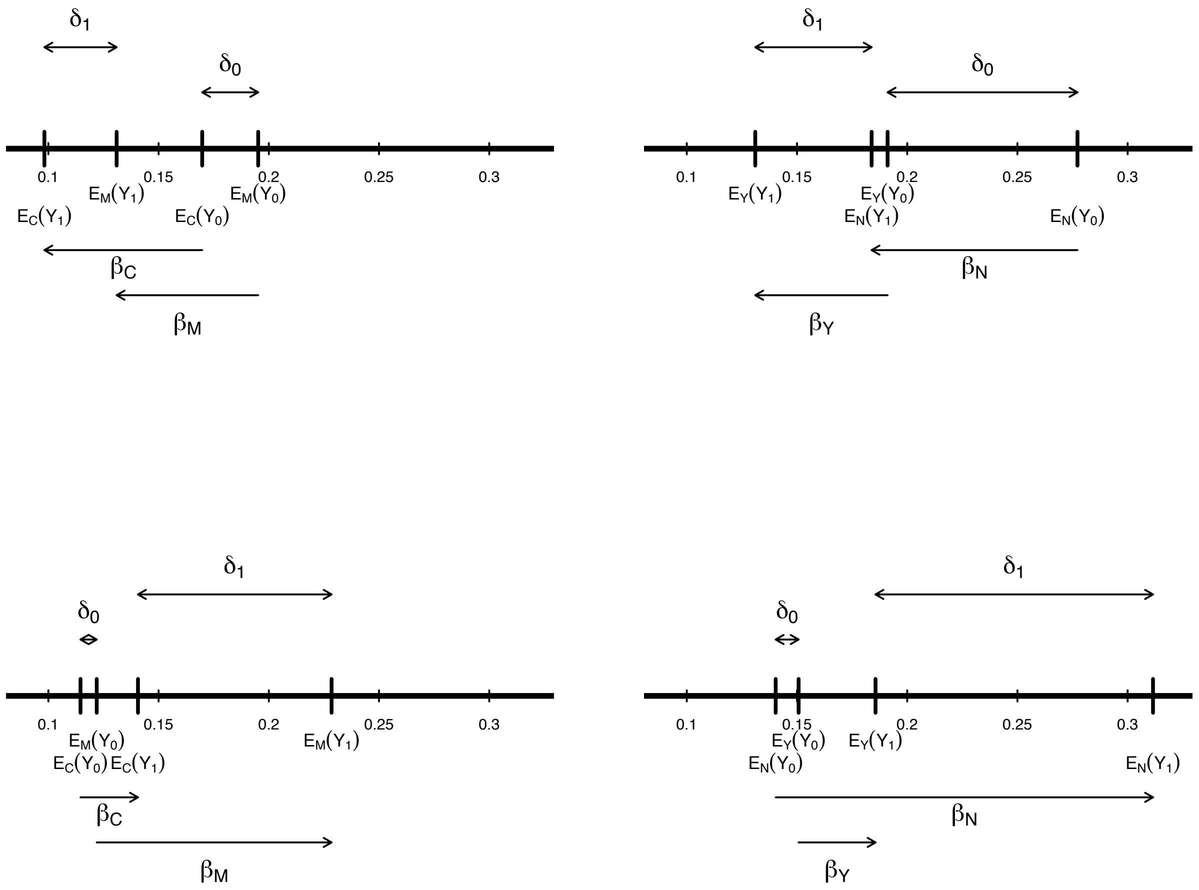


Figure 8: Differences in potential outcome distributions by religious groups (left column) and ethnic groups (right column) for the ethnicity treatment (top row) and class treatment (bottom row). The horizontal axes are the standardized outcome support for prices. Vertical ticks are labeled according to the expected control (Y_0) and treatment (Y_1) potential outcomes for each group (E_{group}). δ indicates the difference in potential outcomes by group and β is the effect estimate of treatment.

distributions highlights that Yoruba coethnic discounts bring prices down to the same level that non-Yorubas charge sans co-ethnic discount. Again, this potentially useful information would be lost in interaction regressions using conventional between-subjects or paired designs.

