Bias in the State: Theory and Evidence from Peru

Michael Best[1,5,6,7], Leonard Goff[3], Jonas Hjort[2,6,7,8], Dafne Murillo[1], and Gastón Pierri[4]

[1]Columbia University
[2]University of Calgary
[3]University College London
[4]Inter-American Development Bank
[5]NBER
[6]BREAD
[7]CEPR
[8]University of Oslo

This version: March 21, 2025

Abstract

We use administrative data and a triage experiment with bureaucrats in Peru to study disparate treatment in the handling of citizen whistleblower allegations of corruption. We develop a novel statistical framework in which to decompose the sources of disparate treatment with two key features. First, it allows us to incorporate three commonly-discussed channels through which disparate treatment occurs: tastes, statistical discrimination, and incorrect beliefs. Second, it features sequential encounters between citizens and state actors. We operationalize the framework through a triage experiment with anti-corruption officials in Peru to study disparate treatment of defendants with Indigenous vs European names. We find that officials overall display bias against Indigenous defendants, with belief-based bias against European defendants and taste-based bias against Indigenous defendants. Analysts systematically overestimate the probability of indictment for cases, but they do so more for European defendants than for Indigenous defendants, underscoring the value of observing beliefs when implementing marginal outcome-based tests for disparate treatment.

1 Introduction

Uniform treatment of citizens in the state's functions is enshrined in most constitutions. Prima facie, the ubiquity of differential treatment—or seemingly differential treatment—of some groups of citizens by states themselves is thus surprising. However, not all forms of seemingly differential treatment of for example ethnic minorities in fact are *bias* in the legal sense: this is part of the reason why the social science and legal literature attempting to tell apart different forms of differential treatment is very large. The other part is that the effective policy responses to different forms of differential treatment oftentimes differ.

In the last decade economists have drawn on other social sciences to begin to account for and attempt to measure a broader spectrum of forms of differential treatment. This is important: inaccurate beliefs or stereotypes about minorities, for example, clearly warrant a different response than the more familiar concepts of "taste-based" and (accurate) "statistical" discrimination (Arnold et al., 2018; Arnold et al., 2022; Bohren et al., 2019; Lang and Lehmann, 2012). A few studies have begun to move towards decomposing overall differential treatment into its components (Bohren et al., 2024; Hurst et al., 2024; Bohren et al., 2025). Bias on part of a decision-maker (such as e.g. a judge or anti-corruption bureaucrat) can only be defined and measured relative to some downstream or upstream outcome (Anwar et al., 2012; Alesina and La Ferrara, 2014; Argyle et al., 2024), however. This matters for measurement: upstream behavior that would look biased if the downstream outcome against which it is benchmarked was unbiased may instead look unbiased if the downstream outcome is biased.[1] It also matters for policy: a planner trying to minimize total, experienced bias over time would need to know how upstream decisions respond to downstream bias and vice versa.

Sequential encounters with a single apparatus or far-reaching organization are a core and often overlooked reason why accurately measuring and decomposing forms of differential treatment is so difficult. They are also the defining feature of individual citizens' interactions with the state to which they belong over their lives. Few citizens, for example, have the option to "change state" by emigrating, and the market forces that constrain discriminatory behavior by firms rarely constrain states effectively.

In this paper we develop a comprehensive framework that includes each broad form of bias (or discrimination) considered in the literature—pure taste-based discrimination (Becker, 1971; Arrow, 1973), accurate statistical discrimination (Aigner and Cain, 1977;

[1] An illustration is the Supreme Court decision in Griggs v. Duke Power Co. (1971), where the court held that the firm's policy requiring a high school diploma for internal transfers unfairly disadvantaged Black employees, who were qualified but often lacked diplomas due to historical educational discrimination. Although the upstream diploma requirement appeared race-neutral, it disproportionately harmed Black employees and had no relation to actual job performance, leading the Court to deem it discriminatory.

Phelps, 1972), belief-based discrimination (Bohren et al., 2025; Canay et al., 2023)—and show how to empirically quantify each.[2] We first do so in a "single-phase" scenario, and thereafter a multi-phase one, explicitly allowing for interdependence between the different forms of bias across the two stages. We then apply the framework to study bias against the indigenous minority in a particular segment of the Peruvian state: its anti-corruption agency's handling of whistleblower reports against public officials.[3] Within the agency there are multiple, nominally independent stages of decision-making, and the accused encounter the state also upstream and downstream (for example in the criminal system in cases that proceed further downstream).

To empirically quantify these different forms of bias, we conduct a lab-in-the-field triage experiment involving analysts from Peru's national anti-corruption agency (CGR). Analysts at CGR evaluate whistleblower complaints in a two-step process: first, they decide whether a case should be *escalated*, meaning advanced internally within CGR for further investigation; second, based on the findings from that investigation, analysts determine whether the case results in an *indictment*, which triggers a full-scale audit against the defendant. We experimentally manipulate the implied ethnicity of defendants by randomly changing surnames from European to Indigenous (and vice-versa) in complaints drawn from CGR's historical data. Participants evaluate these cases and then assign them continuous priority scores. We also ask them to reveal their beliefs about the indictment probability of each case.

Approach and findings. Given the nature of the data in our experimental setting, we make several methodological contributions to the empirical analysis of discrimination. First, we provide a general model of biased decision-making in which a decision-maker takes an action on the basis of their (potential inaccurate) beliefs h. Belief-based bias—including accurate statistical discrimination as a special case—occurs when these beliefs depend directly on race, and taste-based bias occurs when the *mapping* between beliefs and actions—which we denote as ϕ_r—depends directly on race r. Importantly, our framework nests the case of a binary action considered by existing work as well as the continuous action space provided to analysts in our lab experiment.

[2]Our framework combines two predominant methodologies for measuring discrimination: *audit studies*, which experimentally vary group membership (e.g., race or ethnicity) of otherwise comparable subjects to detect differential treatment (Bertrand and Mullainathan, 2004), and *marginal outcome tests*, which statistically infer bias from observed decision-maker thresholds (Arnold et al., 2018; Arnold et al., 2022; Canay et al., 2023).

[3]In Peru, where 25% of the population self-identified as Indigenous Quechua or Aymara (Census, 2017), discrimination remains a pervasive concern, with approximately 60% of Peruvians acknowledging discrimination against Indigenous people (Ministerio de Cultura, 2018). Empirical evidence from a field experiment by Galarza and Yamada, 2014 further highlights this issue, documenting significant labor market discrimination against job applicants perceived as Indigenous.

In the case of a binary action, existing work has estimated taste-based bias through the so-called "marginal outcomes test" (MOT), leveraging the quasi-random assignment of decision-makers to cases in order to overcome the endogeneity that arises from infra-marginal cases, using leniency as an instrumental variable (Arnold et al., 2018; Hull, 2021). We implement the conventional MOT method in our analysis of analysts' historical decisions of whether to escalate a case for further scrutiny, in administrative data. From this we find suggestive evidence of taste-based bias against Indigenous defendants.

We extend the MOT methodology to our lab-experiment setting with a continuous action, focusing on analysts assignment of "priority scores" from 0 to 100 to cases they are shown. In the continuous setting, every case is "marginal" for some level of the action, so infra-marginality bias is avoided quite simply by comparing cases that differ by ethnicity but share the same priority score. This greatly simplifies estimation, as instrumental variables become unnecessary to detect taste-based bias.

However, estimates of taste-based bias using the MOT do require a strong restriction on decision-makers' choice rule, which Canay et al. (2023) refer to as the Extended Roy Model assumption (ERM). Our model introduces and then imposes a relaxed version of ERM, which assumes only that there exist some characteristics of cases that are excluded from the function $\phi_r(\cdot)$, but still affect beliefs h. This notion of ERM is defined in the context of a general action space. We find that in the case in which actions are continuous—as they are in our lab experiment—our weakened version of ERM yields a testable implication if beliefs h are elicited (or otherwise observed). We implement a statistical test for this implication of ERM, and find that the behavior of most of the analysts in our lab experiment are compatible with the ERM restriction.

Extending our analysis of the lab-experiment setting, we then introduce a method to empirically decompose taste-based and belief-based bias under ERM from the total bias that is identified using audit study variation (which will generally affect both h and ϕ by flipping the apparent race of a case). This decomposition requires a definition of each component of bias on a common scale. We achieve this by introducing a concept of "belief-equivalent bias". Belief equivalent bias captures the adjustment to the beliefs that a decision-maker has about a particular case from one ethnic group which would move the decision-maker to the same margin of action as a case from the opposite group that is otherwise observationally equivalent. Differences in belief-equivalent bias across races capture both taste-based and belief-based sources of bias, with the first component recovering the traditional MOT estimand in the case of a binary action.

In our audit study setting with a continuous action space, we show that belief-equivalent bias is identified from an experimental difference-in-means after a particular transformation of the observed priority scores. An alternative transformation captures belief-based bias alone, and we can thus separate this bias from taste-based bias. The

decomposition is possible without observing beliefs provided that analysts' beliefs are accurate and the researcher observes the variable that analysts target with their actions (in our context, whether a case is "indictable"). Given that we are able to observe beliefs, we relax this assumption and can avoid using downstream outcomes in estimating the decomposition. We find that assuming analysts to have accurate beliefs about indictment suggests belief-based bias against Indigenous defendants along with taste-based bias against Europeans, while using elicited beliefs suggests the opposite: belief-based bias against European cases along with taste-based bias against Indigenous defendants, leading to a small total bias against Indigenous defendants.

We reconcile these findings by estimating "true" indictment probabilities from a random forest prediction of indictment given the rich covariates we observe about each case, defined from the full information displayed to analysts in the lab experiment. Our estimates suggest that analysts systematically overestimate the probability of indictment for cases, but they do so more for European defendants than for Indigenous defendants. This implies that belief-based bias against Indigenous cases is smaller than the "accurate statistical discrimination" that analysts would employ if they had accurate beliefs. Incorrectly assuming accurate beliefs attributes this differential to belief based bias, leading to misleading estimates for both belief-based and taste-based bias. This underscores the value of observing beliefs when implementing marginal outcomes tests.

2 Bias in single-phase decisions

2.1 Basic model

We model bias in a setting in which a decision-maker j faces a a single decision in which they choose an action $a \in \mathcal{A}$ after observing r and x about a particular *case*. Here r denotes an identity characteristic (e.g. "race") of the person that the case is about, and x all other variables regarding the case. We refer to the decision-maker as an *analyst* and the person that the case is about as the *accused*.

We assume that decisions a for a given analyst are driven by a scalar index variable $h_j(r, x)$ that can depend generally on case observables (r, x), as well as individual aspects of the analyst j (e.g. individual knowledge, cognitive biases, etc.). We refer to $h_j(r, x)$ as analyst j's *belief* about case (r, x). Decisions can also depend arbitrarily on a strict subset of the observables x that we denote as v. Components of x that are not included in v only affect actions through $h_j(r, x)$. Following Canay et al. (2023) we refer to this restriction as an "Extended Roy Model" or (ERM).

Our notion of an ERM weakens the restriction considered by Canay et al. (2023) by allowing for the function ϕ to depend on a subset of the x through v. Formally, we

suppose that

$$a = \phi_{jrv}(h_j(r, x)) \tag{1}$$

for some increasing and right-continuous function ϕ_{jrv}.[4]

Example 1: binary actions. Most of the existing literature on the econometric evaluation of discrimination considers binary actions $\mathcal{A} = \{0, 1\}$. For binary actions, ϕ_{jrv} cannot be a *strictly* increasing function (provided that $h_j(r, x)$ takes on more than two values). In our analysis of the historical escalation decision, we accordingly assume ϕ_{jrv} only to be *weakly* increasing. In the binary case, we can without loss of generality write $\phi_{jrv}(h) = \mathbb{1}(h \geq t_{jrv})$, where t_{jrv} represents a threshold level of belief at which j takes the action labeled 1, for a case with observables (r, v).[5]

Example 2: continuous actions. When $\mathcal{A}$ is an ordered continuum, such as in our lab experiment for the priority score, we assume ϕ_{jrv} to be *strictly* increasing. The intuitive meaning is that if j's belief about case (r, x) increases, their action must increase as well.

