An Observational Implementation of the Outcome Test with an Application to Ethnic Prejudice in Pretrial Detentions*

Nicolás Grau[†]   Damián Vergara[‡]

First version: January 27, 2020

This version: July 13, 2025

*This paper previously circulated under the title "A Simple Test for Prejudice in Selection Processes: The Prediction-Based Outcome Test." Grau worked on this project during his tenure as a professor in the Department of Economics at the Universidad de Chile before joining the Chilean Government as Minister. The views expressed here, therefore, are those of the authors and do not represent those of the Ministry of Economy, Development, and Tourism, or the Chilean government. We thank the Chilean Public Defender Office (Defensoría Penal Pública), the Mapuche Data Project (http://mapuchedataproject.cl/), and the Director of the Department of Studies of the Supreme Court (Centro de Estudios de la Corte de Suprema), for providing the data. We thank the editor, Andrea Weber, two anonymous referees, and Diego Amador, Hadar Avivi, Aureo De Paula, Marina Dias, Will Dobbie, Joseph Doyle, Peter Hull, Michael Jansson, Felipe Jordán, Patrick Kennedy, Patrick Kline, Attila Lindner, Charles Manski, Pablo Muñoz, Demian Pouzo, James Powell, Jesse Rothstein, Yotam Shem-Tov, Petra Todd, Chris Walters, Harry Wheeler, John Wieselthier, and two anonymous referees, as well as seminar participants at the 11th Annual Conference of the Spanish Association of Law and Economics, the Bolivian Conference on Development Economics, the Chicago/LSE Conference on the Economics of Crime and Justice (2021), CReAM/UCL, the Econometric Society World Congress (2020), the FROGEE Workshop on Gender Economics, King's College, the LACEA/LAMES Annual Meetings (2021), the Online Economics of Crime Seminar (Summer 2020), PUC-Chile (Centro de Justicia y Sociedad), the RIDGE Workshop on the Economics of Crime (2021), Toulouse School of Economics, UC Berkeley, Universidad de Chile, Universidad Mayor, University of Oxford, and Wilfrid Laurier University, for very helpful discussions and suggestions. We also thank Tomás Cortés for excellent research assistance. Nicolás Grau thanks the Centre for Social Conflict and Cohesion Studies (ANID/FONDAP/15130009) for financial support.

[†]Ministry of Economy, Development, and Tourism, Government of Chile, ngrauv@economia.cl.

[‡]Department of Economics, University of Michigan, damiv@umich.edu.

Abstract

We propose an observational implementation of the outcome test that uses the propensity score for selecting samples of treated individuals that are more likely to be close to the margin of treatment, thus attenuating concerns about inframarginality bias. Our approach requires neither instruments nor the random assignment of decision-makers, allows for unrestricted correlation between observables and unobservables, and can accommodate non-monotone patterns of prejudice. We illustrate our method by analyzing prejudice in pretrial detentions against the Mapuche, the largest ethnic minority group in Chile, and find strong evidence of prejudice against them.

1 Introduction

Many selection processes rely on predicted outcomes. For example, bail judges are generally mandated to determine the pretrial detention status of defendants based on the likelihood of pretrial misconduct if released. Researchers and policymakers may be interested in assessing whether selection processes are prejudiced – that is, whether decision-makers routinely establish different selection thresholds for members of a particular group due to animus or systematic mispredictions. A prominent approach for testing for prejudice in selection processes is the outcome test (Becker, 1957, 1993) which is based on the idea that, if bail judges are not prejudiced, marginally released defendants from different groups should exhibit equal pretrial misconduct rates. Thus, testing for prejudice is reduced to comparing the average outcome of marginally selected individuals between groups, which amounts to a simple difference in means.

The outcome test, however, presents a key empirical challenge: the identification of marginally selected individuals. If potential outcome distributions differ between groups, differences in outcomes away from the margin may lead to misleading conclusions regarding prejudice. The literature has proposed various approaches to address this identification problem, known as the *inframarginality bias*. Quasi-experimental solutions usually rely on random assignment of decision-makers, which is unlikely to hold in many settings. On the other hand, observational proposals rely on structural assumptions that may be viewed as overly restrictive for most practical applications. Therefore, a more flexible observational approach for situations in which instruments are unavailable or unlikely to work properly is currently absent from the literature.

This paper proposes a novel observational implementation of the outcome test, the Prediction-Based Outcome Test (P-BOT), that uses the predicted selection status (i.e., the propensity score) to identify samples of treated individuals that are more likely to be close to the margin of treatment, thus attenuating concerns about inframarginality bias. We motivate our approach with a model where judges decide over defendants' pretrial release status based on expected pretrial misconduct.

Our formal notion of prejudice is based on aggregate differences in effective selection thresholds across groups and accounts for judges' preferences and biased beliefs, as in Arnold et al. (2018) and Hull (2021). We formally show that the outcome test is valid under our definition of prejudice.