4.3.3 Effect Decomposition

Confronted with such rich characterizations of effect heterogeneity, the intuitive follow-up question is *why* we see the different effects that we do. Heterogeneous effects that vary by groups of auditees (as is the case with varying effects by ethnicity or religion) can be further explored via effect decomposition which asks how much of the group-level differences in effect sizes can be attributed to group-level differences in auditee characteristics. Following Angrist, Pathak and Walters (2013), we start with a model that interacts pre-treatment auditee characteristics with treatment in each group:

$$
\begin{aligned}
E_G[Y_i|D_i, X_i] &= X_i'\theta_G + \omega_G D_i + D_i X_i'\rho_G \\
\Rightarrow \beta_G &= \omega_G + E[X_G]'\rho_G \\
\beta_Y - \beta_N &= (\omega_Y - \omega_N) + (\bar{X}_Y'\rho_N - \bar{X}_N'\rho_N)
\end{aligned}
$$

where $\bar{X}_G = E_G[X_i|D_i]$ and, as above, subscripts $G \in \{Y, N\}$ reflect Yoruba and non-Yoruba groups. We can use an Oaxaca-Blinder decomposition (see Oaxaca (1973) and Blinder (1973)) to re-write the above as:

$$
\begin{aligned}
\beta_Y &= \beta_N + (\beta_Y - \beta_N) \\
\bar{X}_Y &= \bar{X}_N + (\bar{X}_Y - \bar{X}_N) \\
\Rightarrow \beta_Y - \beta_N &= \underbrace{(\omega_Y - \omega_N)}_{A} + \underbrace{\bar{X}_N'(\rho_Y - \rho_N)}_{B} + \underbrace{(\bar{X}_Y - \bar{X}_N)'\rho_Y}_{C}
\end{aligned}
$$

Substantively, A reflects unexplained differences in treatment, B captures differences in treatment effect due to Yorubas and non-Yorubas responding differently to treatment, and C captures differences in treatment effect due to differences in measured characteristics of Yoruba and non-Yoruba sellers. This decomposition specification is with respect to Yorubas although it can easily be re-written for non-Yorubas.

The dataset we use did not record seller-level information beyond their ethnic and religious identities. To demonstrate effect decomposition, we therefore simulate a pre-treatment covariate ('age') which varies differentially across ethnic groups.[9] Specifically, age is defined to be positively associated with the treatment effect but Yorubas are on average ten years younger than non-Yorubas.

We bootstrap estimates 1,000 times to generate inference on these measures and report the results in Table 3. Column (1) represents the overall difference in the high-class treatment effect between Yorubas and non-Yorubas. Columns (2) and (3) decompose these group-level differences into differences in the covariates (2) and unexplained differences (3). Columns (4) and (5) repeats the analysis loading on the non-Yoruba group.

Table 3: Effect Decomposition for Auditee Ethnic Identity using Simulated Age Data

		Yoruba Loading: Due to...		Igbo/Hausa Loading: Due to...	
	Group-level diffs in TE	...diffs in covs.	...unexpl. diffs in TE	...diffs in covs.	...unexpl. diffs in TE
Treat	(1)	(2)	(3)	(4)	(5)
Class	-0.216	-0.372**	0.156	-0.233	0.017
	(0.26)	(0.104)	(0.27)	(0.266)	(0.236)
Ethnic	-0.018	0.125	-0.143	0.205	-0.223
	(0.13)	(0.101)	(0.143)	(0.146)	(0.196)
Obs	173	173	173	173	173
DF	148	148	148	148	148
Auditee FE	Y	Y	Y	Y	Y

Notes: Effect decomposition by Yorubas and non-Yorubas for the high-class treatment (top row) and the same ethnicity treatment (bottom row). Column (1) shows the overall difference in discrimination between Yorubas and non-Yorubas for each treatment (estimated using simulated covariate data for seller's age). Columns (2) and (3) decompose these group-level differences into differences in the covariates (2) and unexplained differences (3). Columns (4) and (5) repeats the analysis loading on the non-Yoruba group. Inference is conducted via 1,000 bootstrapped simulations. * $p < 0.05$; ** $p < 0.01$.