Remark: Eq. (1) can be derived from a utility-maximization model in which $h_j(r, x)$ represents analyst j's subjective, evaluation that a given decision is "guilty", indicated by a binary random variable Y^*. That is, $h_j(r, x) = \mathbb{E}_j[Y^*|R = r, X = x]$ where $\mathbb{E}_j[\cdot|\mathcal{I}]$ indicates analyst-j's (potentially inaccurate) subjective beliefs given information $\mathcal{I}$. In this case, suppose that j choses a to maximize

$$\mathbb{E}_j[u_j(Y^*, a, r, v)|R = r, X = x] = u_j(0, a, r, v) + h_j(r, x) \cdot \{u_j(1, a, r, v) - u_j(0, a, r, v)\}$$

over $a \in \mathcal{A}$. The maximizing a of expected utility, assuming it is unique, can given the above be written without loss of generality as $\phi_{jrv}(h_j(r, x))$. For $\phi_{jrv}(\cdot)$ to be weakly-increasing, it is sufficient that $\{u_j(1, a, r, v) - u_j(0, a, r, v)\}$ be weakly increasing in a.[6] If the analyst suffers a utility loss $\{u_j(1, a, r, v) - u_j(0, a, r, v)\} < 0$ from the recipient of the decision being guilty rather than innocent, the above property says that this loss is smaller the "larger" the action a is. This is an intuitive property if a is a punishment or mitigating action such as conviction, investigation, or incarceration.

[4]Right-continuity of ϕ_{jrv} is essentially a normalization: by Lebesque's theorem for monotone functions, ϕ_{jrv} can only have isolated points at which it is discontinous. At the countable set of values h that correspond to these discontinuities, we assume $\phi_{jrv}(h)$ to be equal to its left limit at h.

[5]Existing studies consider a binary action, e.g. pre-trial release, that is taken when beliefs (about guilt or misconduct potential) are *low*. For example , the models of Arnold et al. (2018) and Canay et al. (2023) can be written as $a = \mathbb{1}(h_j(r, x) \leq t_{jrv})$, where $a = 1$ indicates release. This can be adapted to the form of Eq. (1) by redefining $a = 1$ to be non-release (while changing the weak inequality into a strict inequality to yield right-continuity rather than left-continuity of the resulting increasing function).

[6]In this case the function $f_{jrv}(a, h) := u_j(0, a, r, v) + h \cdot \{u_j(1, a, r, v) - u_j(0, a, r, v)\}$ satisfies the property of *increasing differences*, i.e. that for $a' \geq a$, $f_{jrv}(a', h) - f_{jrv}(a, h)$ is monotone decreasing in h (Milgrom and Shannon, 1994).

2.2 Defining belief-based and taste-based bias

Consider two values of r, which we denote as e and i for European and Indigenous, respectively.

We say that *belief-based bias* (BBB) holds for the case-analyst pair (j, r, x) when $h_j(i, x) \neq h_j(e, x)$: that is beliefs would differ between two cases with different races but identical other observables. We say that *taste-based-bias* (TBB) holds when the mapping between these beliefs and actions is different: $\phi_{jix} \neq \phi_{jex}$, i.e. the analyst would treat Indigenous and European cases differently even if they shared the same belief about them.

To measure the magnitudes of TBB and BBB both on a common scale, we introduce the concept of *belief-equivalent bias*. Let

$$\delta_{jev} := \sup\{\delta : \phi_{jev}(h_j(e, x) + \delta) \leq \phi_{jiv}(h_j(i, x))\} \tag{2}$$

δ_{jev} is the adjustment to j's belief that would just push them to give a European case with non-race observables x at least as a high an action than an Indigenous case with these same observables. δ_{jev} is strictly positive when the effect of changing r from e to i on the action taken by j given x: $\phi_{jiv}(h_j(i, x)) - \phi_{jev}(h_j(e, x))$, is strictly positive, allowing a strictly positive shift to beliefs δ for a European without leading to a larger action than the one taken against an observationally-equivalent Indigenous case. The condition that $\phi_{jiv}(h_j(i, x)) > \phi_{jev}(h_j(e, x))$ can occur when there is TBB against Indigenous cases: $\phi_{jiv}(h) > \phi_{jev}(h)$ for a given h, when there is BBB against Indigenous cases: $h_j(i, x) > h_j(e, x)$, or a combination of both.

We can similarly define an alternative measure of belief-equivalent bias that puts δ within the ϕ_{jiv} function rather than the ϕ_{jev} function:

$$\delta_{jiv} := \sup\{\delta : \phi_{jev}(h_j(e, x)) \geq \phi_{jiv}(h_j(i, x) + \delta)\} \tag{3}$$

The quantity δ_{jiv} is the adjustment to j s beliefs about an Indigenous case with non-race observables x that would just push it over the edge of delivering as high an action than for an observationally-equivalent European case. We have $\delta_{jiv} > 0$ when $\phi_{jiv}(h_j(i, x)) < \phi_{jev}(h_j(e, x))$, in exactly the opposite situation as when δ_{jev} is positive. We can thus define a symmetric notion of the total belief-equivalent bias against Indigenous cases from the difference:

$$TOT_{jx} := \frac{1}{2}(\delta_{jev} - \delta_{jrv})$$

We can obtain an explicit expression for δ_{jiv}, δ_{jiv} and hence TOT_{jx} by defining $\phi_{jrv}^{-1}(a) := \inf\{h \in \mathbb{R} : \phi_{jrv}(h) > a\}$. In the case of continuous actions, ϕ_{jrv}^{-1} is simply the inverse of the strictly increasing function ϕ_{jrv}. If ϕ_{jrv} is only weakly increasing, then ϕ_{jrv}^{-1} is a generalized inverse function of ϕ_{jrv} thta maintains many of the same properties.

Proposition 1. *The following hold:*

$$\delta_{jev} = \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(e,x)$$

$$\delta_{jiv} = \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x))) - h_j(i,x)$$

Proof. Since ϕ_{jev} is increasing and right-continuous, it follows that $\phi_{jev}(h_j(e,x)+\delta) \leq \phi_{jiv}(h_j(i,x))$ if and only if $h_j(e,x)+\delta \leq \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x)))$ (see e.g. item 5 of Proposition 2.3 in Embrechts and Hofert 2013 for a proof). Thus

$$\delta_{jev} = \sup\{\delta : h_j(e,x)+\delta \leq \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x)))\} = \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(e,x)$$

The case of δ_{jiv} is symmetric.

□

Note that Proposition 1 holds regardless of whether $\mathcal{A}$ is binary, discrete, or continuous. As a consequence of Proposition 1, we have that:

$$TOT_{jx} = \frac{1}{2}\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x))) + h_j(i,x) - h_j(e,x)\} \tag{4}$$

Expression (4) is useful in showing how TOT_{rx} combines TBB and BBB. We see this by considering the binary and continuous cases each in turn.

Example 1: binary actions. Suppose that $\phi_{jrv}(h) = \mathbb{1}(h \geq t_{jrv})$. Then $\phi_{jrv}^{-1}(0) = t_{jrv}$ and $\phi_{jrv}^{-1}(1) = \infty$. Suppose that $\phi_{jiv}(h_j(i,x)) = \phi_{jev}(h_j(e,x)) = 0$ so that δ_{jev} and δ_{jiv} are both well-defined. Then TOT_{jx} provides an average of bias in tastes and bias in beliefs:

$$TOT_{jx} = \frac{1}{2}\underbrace{\{t_{jev} - t_{jiv}\}}_{\text{taste-based bias}} + \frac{1}{2}\underbrace{\{h_j(i,x) - h_j(e,x)\}}_{\text{belief-based bias}},$$

Notice that each term is measured on a comparable scale of beliefs.[7] A describes how as-good-as-random assignment to analysts can be used to isolate an average of $t_{jev} - t_{jiv}$ across judges.

Example 2: continuous actions. Consider the expression for δ_{jev} given in Proposition 1. By adding and subtracting $h_j(i,x)$, we can rewrite it as:

$$\begin{aligned}\delta_{jev} &= \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(e,x) \\ &= \underbrace{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x)}_{\text{taste-based bias}} + \underbrace{h_j(i,x) - h_j(e,x)}_{\text{belief-based bias}}\end{aligned}$$

[7] If instead $\phi_{jiv}(h_j(i,x)) = \phi_{jev}(h_j(e,x)) = 1$ and we replace ∞ with a large fixed number $\bar{H}$ rather infinity in the definition of ϕ_{jrv}^{-1}, then the $\bar{H}$ cancel out and $TOT_{jx} = 1/2 \cdot (h_j(i,x) - h_j(e,x))$ captures belief-based bias alone.

When ϕ_{jev} is invertible, the first term can only differ from zero if $\phi_{jev} \neq \phi_{jiv}$ as a function. Meanwhile $h_j(i,x) - h_j(e,x)$ captures BBB alone. Thus δ_{jev} captures a sum of TBB and BBB. By similar steps one sees that $\delta_{jiv} = \underbrace{\phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x))) - h_j(e,x)}_{\text{taste-based bias}} - \underbrace{\{h_j(i,x) - h_j(e,x)\}}_{\text{belief-based bias}}$

and thus altogether:

$$TOT_{jx} = \underbrace{\frac{1}{2}\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x)\} + \frac{1}{2}\{h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\}}_{\text{taste-based bias}} + \underbrace{h_j(i,x) - h_j(e,x)}_{\text{belief-based bias}}$$

Thus we have seen both in the binary and in the continuous action settings TOT_{jx} captures a linear combination of TBB and BBB: in the continuous action case its two components δ_{jrx} also have this interpretation for each $j \in \{i, e\}$.[8]

3 Single-decision bias within multi-phase systems

Consider the case of the lab experiment decision to assign priority scores to a given case. We can situate our notion of bias introduced above within the many stages of a report in the CGR by explicitly accounting for bias that may occur in the arrival of the associated true report to the CGR, as well as bias in the actual outcome that is observed for that report.

Recall the decomposition of total belief-equivalent bias TOT_{jx} into TBB and BBB:

$$\underbrace{\frac{1}{2}\left\{\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x)\} + \{h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\}\right\}}_{\text{taste-based bias}} + \underbrace{h_j(i,x) + h_j(e,x)}_{\text{belief-based bias}}$$

Let us denote by $A = 1$ the event that a given a "potential case" results in an actual case or a *report*. In this case, we say that the case is "accused" (A=1). Let Y^* denote that a case would be indicted were it escalated, i.e. it is "inidictable". Let Y indicate observed indictment status of the case, so that $Y = Y^*D$ where D indicates whether the case was actually escalated. Finally, let G^* indicate whether the individual accused in a given case was actually guilty. Note that this is latent, and there will generally exist both guilty potential cases that never get accused ($A = 0$ and $G^* = 1$), as well as innocent cases that get accused and perhaps even indicted ($A = D = Y^* = 1$ but $G^* = 0$ in the latter case).

Define $\tilde{f}^*(r,x) = \mathbb{E}[G^*|R = r, X = x]$ to be the true rate of guilt among potential cases with observables r, x, and let $g_A^*(r,x) = \mathbb{E}[G^*|R = r, X = x, A = 1]$ be same among reports. For a given r and x, the gap $f^*(r,x) - \tilde{f}^*(r,x)$ measures selection into accusation.

[8] By contrast, in the binary action case note that for example when $\phi_{jiv}(h_j(i,x)) = 0$ we have that $\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(e,x) = t_{jev} - h_j(e,x)$ which is not interpretable as a measure of TBB: it can differ from zero even if $t_{jev} = t_{jiv}$. It is for this reason that we must appeal to the symmetric measure TOT_{jx} of total belief based bias directly in the binary action case in order to see a decomposition into TBB and BBB, while we can decompose each term δ_{jrv} separately in the continuous case.

Let $h^*(r,x) = \mathbb{E}[Y^*|R=r, X=x, A=1]$ be the rate of indictability among accused cases with observables r, x. The gap $h^*(r,x) - f^*(r,x)$ measures *over-indictment* for cases with observables r, x.

Define $d(r,x) := P(D=1|R=r, X=x, A=1)$ to be the observable escalation probability for a report with r, x, and $\tilde{h}(r,x) := \mathbb{E}[Y^*|R=r, X=x, D=1]$ (note that $D=1$ implies $A=1$ so we can leave that in the conditioning event). Note that in the absence of multiple analysts that can disagree about whether to escalate a case, $d(r,x)$ would be equal to zero or to one for each r, x assuming analysts make deterministic choices given case observables. If $d(r,x) = 0$, we let $\mathbb{E}[Y^*|R=r, X=x, D=1] := 0$ so that it remains well-defined. Let

$$h(r,x) = \mathbb{E}[Y|R=r, X=x, A=1] = \mathbb{E}[Y^*D|R=r, X=x, A=1] = d(r,x) \cdot \tilde{h}(r,x)$$

denote the observable indictment probability among reports with observables r, x. Consider the gap $h(r,x) - h^*(r,x) = (h(r,x) - \tilde{h}(r,x)) + (\tilde{h}(r,x) - h^*(r,x))$. The first term measures how much more likely escalated reports are to be guilty compared with unescalated reports having r, x. This term will tend to be zero if analyst esclation preferences are uncorrelated with case observables. This term will be zero in the limit of no analyst heterogeneity to lead to $d(r,x) \in (0,1)$.