Our main theoretical contribution is to provide conditions under which the released individuals that are more likely to be close to the margin of release also have lower propensity scores. Under these conditions, the challenge of identifying marginal individuals is reduced to a prediction problem, simplifying the implementation of the outcome test: the econometrician has to estimate the propensity score, rank released individuals according to their predicted values, define samples of individuals that are close to the margin of release, compute group-specific pretrial misconduct rates within these samples of individuals, and perform a difference in means.

Relying on the propensity score to predict which released individuals are closer to the margin of release implies that the P-BOT is robust to the presence of correlation between observables and unobservables. Intuitively, bias in the projection coefficients is not problematic because the econometrician needs to know *who* is close to the margin, not *why*. Although the projection error is non-systematic, this error can still generate inframarginality bias because local rank reversals can bias the outcome test if misclassified individuals have different outcome distributions. Hence, the nature of this projection error has two important consequences. First, the potential for inframarginality bias in the P-BOT is proportional to the conditional variance of the projection error since it governs the scope for rank reversals. This implies that the performance of the P-BOT depends on the availability of good predictors. Second, the conditional variance of the unobservables can be estimated with additional assumptions to build an inframarginality bias test.

Our identification argument is based on three assumptions. First, we assume that the outcome distribution is smooth in the latent risk when approaching the margin of release. Under this assumption, the behavior of released individuals that are *close to the margin* is a good approximation of the behavior of marginal individuals, which is the intuitive basis for our identification argument. Second, we assume that the selection equation is additively separable between observables and unobservables. Through the lens of the model, this induces monotonicity on observables in the risk probabilities. Third, while we allow for unrestricted first moments in the joint distribution of observables and unobservables, which is an important improvement relative to the existing observational literature, we impose restrictions on the conditional variance. When the conditional variance is homoskedastic, we show that released individuals with lower propensity scores are closer to the margin of release up to an uncorrelated mean-zero projection noise which can generate the bias discussed above. We also characterize patterns of heteroskedastic conditional variance that deliver the same conclusions.

We view the contribution of our paper as primarily *practical*, as it increases the applicability of outcome tests relative to other approaches while minimizing costs in terms of bias. While some degree of inframarginality bias can be expected from the P-BOT application, we highlight that our empirical test does not rely on random assignment of decision-makers and can accommodate non-monotone patterns of prejudice, appearing as an attractive alternative in situations where preferred methods such as the instrument-based approach of Arnold et al. (2018) are not implementable. Moreover, our test offers avenues to empirically assess the pervasiveness of the inframarginality bias and, as we argue throughout the paper, the concern for bias is substantially attenuated relative to standard observational practices in the literature, such as the use of average outcomes.

As an application of the P-BOT, we test for prejudice against the largest ethnic minority group in Chile, the Mapuche, using nationwide administrative data. Around 10% of the Chilean population identifies as Mapuche. The Mapuche population provides an interesting case study for three reasons. First, a long-running conflict exists between the Mapuche and the Chilean state, dating back more than a century (Cayul et al., 2022). In this context, it is frequently claimed that Chilean institutions are biased against the Mapuche. Second, the Mapuche people are subject to numerous negative stereotypes, such as tendencies towards laziness, violence, and alcoholism, from some quarters of Chilean society (Merino and Quilaqueo, 2003; Merino and Mellor, 2009). Third, Mapuche people are identifiable, mainly because of their surnames but also to some extent due to their physical appearance. Thus, prejudice against members of this group is feasible in this setting.

We use nationwide administrative data that covers more than 95% of criminal cases in Chile between 2008 and 2017. Bail decisions in Chile are effectively binary because monetary bail is not an option. The data contains detailed information on cases and defendants, including judges' and attorneys' identifiers. We merge the administrative records with a register of Mapuche surnames to create measures of ethnicity that combine self-reporting and surname information.

We implement the P-BOT by fitting different projection models for the release status. The results provide evidence of prejudice against Mapuche defendants. Our preferred specification shows that marginal Mapuche defendants are between 3 and 4 percentage points less likely to be engaged in pretrial misconduct relative to marginal non-Mapuche defendants. We also provide evidence of a modest potential for inframarginality bias in our setting. Therefore, the outcome test using the full sample (Knowles et al., 2001) also suggests prejudice against Mapuche defendants, although the implied magnitude is smaller. Since the Chilean setting is characterized by quasi-random assignment of bail judges at the court-by-time level, we also test for prejudice using the instrument-based approach proposed by Arnold et al. (2018). While the LATE for the non-Mapuche sample of defendants is precisely estimated, we show that the estimation is severely underpowered for the Mapuche sample. This prevents us from drawing conclusions from its application. The fragility of the IV estimation in our setting illustrates that the P-BOT is an attractive alternative when the instrument-based approach cannot be properly implemented. Encouragingly, the LATE for the non-minority sample is similar to the P-BOT estimates of non-Mapuche pretrial misconduct rates at the margin, and the non-minority marginal defendants identified by the two methods have similar distributions of observables. This suggests that both approaches can give similar results when both are expected to work properly.