Table 3 highlights the benefits of exploiting the full potential outcomes schedule generated

[9] Researchers using a hierarchical design with the intention of using such effect heterogeneity decompositions should make sure to collect appropriate information on pre-treatment characteristics.

with a hierarchical design. We find a significant portion of the difference between Yoruba and non-Yoruba class-based discrimination that is due to covariate differences between the two groups of sellers, specifically the age of the seller. Substantively, this finding suggests that the majority of the difference between how Yoruba and non-Yoruba sellers respond to high-class customers is due to the fact that Yoruba sellers are generally younger than Igbo and Hausa sellers (histograms of the age breakdown are summarized in Figure 9). Although not significant, the column (3) suggests that Yoruba sellers of the same age as Igbo/Hausa sellers in fact discriminate *more* based on class. Understanding this source of effect heterogeneity could be leveraged to create more impactful policy interventions or target existing policies more efficiently.

4.3.4 Meta Regression

An alternative approach to unpack effect heterogeneity is to use meta regression analysis. With multiple observations per auditee, we are able to estimate auditee-specific effects, ρ_i:

$$\rho_i = E_i[Y_{1i}] - E_i[Y_{0i}]$$

which are visualized as slopes in Figure 10. We use these auditee-specific effect estimates on group-level characteristics in a meta-regression framework where the explanatory variables are pre-treatment auditee covariates.

For our demonstration, we regress the vector of high class and same ethnicity slopes on auditee religion and ethnicity, yielding results summarized in Table 4. Bootstrapped t-statistics presented in brackets highlight the importance in accounting for uncertainty in the creation of these slopes when conducting meta regression analysis. These findings reinforce the analysis above, where class-based discrimination is primarily driven by non-Yoruba Muslims.

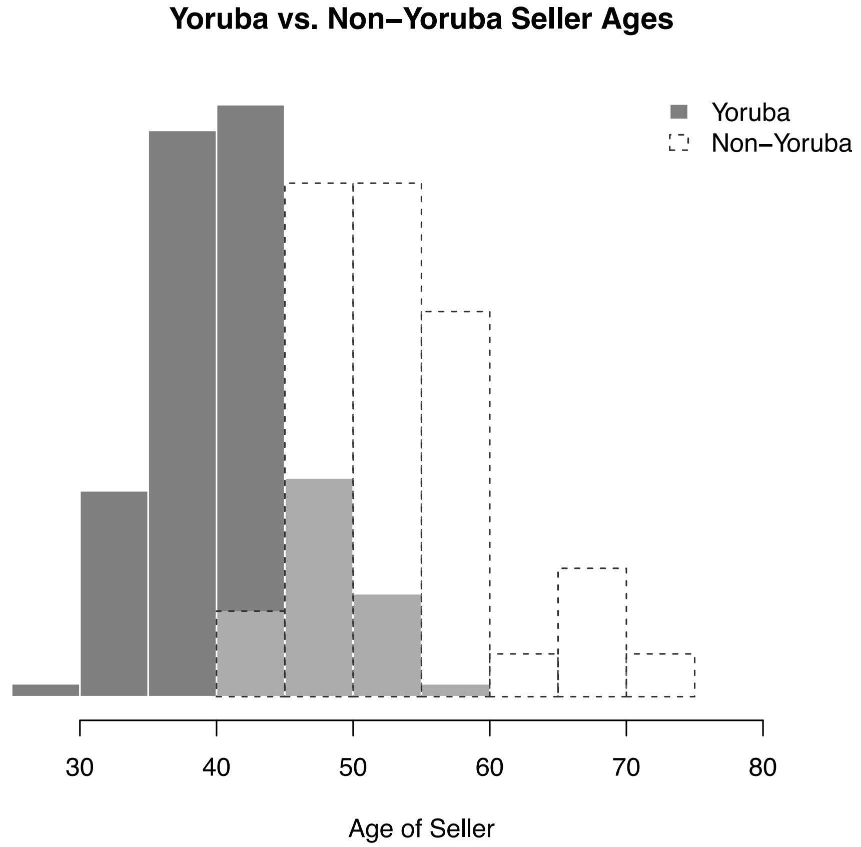


Figure 9: Age differences across sellers by ethnic identity. Data simulated to create stronger effect of class-based discrimination among older sellers.