Collecting, we have that

$$h(r,x) = \underbrace{\tilde{f}^*(r,x)}_{\text{true guilt probability}} + \underbrace{f^*(r,x) - \tilde{f}^*(r,x)}_{\text{accusation selection}} + \underbrace{h^*(r,x) - f^*(r,x)}_{\text{overindictment}} + \underbrace{h(r,x) - h^*(r,x)}_{\text{escalation analyst heterogeneity}} \tag{5}$$

With these definitions, we can write:

$$\begin{aligned} TOT_{jx} &= \underbrace{\frac{1}{2}\left\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x) + h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\right\}}_{\text{taste-based bias}} + \underbrace{h_j(i,x) + h_j(e,x)}_{\text{belief-based bias}} \\ &= \underbrace{\frac{1}{2}\left\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x) + h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\right\}}_{\text{taste-based bias}} \\ &\quad + \underbrace{\{h_j(i,x) - h(i,x)\} - \{h_j(e,x) - h(e,x)\}}_{\text{differentially-incorrect indictment beliefs}} + \underbrace{h(i,x) - h(e,x)}_{\text{accurate statistical discrimination}} \end{aligned} \tag{6}$$

where using Eq. (5) we can unpack the ASD term to write:

$$
\begin{aligned}
TOT_{jx} = & \underbrace{\frac{1}{2}\left\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x) + h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\right\}}_{\text{taste-based bias}} \\
& + \underbrace{\{h_j(i,x) - h(i,x)\} - \{h_j(e,x) - h(e,x)\}}_{\text{differentially-incorrect indictment beliefs}} + \underbrace{\tilde{f}^*(i,x) - \tilde{f}^*(e,x)}_{\text{true difference in rate of guilt}} \\
& + \underbrace{\{f^*(i,x) - \tilde{f}^*(i,x)\} - \{f^*(i,x) - \tilde{f}^*(i,x)\}}_{\text{differential accusation selection}} \\
& + \underbrace{\{h^*(i,x) - f^*(i,x)\} - \{h^*(i,x) - f^*(i,x)\}}_{\text{differential overindictment}} \\
& + \underbrace{\{h(i,x) - h^*(i,x)\} - \{h(i,x) - h^*(i,x)\}}_{\text{differential escalation analyst heterogeneity effect}}
\end{aligned}
$$

Appendix A.0.1 shows that in principle the sign of differential accusation selection can be identified if the econometrician observes all of the variables x that are seen by the analyst, by comparing the distribution of X among reports to that in the larger population of potential cases.

4 Econometric model for the lab experiment

We now map the notions of bias introduced in Section 2 to our observables in the lab experiment. First, we show that the ERM assumption has a testible implication in the case of continuous actions. Then, we show identification of averages of TBB and BBB across analysts under ERM, which affords a test for the presence of each type of bias, as well as a quantitative decomposition of total belief-equivalent bias into both.

4.1 Notation

The unit of analysis is a pair (c, j), where c denotes a real-world case and j an analyst in the lab experiment. We refer to such a pair as a "case-pair", and analysts in the lab experiment as lab-analysts. Each case has observables (R_c, X_c, Y_c), where (R_c, X_c) is the information available to the lab-analyst and Y_c is an indicator for whether the case actually resulted in an indictment (unknown to the lab analyst).

We use capital letters without indices to denote random variables defined over a population of *case-pairs* in the lab experiment (such case pairs, rather than cases c, are the unit of analysis in the lab experiment). The primary variables of interest are (R, X, F, Y, J, S, H), where F indicates whether the ethnicity of the case was "flipped" before showing it to the lab-analyst, J indicates the lab-analyst of that case-pair, S is the priority score assigned by the lab-analyst, and H is the subjective likelihood of indictment elicited from the lab analyst.

We let $\tilde{R}$ denote the "apparent" race of a case by the lab-analyst, where

$$\tilde{R} = \begin{cases} e & \text{if } R = e \text{ and } F = 0 \\ i & \text{if } R = i \text{ and } F = 1 \\ e & \text{if } R = i \text{ and } F = 0 \\ i & \text{if } R = i \text{ and } F = 1 \end{cases}$$

where $R \in \{i, e\}$ are shorthand for Indigenous and European, respectively.

As apparent race $\tilde{R}$ is what is manipulated experimentally, we introduce potential outcomes for our key outcome variables in the lab experiment:

Definition (counterfactual outcomes). *$s_j(r, x)$ denotes the priority score that analyst j would assign to a case with observables (r, x), and thus $S = s_J(\tilde{R}, X)$. Similarly, let $h_j(r, x)$ be potential outcomes for an analysts' subjective likelihood of indictment.*

We do not assume that realized subjective likelihoods $h_J(\tilde{R}, X)$ are directly observable. Rather, we allow each case-pair to have a realization U of measurement error in the observed subjective likelihood scores o that $H = H^* + U$ where $H^* = h_J(\tilde{R}, X)$. We do not allow for such measurement error in scores S, as S represents the action "actually" taken by the lab-analyst in the lab experiment. The existence of U allows for H to have a non-degenerate distribution conditional on S (and $\tilde{R}, J, V$), and vice-versa. Such non-degeneracy is observed in our data.

4.2 Identifying assumptions

Our first and most important assumption for identification is that an analyst's subjective probability that the case will result in an indictment represents an index that drives their priority score decisions in an invertible manner.

Assumption ERM (extended Roy model with a continuous action space). *$s_j(r, x) = \phi_{jrv}(h_j(r, x))$ for some strictly increasing function ϕ_{jrv}, where v is a strict subset of the observable components of x.*

Assumption ERM embeds a few key restrictions. Most substantively, the function ϕ_{jr} does not depend on the full vector x. If v were instead allowed to be equal to the full set of observables x, then assumption ERM would hold trivially.[9] In the other extreme, if no variables are contained in v then Assumption ERM extends the notion from Canay et al. (2023) of an "extended Roy model" (ERM) to our setting in which the action space (priority scores) is continuous rather than binary.

[9]In the case of $v = x$ ERM would amount to assuming that $s_j(r, x) = \phi_{jrx}(h_j(r, x))$ for some strictly increasing function ϕ that can depend directly on x. Then ERM would become trivial because whatever the functions h_j and s_j are, we could simply define $\phi_{jrx}(h)$ as $\phi_{jrx} = s_j(r, x) - h_j(r, x) + h$.

For "intermediate" v (embeding some but not all of the observables X), Assumption ERM allows the preferences of lab-analysts over values of the priority score to assign to depend on certain observables of the case-pair beyond race: for example the type of violation that the defendant is accused of. This is important if lab-analysts see indictments for e.g. illicit enrichment as more valuable than e.g. favoratism.

Assumption ERM does not take a strong stand on what analysts "target": it is for example compatible with analysts attempting to maximize indictments (regardless of guilt), and is also compatible with analysts attempting to maximize indictments among cases in which the analyst feels that the accused is actually guilty, or maximizing indictments in which the accusation is "severe", etc. Whatever the analysts' objective function, ERM requires priority scores to be strictly increasing in the analyst's own perception of the likelihood of indictment, between cases in which case properties v are held fixed. Note that ERM allows for analyst-specific heterogeneity in tastes for priority score, though the set v is taken to be common over all analysts (intuitively, one can take v to be the union of the variables that any analysts care about).

Our second assumption is that randomization was implemented succsefully:

Assumption RAND (randomization of lab-analysts and ethnicity flips). *J, F, and (R, X, Y, U) are mutually independent*

Our third assumption is that measurement error in subjective indictment probabilities is classical in the following sense:

Assumption CLASSICAL (classical measurement error). $\{U \perp\!\!\!\perp X\}|J, \tilde{R}, R$ *and* $\mathbb{E}[U|J, \tilde{R}, R] = 0$

Assumption CLASSICAL permits a idiosyncratic mean-zero gap from analysts' self-reported subjective indictment probability and the one that they actually use in making their decision of S.

4.3 An observable implication of ERM with continuous actions

The key implication of ERM for our analysis is that if a given analyst j were to give the same priority score to two cases that share a race and non-race characteristics v (but differ in other non-race characteristics not included in v), they must also share the same subjective probability of indictment to that analyst. In particular, if x and x' share the same value of v, then:

$$s_j(r, x) = s_j(r, x') \implies h_j(r, x) = h_j(r, x') \tag{7}$$

Eq. (7) is implied immediately by Assumption ERM given that the function ϕ_{jvr} is invertible. In fact in the case Eq. (7) is essentially equivalent to Assumption ERM: it

says that for a given j, r and v, $h_j(r,x)$ can be written as a one-to-one function of $s_j(r,x)$, which if one assumes to be continuous and weakly increasing implies that it is in fact strictly increasing (see e.g. Theorem 18.6 of Ross 2003). Note however that Eq. (7) is not possible in conventional MOT settings in which only a binary action is taken by the agent: invertibility of actions into indictment probabilities is a feature of our setting in which both can take values on a continuum.

Not that the converse of (7), that for any x and x' that share the same value of v:

$$h_j(r,x) = h_j(r,x') \implies s_j(r,x) = s_j(r,x'), \tag{8}$$

is equivalent to the statement that $s_j(r,x) = \phi_{jrv}(h_j(r,x))$ for *some* function ϕ_{jrv} (not necessarily invertible). This weaker implication is natural in settings with a discrete rather than continuous action space, for example the binary decision to release a defendent on bail. Canay et al. (2023) show that if one assumes that (8) holds for *any* vectors of non-race characteristics x and x' (implying that ϕ_{jr} does not depend on v), the conventional marginal outcomes-test can be used to draw inferences about taste-based discrimination.

However, this is a strong assumption, which we relax in Assumption ERM by allowing ϕ_{jrv} to depend on v. However this leaves unresolved the question of what v is: which case observables does one need to hold fixed such that priority scores then only rely on the remaining components of x via indictment beliefs $h_j(r,x)$?

To allow the data to speak to this question, we observe that in cases with a continuous outcome space and beliefs H that are observed (up to classical measurement), Assumption ERM carries a testable implication. In particular, Assumption ERM implies a restriction on observable quantities through Eq. (7), which can be used to test the ERM assumption for a given choice of v.

To begin, we state a general version of this result which may be of use in observational rather than experimental settings. Then we specialize to the setting of our lab-experiment for a version of the result that is more easily implemented. In general, let W be any observable random variable/vector that takes on at least two values w, w' with positive probability, conditional on any values of J, V and $\tilde{R}$.

Proposition 2. *Suppose that ERM holds and W satisfies the following conditions:*

1. *There exist $w, w' \in supp\{W\}$ such that $w, w' \in supp\{W|S, J=j, V=v, \tilde{R}=r\}$ with probability one.*
2. $\mathbb{E}[U|\tilde{R}, J, X, W] = 0$

Then:

$$\begin{aligned}\mathbb{E}[H|\tilde{R}=r, J=j, V=v, W=w] \\ = \mathbb{E}\left\{\mathbb{E}[H|\tilde{R}=r, S, J=j, V=v, W=w']|\tilde{R}=r, J=j, V=v, W=w\right\}\end{aligned} \tag{9}$$

Note that for $w' = w$, Eq. (9) is trivially satisfied without ERM by the law of iterated expectations. Where ERM comes in is establishing the non-trivial Eq. (9) for $w' \neq w$. In typical non-experimental settings in which $R = \tilde{R}$ and only (J, R, X) is observed by the econometrician (possibly for only a subset of the X), finding a candidate W does require V to be a strict subset of X, so that there can be observable varation in W conditional on $V, \tilde{R}$ and J.

In the case of our lab experiment, we can let $W = R$, and the variation in W given $\tilde{R}$ comes from F, the indicator for a flipped case (e.g. $P(R = e|\tilde{R} = i) = \frac{P(F=1,R=e)}{P(F=0,R=i)+P(F=1,R=e)}$). Note that the second condition $\mathbb{E}[U|\tilde{R}, J, X, W] = 0$ of Proposition 2 is then equivalent to Assumption CLASSICAL. This leads to the following test for ERM in the lab experiment setting:

Corollary 1. *If ERM and CLASSICAL hold, then for any v and j:*

$$\mathbb{E}[H|R = e, J = j, V = v, F = 1] = \mathbb{E}\left\{ \mathbb{E}[H|R = i, S, J = j, V = v, F = 0] \,\middle|\, R = e, J = j, V = v, F = 1 \right\}$$

and

$$\mathbb{E}[H|R = i, J = j, V = v, F = 1] = \mathbb{E}\left\{ \mathbb{E}[H|R = e, S, J = j, V = v, F = 0] \,\middle|\, R = i, J = j, V = v, F = 1 \right\}$$

Combining, we have that

$$\mathbb{E}\left\{ H - \mathbb{E}[H|R = r', S, J, V, F = 0] \,\middle|\, R = r, J, V, F = 1 \right\} = 0$$

for each $r \in \{i, e\}$, where r' denotes the opposite value to r.