5 Discussion

Audit experiments and correspondence studies have been instrumental in identifying the existence and magnitude of discrimination across a variety of contexts. For researchers interested in extending this method to speak to theory or evaluation policy, we propose a hierarchical audit design that exposes auditees to the full treatment regime, allowing the observed behavior under control for an individual to serve as her counter-factual behavior for what we record under treatment. This design gives researchers access to the full potential outcomes schedule, allowing for a multitude of analytic results, some of which we demonstrate

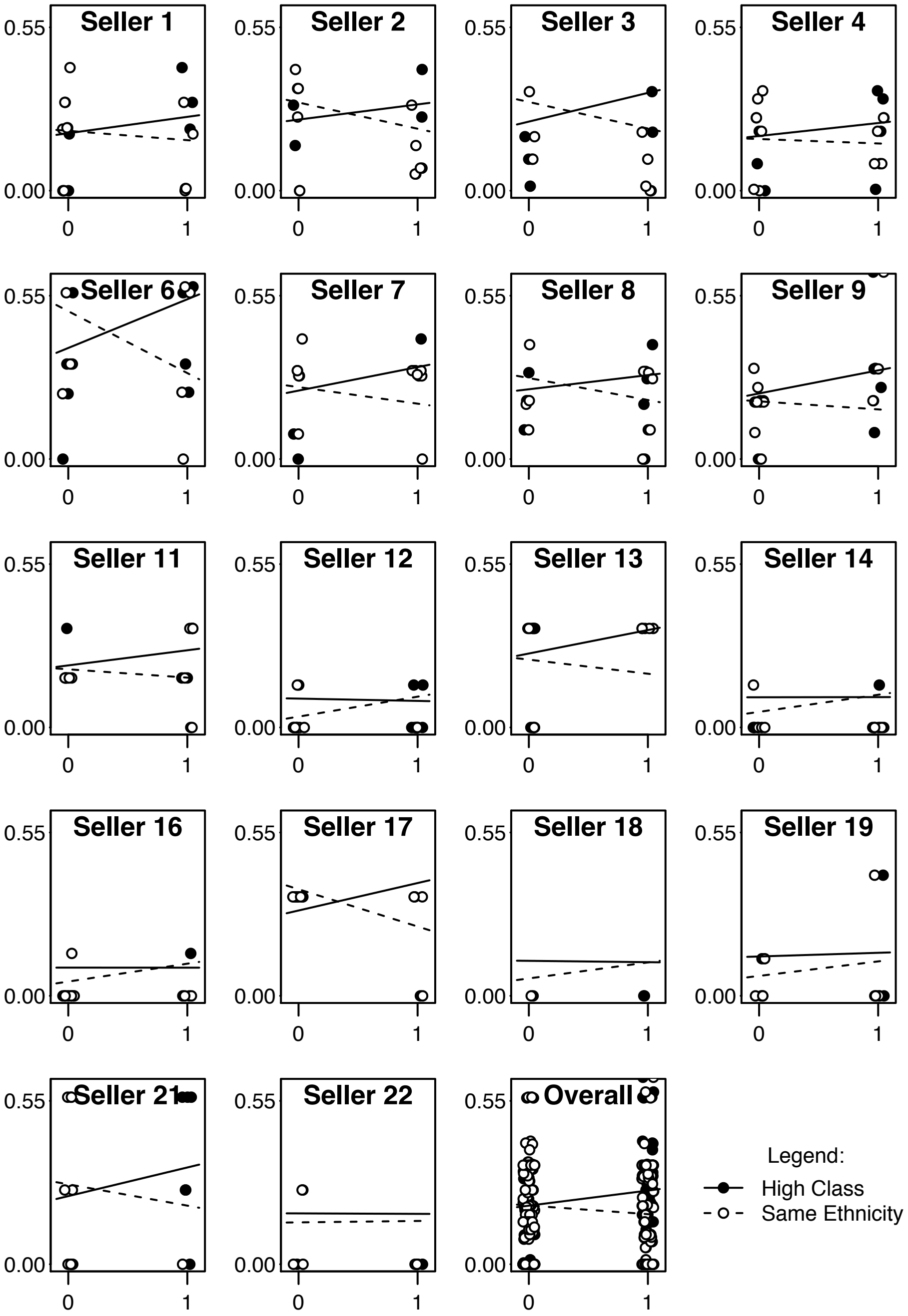


Figure 10: Allowing slopes and intercepts to vary by seller for either the high class treatment (solid points and lines) or the ethnicity treatment (hollow points and dashed lines). Treatment condition indicated on x-axis (points jittered for visual clarity) and price index indicated on y-axis. Slopes are estimated via mixed effects model.