Note that without measurement error U in subjective indictment probabilities, Assumption ERM would imply that H is degenerate conditional on $\tilde{R}, S, J, V$ and thus with probability one $H = \mathbb{E}[H|R, S, J, V, F]$. Among cases with $F = 0$, this would imply that $H = \mathbb{E}[H|R, S, J, V, F = 0]$ (using RAND). However given that U is mean-zero, a version of this relationship still needs to hold between H and P *on average*.

4.4 Implementation of the ERM test

For a given choice of V, define

$$\tilde{g}_{jv}(r, s) := \mathbb{E}[H|R = r, S = s, J = j, V = v, F = 0]$$

The quantity $\tilde{g}_{jv}(r, s)$ denotes the expected subjective likelihood of indictment among (unflipped) cases of ethnicity r and $V = v$, for which analyst j assigns a priority score of s. Then Proposition 1 says that

$$\mathbb{E}[H - \tilde{g}_{jv}(i, S)|R = e, J = j, V = v, F = 1] = 0 \tag{10}$$

and similarly for Indigenous cases.[10]

We test the restriction (10) with $v = \emptyset$, by estimating the function $\tilde{g}_j(r, \cdot) = \mathbb{E}[H|R = r, S = \cdot, J = j, F = 0]$ separately by analyst j and race r, among the cases seen by each of the analysts in the lab experiment. There are 180 such analysts, each of which typically observes around 100 cases in the experiment. For ease of implementation, we assume that $\tilde{g}_j(R, S) := \mathbb{E}[H|R, j, S, F = 0]$ is linear in S and its square, with all coefficients allowed to vary by R. We perform inference via a nonparametric bootstrap of this procedure with resampling clustered by analyst. We call an analyst "ERM-friendly" if we fail to reject implication (10) at the 95% level for *both* Indigenous and European cases.

Overall, we reject implication (10) at the 90% level for at least one of $r \in \{i, e\}$ for 37 of the 180 analysts. This rejection of approximately 20% of analysts is in line with what we would expect if all analysts satisfied ERM (with $v = \emptyset$), given that we are performing two tests for each analyst with 10% size. However, we still have some reason to suspect that some analysts violate ERM. In Appendix A.1, we test an average of (10) across analysts j, for various choices of v among the observable case variables x, using analyst fixed effects. In this test we find significant violations of the averaged implication of ERM whether v is empty or includes indicators for case typology, geography, or both. This suggests that ERM violations do occur among the analysts in our sample, but these violations on average may be driven by relatively few such analysts. For this reason, we present our main results both for all analysts, and for the remaining 143 analysts that pass the analyst-specific test described above.

Appendix A.2 shows that in principle one can implement a systematic search for subsets v of the observables that are sufficient to make Assumption ERM hold. Since v does not seem to be required to pass the analyst-specific test for the majority of analysts, we do not implement this search.

5 Decomposing belief-based and taste-based bias

In this section, we leverage the unique features of our data to estimate the notions of TBB and BBB defined in Section 2 in our lab experiment setting. This is feasible under weak assumptions given i) the continuous action space for the priority score; ii) the elicitation of beliefs about indictment probability H, and iii) the experimental flipping of apparent race using the audit study,

[10]Note that if we considered unflipped, rather than flipped European cases, the analogous equation

$$\mathbb{E}[H|R = e, J = j, V = v, F = 0] = \mathbb{E}[\tilde{g}_{jv}(i, S)|R = e, J = j, V = v, F = 0]$$

would hold immediately by the law of iterated expectations, without requiring Assumption ERM.

5.1 Baseline case: accurate beliefs about indictment

To simplify exposition, suppose first that lab-analysts have perfect knowledge of indictment probabilities, so that $h_j(r, X) = h(r, x)$ where $h(r, x) := \mathbb{E}[Y|R = r, X = x]$.[11] This assumption is made in existing work that implements the marginal outcomes test. We will later relax this assumption and use our data H on beliefs directly.

Note that given Assumption ERM, we could then write

$$S = \phi_{J\tilde{R}V}(h(\tilde{R}, X)) \tag{11}$$

Then, defining $g_{jv}(r, s) := \mathbb{E}[Y|R = r, S = s, J = j, V = v, F = 0]$, we have by RAND and ERM that:

$$\begin{aligned} g_{jv}(r, s) &= \mathbb{E}[h(r, X)|R = r, J = j, S = s, V = v, F = 0] &(12)\\ &= \mathbb{E}[h(r, X)|R = r, \phi_{jrv}(h(r, X)) = s, V = v, J = j] = \phi^{-1}_{jrv}(s) &(13)\end{aligned}$$

where the first equality uses the law of iterated expectations over X and that S is deterministic conditional on J, R, X. Analagously to $\tilde{g}_{jv}(r, s)$, the function $g_{jv}(r, s)$ can be estimated using the sample of un-flipped cases, in which apparent ethnicity $\tilde{R}$ is equal to true ethnicity R. The population "predicted values" from regression (13) are $\hat{g} = g_{JV}(\tilde{R}, S)$, and we can similarly define $\hat{g}_e = g_{JV}(e, S)$ and $\hat{g}_i = g_{JV}(i, S)$, which apply fixed-ethnicity transformations to S.

We will find the following Lemma useful in detecting taste-based-discrimination, using $\tilde{g}_{jv}$ as an extension of the classic MOT from binary-action settings:

Lemma 1. *Under ERM and CLASSICAL, the function ϕ^{-1}_{jrv} is identified from the data by $\tilde{g}_{jv}(r, s) = \phi^{-1}_{jrv}(s)$*

To set the stage, consider the audit study variation with priority scores S as the outcome. The difference in means reveals a combination of taste-based and statistical disrimination. For example a simple difference in means among originally-European cases identifies, by (11) and RAND:

$$\begin{aligned}\mathbb{E}[S|R = e, F = 1] - \mathbb{E}[S|R = e, F = 0] &= \mathbb{E}[\phi_{JiV}(h(i, X))|R = e] - \mathbb{E}[\phi_{JeV}(h(e, X))|R = e]\\ &= \mathbb{E}[\underbrace{\phi_{JeV}(h(i, X)) - \phi_{JeV}(h(e, X))}_{\text{belief-based term}}|R = e] + \mathbb{E}[\underbrace{\phi_{JiV}(h(i, X)) - \phi_{JeV}(h(i, X))}_{\text{taste-based term}}|R = e]\end{aligned}$$

While the first term depends only on belief-based discrimination $h(e, x) \neq h(i, x)$, the second term depends only on taste-based discrimination $\phi_{jev}(\cdot) \neq \phi_{jiv}(\cdot)$ over various lab-analysts j. A similar decomposition can be applied to the originally-Indigenous cases.

[11] Note that $h(r, x)$ is also equal to $\mathbb{E}[Y|R = r, X = x, F = 0]$, by Assumption RAND ($h$ conditions on actual rather than apparent race).

Applying any common transformation to S in both arms of the experiment similarly measures total discrimination. In particular, the choice $g_{JV}(e,\cdot)$ yields:

$$\begin{aligned}
TOT_e &= \mathbb{E}[\hat{g}_e|R=e,F=1] - \mathbb{E}[\hat{g}_e|R=e,F=0] \\
&= \mathbb{E}[g_{JV}(e,S)|R=e,F=1] - \mathbb{E}[g_{JV}(e,S)|R=e,F=0] \\
&= \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JiV}(h(i,X)))|R=e] - \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JeV}(h(e,X)))|R=e] \\
&= \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JiV}(h(i,X))) - h(e,X)|R=e] \\
&= \mathbb{E}[\underbrace{h(i,X) - h(e,X)}_{\text{belief-based bias}}|R=e] + \mathbb{E}[\underbrace{\phi_{JeV}^{-1}(\phi_{JiV}(h(i,X))) - h(i,X)}_{\text{taste-based bias}}|R=e]
\end{aligned}$$

where we have used Eq. (13) and RAND. Compared with the un-transformed difference in means, $TOT_e := \mathbb{E}[\delta_{JeX}|R=e]$ expresses total discrimination in units of indictments, rather than the arbitrary scale of priority score question. This decomposition also holds for $TOT_i := \mathbb{E}[\delta_{JeX}|R=i]$, the analagous quantity among originally-Inidgenous cases. We focus on the measure δ_{jex} of belief-equivalent bias from Section 2 rather than the symmetric form TOT_{jx} solely for simplicity.

Average BBB among originally-European cases can furthermore be isolated by using *apparent-race* predicted values $\hat{g}$, rather than applying fixed-race functions $g_{JV}(e,\cdot)$ to priority scores:

$$\begin{aligned}
BBB_e &= \mathbb{E}[\hat{g}|R=e,F=1] - \mathbb{E}[\hat{g}|R=e,F=0] \\
&= \mathbb{E}[g_{JV}(i,S)|R=e,F=1] - \mathbb{E}[g_{JV}(e,S)|R=e,F=0] \\
&= \mathbb{E}[\phi_{JiV}^{-1}(\phi_{JiV}(h(i,X)))|R=e] - \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JeV}(h(e,X)))|R=e] \\
&= \mathbb{E}[\underbrace{h(i,X) - h(e,X)}_{\text{belief-based bias}}|R=e] \qquad (14)
\end{aligned}$$

TBB is thus identified by the difference between TOT_e and BBB_e:

$$\begin{aligned}
TBB_e &:= \mathbb{E}[\underbrace{\phi_{JeV}^{-1}(\phi_{JiV}(h(i,X))) - h(i,X)}_{\text{taste-based bias}}|R=e] = TOT_e - BBB_e \\
&= \{\mathbb{E}[g_{jV}(e,S)|R=e,J=j,F=1] - \cancel{\mathbb{E}[g_{jV}(e,S)|R=e,J=j,F=0]}\} \\
&\quad - \{\mathbb{E}[g_{jV}(i,S)|R=e,J=j,F=1] - \cancel{\mathbb{E}[g_{jV}(e,S)|R=e,J=j,F=0]}\} \\
&= \mathbb{E}[\hat{g}_i|R=e,J=j,F=1] - \mathbb{E}[\hat{g}_e|R=e,J=j,F=1] \qquad (15)
\end{aligned}$$

The same measures can be computed analogously for originally-Indigenous cases, and then the two can be averaged together to get overall measures TOT, BBB, and TBB.

To increase efficiency, regressions (14) and (15) are both implemented as fixed effects regressions on F with fixed effects for analyst J. That is, we estimate for each r:

$$TOT_r = \mathbb{E}[\hat{g}_r|R=r,F=1] - \mathbb{E}[\hat{g}_r|R=r,F=0] = \mathbb{E}\left\{\mathbb{E}[\hat{g}_r|R=r,J,F=1] - \mathbb{E}[\hat{g}_r|R=r,J,F=0]\right\}$$

from the rightmost expression (which can be obtained using that $J \perp\!\!\!\perp R \perp\!\!\!\perp F$ by RAND), and we assume that $\mathbb{E}[\hat{g}_r|R = r, J, F]$ is additively separable in F and J. Similarly we estimate BBB_r from

$$BBB_r = \mathbb{E}[\hat{g}|R = r, F = 1] - \mathbb{E}[\hat{g}|R = r, F = 0] = \mathbb{E}\left\{\mathbb{E}[\hat{g}|R = r, J, F = 1] - \mathbb{E}[\hat{g}|R = r, J, F = 0]\right\}$$

where again $\mathbb{E}[\hat{g}|R = r, J, F]$ is taken to be additively separable in F and J. We then compute TBB_r as $TOT_r - BBB_r$.

To implement estimation, we assume that $\tilde{g}_{JV}(R, S) := \mathbb{E}[H|R, J, S, V, F = 0]$ is linear in S and its square, V, and analyst-specific fixed effects, with all coefficients allowed to vary by R. This leads to a fixed-effects regression of the form $H = \alpha_{JR} + \beta_{0R}S + \beta_{1R}S^2 + \gamma_R'V + \epsilon$ where $\mathbb{E}[\epsilon|J, R, S, V, F = 0] = 0$. This equation is estimated separately by race.

5.2 General case with potentially inaccurate beliefs

More generally, now let $\tilde{h}_j(r, x) = h_j(r, x) - h(r, x)$ so that

$$S = \phi_{J\tilde{R}V}(h(\tilde{R}, X) + \tilde{h}_j(\tilde{R}, X))$$

instead of Eq. (11). If $\tilde{h}_j(r, x) \neq 0$ for some r and x, then analysts do not always have accurate beliefs about indictment probabilities.

5.2.1 What goes wrong with the decomposition using Y

In this setting, we can derive the following generalization of (13):

$$\begin{aligned}
g_{jv}(r, s) &= \mathbb{E}[h(r, X)|R = r, J = j, S = s, V = v, F = 0] \\
&= \mathbb{E}[h(r, X)|R = r, \phi_{jrv}(h(r, X) + \tilde{h}_j(r, X)) = s, V = v, J = j] \\
&= \mathbb{E}[h(r, X)|R = r, h(r, X) = \phi_{jrv}^{-1}(s) - \tilde{h}_j(r, X), V = v, J = j] \\
&= \phi_{jrv}^{-1}(s) - \mathbb{E}[\tilde{h}_j(r, X)|R = r, S = s, V = v, J = j]
\end{aligned}$$

and thus $g_{jv}(e, s) - g_{jv}(i, s)$ does not simply reflect $\phi_{jev}^{-1}(s) - \phi_{jiv}^{-1}(s)$, but also a second term that depends on the relative over-estimation of indictment probabilities among cases of each race. In particular, the MOT at a given point s (but averaged over lab-analysts j and observables v) yields:

$$\begin{aligned}
&\mathbb{E}[g_{JV}(e, s) - g_{JV}(i, s)|S = s] = \mathbb{E}[\phi_{JeV}^{-1}(s) - \phi_{JiV}^{-1}(s)|S = s] \\
&\quad + \mathbb{E}\left\{\mathbb{E}[\tilde{h}_J(i, X)|R = i, S = s, V, J]|S = s\right\} - \mathbb{E}\left\{\mathbb{E}[\tilde{h}_J(e, X)|R = e, S = s, V, J]|S = s\right\}
\end{aligned}$$

If the distribution of scores among originally European and originally Indigenous cases are not too different, then the final two terms above evaluate to $\mathbb{E}[\tilde{h}_J(i, X)|R = i, S = s] - \mathbb{E}[\tilde{h}_J(e, X)|R = e, S = s]$. If lab-analysts tend to over-estimate the probability of

indictment more for European cases than they do for Indigenous cases, for example, this second term will be negative which may misleadingly suggest taste-based discrimination against European defendants even if taste-based bias exists against Indigenous defendants. In general, the MOT with Y on the left-hand side is no longer valid when $h_j(r,x)$ differs from $h(r,x) = \mathbb{E}[Y|R=r, X=x]$.

5.2.2 Recovering the decomposition using H instead of Y

We can still implement the MOT (and decomposition into BBB and TBB) when analysts may have inaccurate beliefs about indictment probabilities, by using the data on H (which was not used in the exercise of Section 5.1). In fact, the fix is simple: replace the $g_{jv}(r,\cdot)$, which are condition expectations of Y, with the $\tilde{g}_{jv}(r,\cdot)$, which are conditional expectations of H.[12]

Average belief-based discrimination can be defined in the general setting as:

$$\begin{aligned} BBB_e &:= \mathbb{E}[h_J(i,X) - h_J(e,X)|R=e] = \mathbb{E}[H-U|R=e,F=1] - \mathbb{E}[H-U|R=i,F=0] \\ &= \mathbb{E}[H|R=e,F=1] - \mathbb{E}[H|R=i,F=0] \end{aligned} \tag{16}$$

where the final equality uses CLASSICAL and RAND. THis shows that BBE_e can be identified quite simply as a difference in means for originally-European cases and using H as the outcome. For contiguity with the estimates of Section 5.1 that assume accurate indictment beliefs, we base estimation on rewriting Eq. (16) as:

$$\begin{aligned} BBB_e &= \mathbb{E}\left\{\mathbb{E}[H|R=e,J,S,V,F=1]|R=e,F=1\right\} \\ &\qquad - \mathbb{E}\left\{\mathbb{E}[H|R=i,J,S,V,F=0]|R=e,F=0\right\} \\ &= \mathbb{E}[\tilde{g}_{JV}(i,S)|R=e,F=1] - \mathbb{E}[\tilde{g}_{JV}(e,S)|R=e,F=0] \end{aligned}$$

using the Eq. (10) form of Proposition 1, which shows that BBB_e can be estimated by simply replacing g in Eq (14) with $\tilde{g}$.

When replacing g in Eq (15) with $\tilde{g}$, the taste-based discrimination estimand becomes:

$$\begin{aligned} TBB_e &= \mathbb{E}[\tilde{g}_{jV}(e,S)|R=e,J=j,F=1] - \mathbb{E}[\tilde{g}_{jV}(i,S)|R=e,J=j,F=1] \\ &= \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JiV}(\tilde{h}_J(i,X))) - \tilde{h}_J(i,X)|R=e] \end{aligned} \tag{17}$$

using Lemma 1. This more general measure of TBB_e again only differs from zero due to taste-based discrimination $\phi_{jev}(\cdot) \neq \phi_{jiv}(\cdot)$ over various lab-analysts j. An analogous decomposition holds for Indigenous cases. As in the case for estimates based on accurate beliefs from Section 5.1, estimates of BBB_r and TOT_r (and hence TBB_r) are implemented as fixed effects regressions on F with fixed effects for analyst J.

[12]Recall from Section 4.4 that $\tilde{g}_{jv}(r,s) := \mathbb{E}[H|R=r,S=s,J=j,V=v,F=0]$ is the expected subjective likelihood of indictment among (unflipped) cases of ethnicity r and $V=v$, for which analyst j assigns a priority score of s.

5.2.3 Separating accurate and incorrect beliefs

As we saw in Equation (6) of Section 3, belief-based bias can further be separated into accurate statistical discrimination and a term due to differentially incorrect beliefs

$$\begin{aligned} BBB_e &= \mathbb{E}[h_J(i,X) - h_J(e,X)|R=e] \\ &= \underbrace{\mathbb{E}[h(i,X) - h(e,X)|R=e]}_{\text{accurate stat. disc.}:=ASD_e} + \underbrace{\mathbb{E}[\tilde{h}_J(i,X) - \tilde{h}_J(e,X)|R=e]}_{\text{incorrect belief term}:=BBT_e} \end{aligned} \tag{18}$$

where $\tilde{h}_j(r,x) := h_J(r,x) - h(r,x)$, with $h(r,x) = \mathbb{E}[Y|R=r, X=x]$. An analogous decomposition to (18) holds for Indigenous cases.

The inaccuracy in $\tilde{h}_j(r,x)$ in analysts' *beliefs* can be estimated provided that the function $h(r,x)$ is known. Note that even given the function h, this measurement for a given case-analyst pair is only up to the measurement error term U, since we only observe $H = h_J(\tilde{R}, X) + U$ and not $h_J(\tilde{R}, X)$ directly.

In principle, we can estimate $h(r,x)$ given that we observe all of the variables x that are shown to analysts in the lab-experiment, along with the actual indictment status Y from the associated historical case. However, the estimation problem is high-dimensional, given the richness of the x. We rely on a random forest (RF) approach to provide estimates $\hat{h}(r,x)$ of the function $h(r,x)$. We begin with roughly $2,000$ potential explanatory variables after standardizing the universe of historical complaints from 2014 to 2021. These include summary statistics of the denounced entity's record (previous complaints, previous successful complaints etc), transformations of the text of the complaint, and summaries of this text (e.g. length, readability, word TF-IDF).

To limit the number of variables fed into the RF algorithm, we perform the following steps. First, we use a clustering algorithm to group variables into 70 clusters. Then, we select one variable within each cluster, and run RF with the resulting 70 variables, recording each variable's explanatory power. This procedure is repeated many times with different draws of the 70 explanatory variables (one from each cluster). We then perform a final variable selection by choosing the variable from each cluster with the highest average explanatory power across those draws. RF predictions from this final set of variables define $\hat{h} := \hat{h}(R,X)$ for each case in the historical data.

6 Results

6.1 Binary Marginal Outcome Test on the historical data

Table 1 shows the results from implementing the historical MOT outlined in Appendix A. We estimate, separately by race, two-stage-least squares regressions of indictment ($Y=1$)

	(1)	(2)	(3)	(4)
Panel A: European (N=9,296)				
Escalation	0.2481***	0.2321***	0.2474***	0.2302***
	(0.0182)	(0.0214)	(0.0182)	(0.0215)
Adjusted R^2	0.2677	0.2885	0.2678	0.291
Panel B: Indigenous (N=3,589)				
Escalation	0.1582***	0.1613***	0.1595***	0.1650***
	(0.0349)	(0.0349)	(0.0349)	(0.0349)
Adjusted R^2	0.1928	0.2145	0.1947	0.2274
Panel C: Difference (N=12,885)				
Escalation	0.0899**	0.0708*	0.0879*	0.0652
	(0.0391)	(0.0409)	(0.0391)	(0.0410)
Adjusted R^2	0.2539	0.2716	0.2543	0.2736
Fixed Effects	-	Region	Month	Region×Month

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1: Results of the historical MOT. Positive values in Panel C represent discrimination against Indigenous defendants.

on escalation ($D = 1$) using the leniency[13] of the matched analyst as an instrument:

$$Y = \alpha_2 + \beta_2 \widehat{D} + \mathbf{X}_i \boldsymbol{\gamma}_2 + \varepsilon$$
$$D = \alpha_1 + \beta_1 \text{Leniency} + \mathbf{X}_i \boldsymbol{\gamma}_1 + \nu$$

The columns of 1 indicate various choices of fixed effects to represent strata within which historical analysts are assumed to be as-good-as-randomly assigned to cases. While controlling for region by month fixed effects (column (4)) represents the weakest assumption in terms of identification, these rich fixed effects increase standard errors enough to render the difference in Panel C insignificant at the 10% level. Given the positive magnitude of the difference for less rich sets of fixed effects, we interpret Table 1 as providing suggestive but relatively weak evidence of taste based bias against Indigenous defendents in the historical escalation decisions at CGR.

[13]We construct a scalar measure Z_i representing the assigned analyst's leniency (defined as propensity for escalation of report i), taking ordered values $z_0, \ldots, z_J$, where $J+1$ is the total number of analysts in the CGR system. For instance, a value of $j = 0.5$ indicates that analyst j escalates 50% of all reports. We construct Z_i using a standard leave-out procedure capturing analysts' leniency separately by ethnicity.

6.2 Triage Experiment

Table 2 reports our results from the experiment outlined in Appendix C. We estimate the impact of experimentally manipulating defendants' surnames on analysts' assigned scores with report-level and analyst-level controls:

$$\begin{aligned}\text{score}_i = \alpha + \beta_1 \text{Switch Euro to Ind}_i + \beta_2 \text{Switch Ind to Euro}_i + \beta_3 \text{Original Ind}_i \\ + \mathbf{X_i} + \mathbf{X_j} + \varepsilon_i\end{aligned}$$

where score_i is the score assigned to report i, $\text{Switch Euro to Ind}_i$ and $\text{Switch Ind to Euro}_i$ indicate whether the surname was experimentally switched from European to Indigenous or from Indigenous to European, respectively, and Original Ind_i indicates the defendant's original surname was Indigenous. $\mathbf{X_i}$ and $\mathbf{X_j}$ are vectors of report and analyst controls.[14]

Our results show that on average, analysts who see a complaint presented as accusing an Indigenous defendant assign it a higher indictment probability score than those who see it as a European defendant.

	(1)	(2)	(3)	(4)
Switching surname from European to Indigenous	1.544***	1.255***	1.063**	0.835**
	(0.484)	(0.416)	(0.461)	(0.390)
Switching surname from Indigenous to European	-1.223	-0.346	-0.282	0.584
	(0.907)	(0.807)	(0.855)	(0.742)
Original Indigenous surname	-4.225***	-4.647***		
	(0.574)	(0.503)		
Analyst-level controls	Yes	No	Yes	No
Complaint-level controls	Yes	Yes	No	No
Analysts FEs	No	Yes	No	Yes
Complaints FEs	No	No	Yes	Yes
Clusters (analyst-level SEs)	177	182	177	182
Adjusted R-squared	0.029	0.427	0.130	0.530
Observations	18,706	19,112	18,706	19,112

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Results of lab-in-the-field triage experiment. Our dependent variable is Indictment Probability Score, with mean=63.01 and S.D.=28.96.

[14] We can modify our specification to include report and analyst fixed effects:

$$\text{Score}_i = \alpha + \beta_1 \text{Switch_Euro_to_Ind}_i + \beta_2 \text{Switch_Ind_to_Euro}_i + \phi_i + \phi_j + \varepsilon_i$$

where ϕ_i are report fixed effects and ϕ_j are analyst fixed effects. This approach controls for all time-invariant characteristics at both the report and analyst levels.

6.3 Results of the decomposition into BBB and TBB

Table 3 reports for belief based and taste-based bias, based on the identification results established in Section 5. Column (4) represents our prefered estimates, which use analysts stated beliefs H rather than observed indictments Y, and restrict the analysis to "ERM-friendly" analysts that pass the overidentification restriction described in Section 4.3. The estimates suggest that on average, analysts exhibit taste-based bias against Indigenous defendants along with belief-based bias in favor of Indigenous defendants, leading to an overall total bias weakly against Indigenous defendants. These patterns hold whether averaging over originally-European or originally-Indigenous cases. Table 4 shows a similar pattern holds among Indigenous and European analysts.