Table 4: Meta Analysis of Effect Heterogeneity

	Ethnicity Effect	Class Effect	Religion Effect
	(1)	(2)	(3)
Yoruba Seller	0.060	-0.168^{*}	0.102
	(0.102)	(0.073)	(0.086)
	[0.44]	[-1.65]	[-0.91]
Muslim Seller	0.014	0.167^{*}	-0.036
	(0.086)	(0.061)	(0.072)
	[-0.15]	[1.93]	[-0.39]
Constant	-0.113	0.132^{*}	-0.109
	(0.080)	(0.057)	(0.067)
	[-1.08]	[1.69]	[-1.26]
Observations	20	20	20
R^2	0.033	0.343	0.077
Adjusted R^2	-0.080	0.266	-0.031
Residual Std. Error (df = 17)	0.493	0.350	0.413
F Statistic (df = 2; 17)	0.293	4.444^{**}	0.711

Notes: Effect heterogeneity predicted by seller ethnic (top row) and religious (middle row) identity. Bootstrapped t-statistics presented in brackets. * $p < 0.05$; ** $p < 0.01$.

above.

In many empirical contexts, we believe the adjustments to the conventional paired design are small and carry minimal risk. However, we acknowledge that hierarchical designs are not always feasible. There are several good reasons for limiting the number of times the experiment is run on a given auditee. All audit studies rely on deception which carries substantial risks should the auditee realize the experiment. Beyond simply souring the particular dataset, such auditee awareness can potentially poison the well for future research seeking to run experiments on the same or similar populations. As such, limiting the number of interactions per auditee is defensible for minimizing these risks.

There are other ethical concerns about wasting an auditee's time or distracting them from serving actual customers / clients / citizens. In addition to reducing the number of interactions per auditee, researchers are careful to design treatments that use only a minimum of the auditee's time, occur during off-peak hours, and are flexible enough to be discontinued should another individual request the auditee's attention.

We do not discount these ethical and experimental concerns. Our hierarchical model requires multiple observations per auditee and we recognize that it may not always be feasible. However, for cases where a hierarchical design is logistically and experimentally feasible, we believe the benefits outweigh the costs of collecting additional observations per auditee. In addition to the analytic benefits documented above, we posit that hierarchical experiments are particularly well-suited to partnering with policy-makers who may be more willing to collaborate given the potential for findings that speak to particular policy actions that could then be taken, or evaluate proposed policy impacts. The oversight provided by such partnerships would help mitigate the broader ethical concerns involving all types of audit designs, ranging from poisoning research pools to wasting the time of auditees.

The analyses summarized above demonstrate the richness afforded to researchers who use hierarchical audit designs. We find evidence consistent with both theories of discrimination but also find that taste-based discrimination dominates statistical discrimination, with high class auditors who share an ethnic identity with the seller offered the same prices as low class auditors irrespective of ethnic match. In addition, we find evidence of religious discrimination when this identity is salient, as we argue is the case for Muslim sellers on Fridays.

Beyond theory testing, we also demonstrate the ability of hierarchical designs to yield substantively important estimates for policy-makers attempting to maximize an intervention's impact. Using pilot data from an unregistered experiment, we show that Yoruba sellers disproportionately engage in discrimination against high-class buyers relative to non-Yoruba sellers. However, careful analysis of the potential outcome distributions illuminates equilibria outcomes that would be missed with a simple analysis of marginal effects. Specifically, although Yoruba sellers give co-ethnic discounts commensurate in magnitude to Igbo and Hausa sellers, this discount only brings the price down to the non-discounted level of non-Yoruba sellers.

Second, we decompose this effect into the how much can be explained by differences in the covariate profile of these groups and how much cannot. Using simulated data, we demonstrate the method by artificially making a significant portion of the difference between Yoruba and non-Yoruba sellers due to differences in their ages. Finally, we demonstrate the ability of researchers to exploit the full potential outcomes schedule in the context of a meta regression in which we estimate individual-specific discrimination behaviors for each seller and regress this vector of estimates on pre-treatment covariates.

These substantive findings warrant more careful analysis, ideally with a pre-registered hierarchical audit designed to adjudicate between the mechanisms our post-hoc application suggests. The main contribution of this paper is to summarize the benefits associated with a hierarchical audit design. While not always feasible, where appropriate, hierarchical audits dramatically expand the sophistication of analysis available to researchers. These designs are particularly well-suited to inquiries into the mechanisms of discrimination or the effect of policy interventions designed to reduce said discrimination.