	Y (indictment)	Y (indictment)	H (belief)	H (belief)
Belief-based bias (Euro. cases)	2.191**	2.197**	-1.542**	-1.530*
	(0.720)	(0.816)	(0.494)	(0.597)
Taste-based bias (Euro. cases)	-2.038**	-2.122**	2.253***	1.908***
	(0.710)	(0.821)	(0.355)	(0.382)
Total bias (Euro. cases)	0.153*	0.0745	0.711*	0.378
	(0.0757)	(0.107)	(0.350)	(0.429)
Belief-based bias (Indig. cases)	2.265***	2.410**	-1.549*	-0.957
	(0.663)	(0.820)	(0.644)	(0.764)
Taste-based bias (Indig. cases)	-2.131**	-2.234**	2.035***	1.732***
	(0.653)	(0.785)	(0.332)	(0.363)
Total bias (Indig. cases)	0.134	0.176	0.486	0.775
	(0.136)	(0.171)	(0.613)	(0.714)
Restricted to ERM-friendly analysts		X		X
Number analysts	180	143	180	143
Number of obs.	19089	15028	19089	15028

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Results of the bias decomposition: first with actual indictments and then with perceived indictment probabilities.

	H (Euro. analysts)	H (Indig. analysts)
Belief-based bias	-1.546**	-0.949
	(0.564)	(1.082)
Taste-based bias	1.808***	2.101**
	(0.417)	(0.717)
Total bias	0.262	1.153
	(0.393)	(0.748)
Restricted to ERM-friendly analysts	X	X
Number analysts	111	32
Number of obs.	11569	3459

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Decomposition results by race of historical analyst. . Relative to Table 3, this table is simplified by only considering estimates based on stated beliefs H, only estimating on the restricted sample of ERM-friendly analysts, and averaging bias across originally-European and originally-Indigenous cases.

6.4 Results on inaccuracy in analysts' beliefs

Table 5 shows estimates of $\mathbb{E}[H - \hat{h}(R, X)|R = r, J \in \mathcal{J}]$ for given races $r \in \{i, e\}$ of the original historical case, and sets $\mathcal{J}$ of analysts (Indigenous analysts, European analysts, and all analysts). Given Assumption CLASSICAL, and assuming consistency of the random forest predictions, $\mathbb{E}_n[H - \hat{h}(R, X)|R = r, J \in \mathcal{J}]$ is a consistent estimator for $\mathbb{E}[\tilde{h}_j(r, x)|R = i, J \in \mathcal{J}] := \mathbb{E}[h_J(r, x) - h(r, x)|R = i, J \in \mathcal{J}]$, where $\mathbb{E}_n$ is the sample expectation function.

In particular, for a given choice of $\mathcal{J}$, the second row of Table 5 shows our estimate of $\mathbb{E}[\tilde{h}_j(r, x)|R = i, J \in \mathcal{J}]$ (average inaccuracy in beliefs among Indigenous cases). The second row displays estimates of $\mathbb{E}[\tilde{h}_j(r, x)|R = e, J \in \mathcal{J}] - \mathbb{E}[\tilde{h}_j(r, x)|R = i, J \in \mathcal{J}]$. The results indicate that analysts *overestimate* indictment probabilities (by 0.4 or 40 points for Indigenous cases, and by roughly 0.5 or 50 points for European cases). The gap of roughtly 0.1 (or 10 points on a 100 point scale) is strongly statistically in all three columns. This suggests that the term labeled "differentially-incorrect indictment beliefs"

in Eq. (6) is *negative*, meaning that belief-based bias against Indigenous accused entities $h_j(i,x) - h_j(e,x)$ tends to be lower than accurate statistical discrimination $h(i,x) - h(e,x)$.

Incorrectly assuming that analysts' beliefs are accurate and using Y in the decomposition of total bias into BBB and TBB—as columns (1)-(2) of Table 3 do—attributes this gap to BBB, yielding overestimates of BBB and underestimates of TBB. Columns (3)-(4) of Table 3 instead use observed beliefs H, and accordingly find a lower value of BBB and a higher value of TBB, in fact flipping the signs of each. See Section 5.2.1 for further discussion of how the positive gap reported in Table 5 between $\tilde{h}_j(e,x)$ and $\tilde{h}_j(i,x)$ biases estimates of TBB using Y away from finding bias against Indigenous defendents.

	(1)	(2)	(3)
	All analysts	European analysts	Indigenous analysts
Difference Euro.-Indig. cases	10.11***	10.23***	9.733***
	(0.786)	(0.904)	(1.583)
Indigenous cases	41.83***	40.87***	44.83***
	(0.718)	(0.827)	(1.447)
N	14506	10971	3535

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Estimates of the average bias in beliefs about the indictment probability for cases of a given ethnicity, measured out of 100 points. See text above for description.

7 Conclusion

In this paper we have provided a comprehensive framework within which to study biases in how citizens are treated by the state. The framework incorporates the three main sources of bias discussed in the economics literature: Biases based on tastes, on statistical discrimination, and on inaccurate beliefs (Arnold et al., 2018; Arnold et al., 2022; Bohren et al., 2019; Lang and Lehmann, 2012; Hurst et al., 2024; Bohren et al., 2025). It also incorporates citizens' sequential encounters with state actors and the interlinkages between the biases at each stage. We operationalize the framework using a triage experiment conducted with Peruvian anti-corruption officials, finding that assuming analysts to have accurate beliefs about indictment suggests belief-based bias against Indigenous defendants along with taste-based bias against Europeans, while using elicited beliefs suggests the opposite: belief-based bias against European cases along with taste-based bias against Indigenous defendants, leading to a small total bias against Indigenous defendants.

We reconcile these findings by estimating "true" indictment probabilities from a ran-

dom forest prediction of indictment given the rich covariates we observe about each case, defined from the full information displayed to analysts in the lab experiment. Our estimates suggest that analysts systematically overestimate the probability of indictment for cases, but they do so more for European defendants than for Indigenous defendants. This implies that belief-based bias against Indigenous cases is smaller than the "accurate statistical discrimination" that analysts would employ if they had accurate beliefs. Incorrectly assuming accurate beliefs attributes this differential to belief based bias, leading to misleading estimates for both belief-based and taste-based bias. This underscores the value of observing beliefs when implementing marginal outcomes tests.

References

AIGNER, D. J. and CAIN, G. G. (1977). "Statistical Theories of Discrimination in Labor Markets". *ILR Review* 30, pp. 175–187.

ALESINA, A. and LA FERRARA, E. (2014). "A test of racial bias in capital sentencing". *American Economic Review* 104 (11), pp. 3397–3433.

ANWAR, S., BAYER, P. and HJALMARSSON, R. (2012). "The impact of jury race in criminal trials". *The Quarterly Journal of Economics* 127 (2), pp. 1017–1055.

ARGYLE, B., INDARTE, S., IVERSON, B. and PALMER, C. (2024). *Racial Disparities and Bias in Consumer Bankruptcy*. URL: `https://www.nber.org/papers/w33575`.

ARNOLD, D., DOBBIE, W. and HULL, P. (2022). "Measuring Racial Discrimination in Bail Decisions". *American Economic Review* 112 (9), pp. 2992–3038.

ARNOLD, D., DOBBIE, W. and YANG, C. S. (May 2018). "Racial Bias in Bail Decisions*". *The Quarterly Journal of Economics* 133 (4), pp. 1885–1932. URL: `https://doi.org/10.1093/qje/qjy012`.

ARROW, K. J. (1973). "The Theory of Discrimination". *Discrimination in Labor Markets*. Ed. by O. ASHENFELTER and A. REES. Princeton University Press.

ARTILES, M. (2024). "Ethnic Diversity, Historical Economic Exchange, and Development: Evidence from Andean Peru".

BECKER, G. S. (1971). *The economics of discrimination*. University of Chicago press.

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

BOHREN, J. A., HAGGAG, K., IMAS, A. and POPE, D. G. (2025). "Inaccurate Statistical Discrimination: An Identification Problem". *The Review of Economics and Statistics*, pp. 1–16.

BOHREN, J. A., HULL, P. and IMAS, A. (2024). *Systemic Discrimination: Theory and Measurement*. URL: https://www.nber.org/papers/w29820.

BOHREN, J. A., IMAS, A. and ROSENBERG, M. (2019). "The Dynamics of Discrimination: Theory and Evidence". *American Economic Review* 109 (10), pp. 3395–3436.

CANAY, I. A., MOGSTAD, M. and MOUNTJOY, J. (Aug. 2023). "On the Use of Outcome Tests for Detecting Bias in Decision Making". *The Review of Economic Studies* 91 (4), pp. 2135–2167. URL: https://doi.org/10.1093/restud/rdad082.

EMBRECHTS, P. and HOFERT, M. (2013). "A note on generalized inverses". *Mathematical Methods of Operations Research* 77, pp. 423 – 432. URL: https://api.semanticscholar.org/CorpusID:13152899.

FRANDSEN, B., LEFGREN, L. and LESLIE, E. (2023). "Judging Judge Fixed Effects". *American Economic Review* 113 (1), 253–77. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20201860.

GALARZA, F. B. and YAMADA, G. (2014). "Labor Market Discrimination in Lima, Peru: Evidence from a Field Experiment". *World Development* 58, pp. 83–94.

HODERLEIN, S., HOLZMANN, H., KASY, M. and MEISTER, A. (2017). "Corrigendum: Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment". *The Review of Economic Studies* 84 (2 (299)), pp. 964–968. URL: http://www.jstor.org/stable/45106746 (visited on 01/25/2025).

HULL, P. (2021). *What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making*. Working Paper 28503. National Bureau of Economic Research. URL: http://www.nber.org/papers/w28503.

HURST, E., RUBINSTEIN, Y. and SHIMIZU, K. (2024). "Task-Based Discrimination". *American Economic Review* 114 (6), pp. 1723–1768.

LANG, K. and LEHMANN, J.-Y. K. (2012). "Racial Discrimination in the Labor Market: Theory and Empirics". *Journal of Economic Literature* 50 (4), pp. 959–1006.

MILGROM, P. and SHANNON, C. (1994). "Monotone Comparative Statics". *Econometrica* 62 (1), pp. 157–180. URL: http://www.jstor.org/stable/2951479 (visited on 03/05/2025).

MINISTERIO DE CULTURA (2018). *I Encuesta Nacional Percepciones y actitudes sobre diversidad cultural y discriminación étnico-racial. Principales resultados*.

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

ROSS, K. A. (2003). *Elementary Analysis: The Theory of Calculus*. 2nd. New York: Springer.

SIGSTAD, H. (2023). "Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs". *Mimeo:*

VYTLACIL, E. (2002). "Independence, Monotonicity, and Latent Index Models: An Equivalence Result". *Econometrica* 70 (1), pp. 331–341. URL: http://www.jstor.org/stable/2692171 (visited on 03/20/2025).

Appendices

A Econometric framework for the historical MOT

Suppose that $\phi_{jrv}(h) = \mathbb{1}(h \geq t_{jrv})$ and $h_j(r,x) = h^*(r,x)$ for some common function h^* (i.e. beliefs are common, whether or not they are correct). Suppose that conditional on any value R, $h^*(R,X)$ is continuously distributed with a positive density such that

$$D = \mathbb{1}(h^*(R,X) \geq t_{JRV}) = \mathbb{1}(U \geq G_{RV}(t_{JRV}))$$

where $G_{RV}(t) := F_{h^*(R,X)|R,V}(t)$ and $U = G_{RV}(h^*(R,X))$. Notice that $U|R,V = Unif[0,1]$ by assumption and thus $U|J,R,V = Unif[0,1]$ since $J \perp\!\!\!\perp (R,X)$ by quasi-random assignment of analysts. For ease of exposition, we implicitly condition on the strata within which this assumption is maintained. The above decision rule is equivalent to the LATE monotonicity assumption holding for cases of a given race (Vytlacil, 2002).