References

Adida, Claire L., David Laitin and Marie-Anne Valfort. 2013. "Muslims in France: Identifying a Discriminatory Equilibrium." *Working Paper* .

Angrist, Joshua D, Parag A Pathak and Christopher R Walters. 2013. "Explaining charter school effectiveness." *American Economic Journal: Applied Economics* 5(4):1–27.

Ayres, Ian and Peter Siegelman. 1995. "Race and gender discrimination in bargaining for a new car." *The American Economic Review* pp. 304–321.

Barr, Abigail and Abena Oduro. 2002. "Ethnic Fractionalization in an African Labour Market." *Journal of Development Economics* 68(2):355–379.

Becker, Gary. 1957. *The Economics of Discrimination*. University of Chicago Press.

Bergmann, Barbara R. 1997. *In defense of affirmative action*. Basic Books.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *The American Economic Review* 94(4):991–1013.

Blinder, Alan S. 1973. "Wage discrimination: reduced form and structural estimates." *Journal of Human resources* pp. 436–455.

Feagin, Joe R. 1991. "The continuing significance of race: Antiblack discrimination in public places." *American Sociological Review* pp. 101–116.

Gill, Andrew M. 1989. "The role of discrimination in determining occupational structure." *ILR Review* 42(4):610–623.

Grossman, Shelby and Dan Honig. 2017. "Evidence from Lagos on Discrimination across Ethnic and Class Identities in Informal Trade." *World Development* 96:520–528.

Hakken, Jon. 1979. *Discrimination against Chicanos in the Dallas rental housing market: An experimental extension of the housing market practices survey.* Vol. 469 Division of Evaluation, US Department of Housing and Urban Development, Office of Policy Development and Research.

Heckman, James. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12(2):101–116.

Imai, Kosuke, Luke Keele and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological methods* 15(4):309.

Kahn, Lawrence M. 1991. "Discrimination in professional sports: A survey of the literature." *ILR Review* 44(3):395–418.

Kirschenman, Joleen and Kathryn M Neckerman. 1991. "We'd love to hire them, but...": The meaning of race for employers." *The urban underclass* 203:203–32.

Knowles, John, Nicola Persico and Petra Todd. 2001. "Racial bias in motor vehicle searches: Theory and evidence." *Journal of Political Economy* 109(1):203–229.

Michelitch, Kristin. 2015. "Does electoral competition exacerbate interethnic or interpartisan economic discrimination? Evidence from a field experiment in market price bargaining." *American Political Science Review* 109(01):43–61.

Murnane, Richard J, John B Willett and Frank Levy. 1995. The growing importance of cognitive skills in wage determination. Technical report National Bureau of Economic Research.

Neal, Derek A and William R Johnson. 1996. "The role of premarket factors in black-white wage differences." *Journal of political Economy* 104(5):869–895.

Neumark, David. 2012. "Detecting discrimination in audit and correspondence studies." *Journal of Human Resources* 47(4):1128–1157.

Neumark, David, Roy J Bank and Kyle D Van Nort. 1996. "Sex discrimination in restaurant hiring: An audit study." *The Quarterly Journal of Economics* 111(3):915–941.

Oaxaca, Ronald. 1973. "Male-female wage differentials in urban labor markets." *International economic review* pp. 693–709.

O'Neill, June. 1990. "The role of human capital in earnings differences between black and white men." *The Journal of Economic Perspectives* 4(4):25–45.

Phelps, Edmund S. 1972. "The statistical theory of racism and sexism." *The American Economic Review* 62(4):659–661.

Siegelman, Peter and J Heckman. 1993. "The Urban Institute audit studies: Their methods and findings.".

Turner, Margery Austin, Raymond J Struyk and John Yinger. 1991. *Housing discrimination study: Synthesis.* US Dept. of Housing and Urban Development, Office of Policy Development and Research.

Wienk, Ronald E et al. 1979. "Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey.".

Yinger, John. 1995. Opening Doors: How to Cut Discrimination by Support Neighborhood Integration. Policy Brief No. 3 Syracuse University, Maxwell School of Citizenship and Public Affiairs; Center for Policy Research.

Yinger, John. 1998. "Evidence on discrimination in consumer markets." *The Journal of Economic Perspectives* 12(2):23–40.