Consequently, we can identify the quantity $G_{rv}(t_{jrv})$ by noticing that

$$P(D = 0|J = j, R = r, V = v) = P(U < G_{rv}(t_{jrv})) = G_{rv}(t_{jrv})$$

Define $p_{RV}(j) = P(D = 0|J = j, R, V) = G_{RV}(t_{jRV})$, where $p_{rv}(j)$ is the leniency of judge j among cases of race r and $V = v$. Consider now

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[Y|p_{RV}(J) = t, R, V] &= \frac{d}{dt}\mathbb{E}[Y^* \cdot D|p(J) = t, R, V] \\
&= \frac{d}{dt}\mathbb{E}[Y^* \cdot \mathbb{1}(U \geq G_{RV}(t_{JRV}))|G_{RV}(t_{JRV}) = t, R, V] \\
&= \frac{d}{dt}\mathbb{E}[Y^* \cdot \mathbb{1}(U \geq t)|R, V] \\
&= \frac{d}{dt}\int_t^1 \mathbb{E}[Y^*|U = u, R, V] \cdot du = -\mathbb{E}[Y^*|U = t, R, V] \\
&- \mathbb{E}[Y^*|U = t, R, V] = -\mathbb{E}[Y^*|h^*(R,V) = G_{RV}^{-1}(t), R, V] \\
&= -\mathbb{E}[Y^*|h^*(R,V) = G_{RV}^{-1}(t), R, V, p_{RV}(J) = t] \\
&= -\mathbb{E}[Y^*|h^*(R,V) = G_{RV}^{-1}(p_{RV}(J)), R, V, p_{RV}(J) = t] \\
&= -\mathbb{E}[Y^*|h^*(R,V) = t_{JRV}, R, V, p_{RV}(J) = t]
\end{aligned}$$

Suppose that $h^*(R,X) = \mathbb{E}[Y^*|R,X]$. Then in turn, by the law of iterated expectations:

$$\begin{aligned}
\mathbb{E}[Y^*|h^*(R,X) = t_{JRV}, R, V, p_{RV}(J) = t] &= \mathbb{E}\left\{\mathbb{E}[Y^*|R,X,J]|h^*(R,X) = t_{JRV}, R, V, p_{RV}(J) = t\right\} \\
&= \mathbb{E}\left\{\mathbb{E}[Y^*|R,X]|h^*(R,X) = t_{JRV}, R, V, p_{RV}(J) = t\right\} \\
&= \mathbb{E}\left\{h^*(R,X)|h^*(R,X) = t_{JRV}, R, V, p_{RV}(J) = t\right\} \\
&= \mathbb{E}\left\{t_{JRV}|h^*(R,X) = t_{JRV}, G_{RV}(t_{JRV}) = t, R, V\right\} \\
&= \mathbb{E}\left\{t_{JRV}|h^*(R,X) = t_{JRV} = G_{RV}^{-1}(t), R, V\right\} \\
&= \mathbb{E}[t_{JRV}|R,V]
\end{aligned}$$

Consider the case with no V for simplicity. Then the difference in MTE's:

$$\frac{d}{dt}\mathbb{E}[Y|p_i(J)=t,R=i]-\frac{d}{dt}\mathbb{E}[Y|p_e(J)=t,R=e]=\mathbb{E}[t_{Je}|R=e]-\mathbb{E}[t_{Ji}|R=i]=\mathbb{E}[t_{Je}-t_{Ji}]$$

We follow Arnold et al. (2018) in assuming that the LHS of the above expression is well-approximated by a two-stage least squares regression of Y on escalation D, instrumented by leniency $p_r(j)$ within race, and including strata fixed effects.

A.0.1 Identifying bias in accusations

The distribution of observables X over the sample of accused cases of a given race r is

$$f_{X|A=1,R=r}(x)=f_{X|R=r}(x)\cdot\frac{P(A=1|R=r,X=x)}{P(A=1|R=r)}$$

Consequently, we can identify $P(A=1|R=i,X=x)-P(A=1|R=e,X=x)/P(A=1)$ as

$$\frac{f_{X|A=1,R=i}(x)}{f_{X|R=i}(x)}\cdot\frac{P(R=i|A=1)}{P(R=i)}-\frac{f_{X|A=1,R=e}(x)}{f_{X|R=e}(x)}\cdot\frac{P(R=e|A=1)}{P(R=e)} \tag{19}$$

All of the quantities in (19) are identified. Although we do not know the unconditional probability of accusation $P(A=1)$, it is positive and hence we can infer the sign of $P(A=1|R=i,X=x)-P(A=1|R=e,X=x)$, the difference in accusation-rates among observably-equivalent Indigenous and European cases.

A.1 A pooled test for ERM violations averaged across analysts

Note that by the law of iterated expectations over analysts J and over case observables V, Corollary 1 implies that $\mathbb{E}[H|R=e,F=1]=\mathbb{E}[\tilde{g}_{JV}(i,S)|R=e,F=1]$ and $\mathbb{E}[H|R=i,F=1]=\mathbb{E}[\tilde{g}_{JV}(e,S)|R=i,F=1]$. Thus in implementing an "averaged" test of ERM, we only need to form regression predictions $\tilde{g}_{JV}(r,S)=\mathbb{E}[H|R=r,J,S,V,F=0]$ for each r and test that $H-\tilde{g}_{JV}(r,S)$ is mean zero among the sample of flipped cases for the opposite ethnicity. For ease of implementation, we assume that $\tilde{g}_{JV}(R,S):=\mathbb{E}[H|R,J,S,V,F=0]$ is linear in S and its square, V, and analyst-specific fixed effects, with all coefficients allowed to vary by R. This leads to a fixed-effects regression of the form $H=\alpha_{JR}+\beta_{0R}S+\beta_{1R}S^2+\gamma_R'V+\epsilon$ where $\mathbb{E}[\epsilon|J,R,S,V,F=0]=0$. This equation is estimated separately by race, and predicted values formed across all observations. Results are presented in Table 6.

A.2 A systematic search over v satisfying ERM

Let $\mathcal{X}$ denote the set of non-race variables visible to the analyst, and let $\mathcal{V}^*\subseteq\mathcal{X}$ be the "true" subset of these variables that enter into analysts decisions of S.

	(1)	(2)	(3)	(4)
Moment value (European cases)	0.0278***	0.0284***	0.0260***	0.0277***
	(0.00384)	(0.00394)	(0.00679)	(0.00702)
Moment value (Indigenous cases)	-0.0184***	-0.0167***	-0.0172*	-0.0175*
	(0.00418)	(0.00364)	(0.00676)	(0.00818)
Typology controls		X		X
Department controls			X	X
Number of obs.	19089	19089	19089	19089

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Tests of Assumption ERM via the overidentification restriction in Proposition 1. Columns correspond to various choices of V: either indicators for case typology or for geographical indicators (Peruvian *departamentos*). Under ERM for a given choice of V, the value of the moment should be zero for each case race.

We've seen above that ERM would hold trivially if $\mathcal{V} = \mathcal{X}$, the full set of case observables. Note that if ERM holds with $\mathcal{V}$, then it also holds for any larger set of variables $\mathcal{V} \subset \mathcal{V}'$, because the function $\phi_{jrv'}$ need not depend on the components of v' that are not in v. Thus, the assumption that ERM holds with $\mathcal{V} = \emptyset$ (analagous to the assumption of Canay et al. (2023) but for a continuous action) is the strongest version of ERM possible. For concreteness, we assume that Assumption ERM holds for no strict subsets of $\mathcal{V}^*$, i.e. it is a *minimal* subset of $\mathcal{X}$ that makes Assumption ERM hold.

Suppose that for any $\mathcal{V} \subset \mathcal{X}$: $\tilde{g}_{jv}(r,s) = \alpha_{jr}^{\mathcal{V}} + \beta_{0r}^{\mathcal{V}} s + \beta_{1r}^{\mathcal{V}} s^2 + \gamma_r^{\mathcal{V}\prime} v$, where note that all coefficients can depend on the set $\mathcal{V}$. Let $g_{jv}^*(r,s)$ be the function $\tilde{g}_{jv}(r,s)$ when $\mathcal{V} = \mathcal{V}^*$, and denote its coefficients as $g_{jv}^*(r,s) = \alpha_{jr}^* + \beta_{0r}^* s + \beta_{1r}^* s^2 + \gamma_r^{*\prime} v$

Lemma 2. *$\lambda_j = 0$ for any $v_j \notin \mathcal{V}^*$.*

Proof. Let v^* be the part of v that corresponds to $\mathcal{V}^*$

$$\begin{aligned}
\tilde{g}_{jv}(r,s) &= \mathbb{E}[H|R=r, J=j, S=s, V=v, F=0] \\
&= \mathbb{E}[h_j(r,X) + U|R=r, J=j, S=s, V=v, F=0] \\
&= \mathbb{E}[h_j(r,X)|R=r, J=j, s_j(r,X)=s, V=v, F=0] \\
&= \mathbb{E}[h_j(r,X)|R=r, J=j, s_j(r,X)=s, V^*=v^*, V^c=v^c] \\
&= \mathbb{E}[h_j(r,X)|R=r, J=j, h_j(r,X)=\phi_{jrv^*}^{-1}(s), V^*=v^*, V^c=v^c] = \phi_{jrv^*}^{-1}
\end{aligned}$$

which does not vary with v^c. □

B Testing the monotonicity assumption of the historical MOT

Let D_{ij} denote whether analyst j would escalate case i. ("Unconditional") monotonicity says that for any two analysts j', j one of the analysts (j) is more "lenient" than the other in the sense that

$$D_{ij'} \geq D_{ij} \quad \text{for all cases } i$$

That is: if the more lenient analyst j would escalate the case, then so would the less-lenient analyst j' (i.e. $D_{ij} = 1 \implies D_{ij'} = 1$). Given this assumption and random assignment of analysts to cases, we can detect empirically which analyst from a pair is the more lenient, since the ordering of j' and j above is the same as the ordering of analysts based on their observed leniency $z_j = P(D_i = 0|J_i = j)$, where J_i indicates the analyst assigned to case i. To see this, use that $D_i = D_{i,J_i}$ and thus $z_j = P(D_{ij} = 0)$ and $z_j - z_{j'} = P(D_{ij} = 0 \text{ but } D_{ij'} = 1) \geq 0$, so $z_j \geq z_{j'}$.

Within-ethnicity monotonicity instead assumes that

$$D_{ij'} \geq D_{ij} \quad \text{for all cases in which } R_i = r$$

Under this assumption we instead define an analysts' leniency only using cases of a given race, i.e. $z_{jr} = P(D_i = 0|J_i = j, R_i = r)$ where R_i the ethnicity of the accused in case i. The ordering of analysts by their leniencies may differ by race. Then we say that analyst j is more lenient than analyst z' for cases of type r if $z_{jr} \geq z_{j'r}$, which follows from the above given random assignment of cases to analysts $J_i \perp\!\!\!\perp (R_i, X_i) \implies \{D_{ji} \perp\!\!\!\perp J_i\}|R_i$.

B.1 Testing for monotonicity

Following Frandsen et al. (2023), we conduct a nonparametric test for the monotonicity assumption required for our instrumental variable (IV) design based on quasi-random assignment of analysts to cases. Test rejection indicates that either the exclusion restriction, the monotonicity assumption, or both, may not be upheld. Table 7 shows the results of implementing the test. For Indigenous defendants, we fail to reject (p-value $\sim$1) at any confidence level, while for European defendants we fail to reject with 95% confidence but we reject with 90% confidence (p-value = 0.054).

B.2 Direct enumeration of monotonicity violations

Similar to Sigstad (2023), we can also test monotonicity more directly making use of the fact that in our experiment, multiple analysts see overlapping pairs of cases. We use the priority scores P_{ij} assigned to case i by analyst j. We assume that these represent an

	European	Indigenous
Test statistic	671	158
d.f.	(602)	(274)
P-value	[0.054]	[1.000]

Table 7: Implementing the Frandsen et al. (2023) nonparametric test for monotonicity

"ordinal proxy" for the counterfactual escalation decisions D_{ij} they would make, in the sense that

$$S_{ij'} > S_{ij} \implies D_{ij'} \geq D_{ij} \tag{20}$$

Let analysts j' and j see sets $C_j(r)$ and $C_{j'}(r)$ cases of ethnicity r, with intersection $I_{jj'}(r) = C_{jr} \cap C_{j'}(r)$. We measure violations of monotonicity separately for Indigenous and European cases (and restricting attention to cases that have not been "flipped".

For any two cases $i \neq i'$ in $I_{jj'r}$, we say that there exists a monotonicity violation if $S_{ij'} > S_{ij}$ but $S_{i'j'} < S_{i'j}$. The overall number of such violations is

$$n(r) = \sum_{\substack{j,j' \in \mathcal{J} \\ j' \neq j}} \sum_{\substack{i,i' \in I_{jj'}(r) \\ i' \neq i}} \mathbb{1}(S_{ij'} > S_{ij} \text{ and } S_{i'j'} < S_{i'j})$$

where $\mathcal{J}$ is the set of all analysts. The number of such comparisons is

$$d(r) = \sum_{\substack{j,j' \in \mathcal{J} \\ j' \neq j}} \sum_{\substack{i,i' \in I_{jj'}(r) \\ i' \neq i}} 1$$

and thus the overall rate of monotonicity violations among cases of race r can be defined as $n(r)/d(r)$. Given (20), this serves as an upper bound on the rate of monotoncity violations with respect to escalation D_{ij}, because the strict inequalities in S only imply weak inequalities in D. We find this rate to be 11.8% among Indigenous cases and 11.6% among European cases.

C Experimental Design

During July and August 2021, 182 analysts from CGR (approximately 90% of the triaging staff) evaluated 196 whistleblower reports previously assessed by CGR. Analysts assigned scores from 0 to 100, indicating the priority and likelihood of indictment, with 100 representing the highest priority. The experiment was conducted virtually following CGR Covid-19 safety protocols. Figure A.1 shows the experiment's interface.

Sample Selection. Reports were selected based on specific criteria: (i) text length between 50 and 350 words; (ii) issued between 2018 and 2020; (iii) allegations involving

illicit enrichment, favoritism, or deficient public works execution; (iv) defendants identified as municipal mayors; and (v) defendant's full name explicitly mentioned. The final sample aimed to reflect Peru's ethnic composition.

Participants. Analysts were predominantly male, aged 30-40, university-educated, primarily trained in Law or Engineering, and employed at various regional CGR offices. Participation was mandated by CGR management, with analysts informed their evaluations would assist in designing a prioritization algorithm in collaboration with the Inter-American Development Bank and Columbia University. Analysts were unaware that surnames had been altered in some reports.

Randomization Procedure. Reports were randomly assigned to analysts ensuring participants would not receive cases from their own regional offices, to prevent familiarity with the cases. To maintain balance across key characteristics, randomization was stratified by defendant ethnicity (Indigenous or European) and outcomes (dismissed at escalation, escalated but dismissed at indictment, or indicted).

Surname Randomization. To experimentally manipulate implied defendant ethnicity, surnames were randomly changed between Indigenous (Quechua and Aymara) and European (Spanish) origins based on regional surname dictionaries derived from historical CGR data and Artiles, 2024. Approximately 24% of evaluations involved altered surnames, ensuring realistic and regionally appropriate replacements.

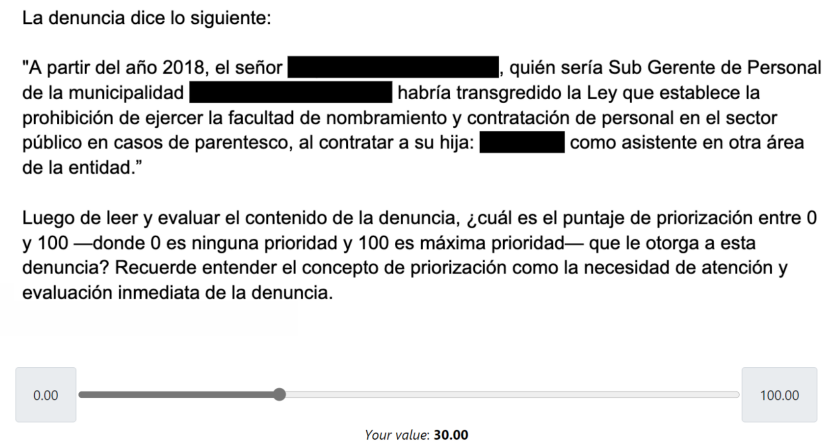


Figure A.1: Experiment Interface

D Further proofs

D.1 Proof of Proposition 2

Let $\tilde{g}_{jvw}(r,S) := \mathbb{E}[H|\tilde{R}=r,S,J=j,V=v,W=w]$. Then

$$\begin{aligned}
&\mathbb{E}[\tilde{g}_{jvw'}(r,S)|\tilde{R}=r,J=j,V=v,W=w] = \mathbb{E}[g_{jvw'}(r,s_j(r,X))|\tilde{R}=r,J=j,V=v,W=w] \\
&\quad= \int dF_{X|rjvw}(x)\cdot\tilde{g}_{jvw'}(r,s_j(r,x)) \\
&\quad= \int dF_{X|rjvw}(x)\cdot\mathbb{E}[h_j(r,X)+U|\tilde{R}=r,J=j,s_j(r,X)=s_j(r,x),V=v,W=w'] \\
&\quad= \int dF_{X|rjw}(x)\cdot h_j(r,x) + \int dF_{X|rjvw}(x)\cdot\mathbb{E}[U|\tilde{R}=r,J=j,s_j(r,X)=s_j(r,x),V=v,W=w'] \\
&\quad= \mathbb{E}[h_j(r,X)|R=e,J=j,V=v,W=w]+0 \\
&\quad= \mathbb{E}[H|\tilde{R}=r,J=j,V=v,W=w]
\end{aligned}$$

Direct proof in the lab experiment setting

Letting $\tilde{g}_{jv}(r,s) := \mathbb{E}[H|R=r,S=s,J=j,V=v,F=0]$:

$$\begin{aligned}
&\mathbb{E}[\tilde{g}_{jv}(i,S)|R=e,J=j,V=v,F=1] = \mathbb{E}[\tilde{g}_{jv}(i,s_j(i,X))|R=e,J=j,V=v,F=1] \\
&\quad= \int dF_{X|ejv1}(x)\cdot\tilde{g}_{jv}(i,s_j(i,x)) \\
&\quad= \int dF_{X|ejv1}(x)\cdot\mathbb{E}[h_j(i,X)+U|R=i,J=j,s_j(i,X)=s_j(i,x),V=v,F=0] \\
&\quad= \int dF_{X|ej1}(x)\cdot h_j(i,x) + \int dF_{X|ejv1}(x)\cdot\mathbb{E}[U|R=i,J=j,s_j(i,X)=s_j(i,x),V=v,F=0] \\
&\quad= \mathbb{E}[h_j(i,X)|R=e,J=j,V=v,F=1]+0 = \mathbb{E}[H|R=e,J=j,V=v,F=1]
\end{aligned}$$

where we use (7) in the fourth step to replace $\mathbb{E}[h_j(i,X)|R=i,J=j,s_j(i,X)=s_j(i,x),V=v,F=0]$ with $h_j(i,x)$, because (7) and that X and x share a v means that with probability one $s_j(i,X)=s_j(i,x)$ implies that $h_j(i,X)=h_j(i,x)$. The proof for Indigenous cases is identical.

D.2 Proof of Lemma 1

$$\begin{aligned}
\tilde{g}_{jv}(r,s) &= \mathbb{E}[h_J(\tilde{R},X)+U|R=r,\phi_{J\tilde{R}V}(h_J(\tilde{R},X))=s,J=j,V=v,F=0] \\
&= \mathbb{E}[h_j(r,X)|R=r,h_j(r,X)=\phi^{-1}_{jrv}(s),J=j,V=v,F=0] \\
&\qquad + \mathbb{E}[U|R=r,h_j(r,X)=\phi^{-1}_{jrv}(s),J=j,V=v,F=0] \\
&= \phi^{-1}_{jrv}(s)+\mathbb{E}[U|R=r,J=j,F=0] = \phi^{-1}_{jrv}(s)+0
\end{aligned}$$

Technical note: Given that the event $\phi_{J\tilde{R}V}(h_J(\tilde{R},X))=s$ has probability zero, some care is required in replacing the conditional expectation $\mathbb{E}[h_j(r,X)|R=r,\phi_{jrv}(h_j(r,X))=$

$s, J = j, V = v, F = 0]$ with $\mathbb{E}[h_j(r, X)|R = r, h_j(r, X) = \phi_{jrv}^{-1}(s), J = j, V = v, F = 0]$. The reason is that although $\phi_{jrv}(h_j(r, X)) = s$ occurs if an only if $h_j(r, X) = \phi_{jrv}^{-1}(s)$, conditional distributions that condition on $\phi_{jrv}(h_j(r, X)) = s$ are defined from a shrinking sequence of intervals $[s, s + \epsilon]$ as $\epsilon \downarrow 0$, while conditional distributions that condition on $h_j(r, X) = \phi_{jrv}^{-1}(s)$ are defined from a shrinking sequence of intervals $[\phi_{jrv}^{-1}(s), \phi_{jrv}^{-1}(s) + \epsilon]$ as $\epsilon \downarrow 0$. If instead of holding $V = v$ fixed, we held $X = x$ fixed, the two conditional expectations would not need to be the same for multivariate V (HODERLEIN et al., 2017).

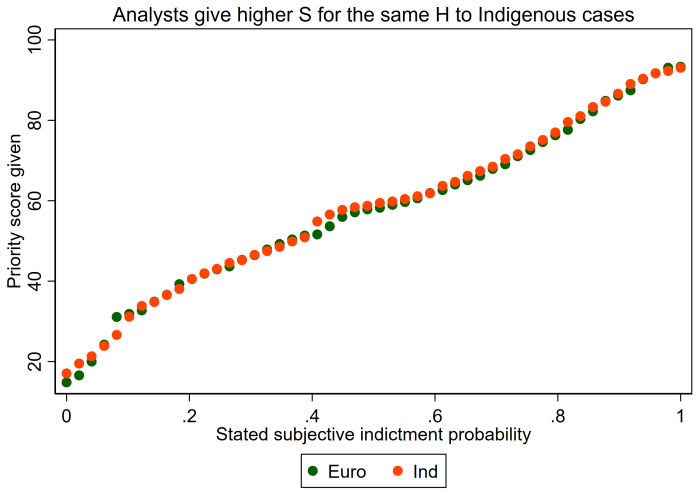
D.3 Direct proof of Corollary 1 using Lemma 1

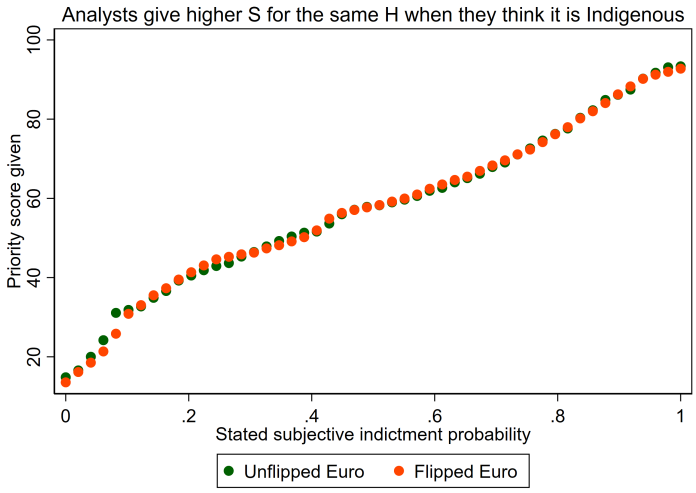
Without going through Proposition 2, a more intuitive proof of Corollary 1 can be obtained via Lemma 1. Given Lemma 1, we can prove Corollary 1 by noting then that

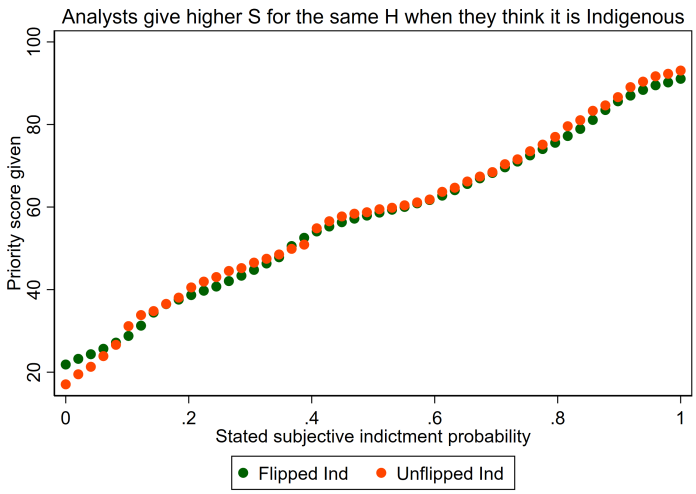
$$\begin{aligned}
\mathbb{E}[\tilde{g}_{jv}(i, S)|R = e, J = j, V = v, F = 1] &= \mathbb{E}[\phi_{jiv}^{-1}(S)|R = e, J = j, V = v, F = 1] \\
&= \mathbb{E}[\phi_{jiv}^{-1}(\phi_{J\tilde{R}V}(h_J(\tilde{R}, X)))|R = e, J = j, V = v, F = 1] \\
&= \mathbb{E}[\phi_{jiv}^{-1}(\phi_{jiv}(h_j(i, X)))|R = e, J = j, V = v, F = 1] \\
&= \mathbb{E}[h_j(i, X)|R = e, J = j, V = v, F = 1] \\
&= \mathbb{E}[h_J(\tilde{R}, X)|R = e, J = j, V = v, F = 1] \\
&= \mathbb{E}[H - U|R = e, J = j, V = v, F = 1] = \mathbb{E}[H|R = e, J = j, V = v, F = 1]
\end{aligned}$$

E Additional tables and figures

E.1 Prima-facie evidence for taste-based bias


Analysts give higher S for the same H to Indigenous cases
Priority score given
20
40
60
80
100
0
.2
.4
.6
.8
1
Stated subjective indictment probability
Euro
Ind


Analysts give higher S for the same H when they think it is Indigenous
Priority score given
20
40
60
80
100
0
.2
.4
.6
.8
1
Stated subjective indictment probability
Unflipped Euro
Flipped Euro


Analysts give higher S for the same H when they think it is Indigenous
Priority score given
20
40
60
80
100
0
.2
.4
.6
.8
1
Stated subjective indictment probability
Flipped Ind
Unflipped Ind