We conclude with extensions that further illustrate the practical appeal of the P-BOT. First, we explore more complex patterns of prejudice by including additional regressors in the outcome equation. We present two examples. In the first, we group defendants by income level, conjecturing that the prejudice patterns may interact with socioeconomic status. In the second, considering the geographical component of the Mapuche conflict, we test for geographical heterogeneities in the prejudice patterns. Our results show that prejudice patterns are stronger for Mapuche defendants that live in low-income municipalities and that, while there is prejudice against Mapuche defendants in all Chilean courts, it is modestly stronger in the region that has been historically associated with the Mapuche conflict. These results suggest that non-monotone patterns of discrimination are likely to occur in practice. We also estimate the outcome equations controlling by court-by-time fixed effects (the level at which judges are randomly assigned) and find that around half of the overall prejudice is explained by the assignment rule of judges to defendants (i.e., by Mapuche defendants being systematically assigned to courts with stricter judges). This mediator relates to the notion of *institutional discrimination* formalized in Bohren et al. (2025b).

This paper contributes to the literature on discrimination by proposing a simple method to test for prejudice (Guryan and Charles, 2013; Lang and Kahn-Lang, 2020; Small and Pager, 2020). More specifically, it adds to the literature that discusses the properties and the implementation of the outcome test (Knowles et al., 2001; Anwar and Fang, 2006; Simoiu et al., 2017; Arnold et al., 2018; Gelbach, 2021; Hull, 2021; Feigenberg and Miller, 2022; Marx, 2022; Canay et al., 2024). Throughout the paper, we argue that our approach is appealing in settings where instrument-based approaches are weak or infeasible, thus constituting a complement to the existing literature. Our empirical application also adds to a vast body of evidence on racial bias in the criminal justice system (e.g., Antonovics and Knight, 2009; Abrams et al., 2012; Anwar et al., 2012; Rehavi and Starr, 2014; Anwar et al., 2018; Cohen and Yang, 2019; Rose, 2021; Arnold et al., 2022). Related to our paper, Arnold et al. (2018) find that bail judges are prejudiced against black defendants. Understanding racial disparities in pretrial detentions matters beyond the normative concerns they raise because pretrial detention affects conviction rates, employment, and the use of state benefits (Leslie and Pope, 2017; Dobbie et al., 2018; Dobbie and Yang, 2021a,b; Grau et al., 2021).

The rest of the paper is organized as follows. Section 2 describes and discusses our definition of prejudice and the outcome test. Section 3 introduces our approach, the P-BOT, and discusses identification and estimation. Section 4 describes the institutional setting and the data used in our empirical application. Section 5 presents the results. Finally, Section 6 concludes.

2 Preliminaries: Prejudice and the Outcome Test

This section describes and discusses our definition of prejudice, and the outcome test and its empirical challenges. We formally show that the outcome test identifies our definition of prejudice.

2.1 Prejudice

We analyze prejudice in selection rules that are based on expected outcomes. To fix ideas, consider a situation where judges decide whether or not to grant pretrial release for a defendant. Each judge is mandated to predict the likelihood that the defendant will be engaged in pretrial misconduct (non-appearance in court or pretrial recidivism) if released during the investigation, compare that to a threshold, and make a decision. The legal mandate requires judges to release defendants unless the expected risk of pretrial misconduct is *large*. The question we address is whether there is prejudice against a specific group (e.g., minority defendants) in the release decision.

A non-prejudiced judge makes unbiased predictions of the probability of pretrial misconduct and releases defendants whenever that probability is smaller than a leniency threshold. From the perspective of an individual judge, however, prejudice can arise because of two (non-exclusive) sources. Because of preferences, the judge can set different thresholds to different groups of defendants, a practice commonly known as *taste-based discrimination*. Judges can also engage in *inaccurate statistical discrimination* or stereotypes, for example, by systematically overestimating misconduct risk for minority defendants. In both cases, a prejudiced judge sets higher effective thresholds for minority defendants, as a smaller *true* pretrial misconduct probability is required for release. The definition of prejudice we use in this paper is the composite effect of preferences and stereotypes. As common in the literature, the framework we develop is not able to separately identify both sources of prejudice (Arnold et al., 2018; Hull, 2021; Bohren et al., 2025a).[1]

[1] The – statistically accurate – use of group identity for computing pretrial misconduct probabilities is usually referred to as *accurate statistical discrimination*, which is also a normatively relevant source of discrimination (Kleinberg et al., 2019; Kline and Walters, 2021; Yang and Dobbie, 2020; Arnold et al., 2022). While being robust to its

In general terms, the stylized selection rule for an individual defendant can be formalized by:

$$R_i \;=\; 1\{m(G_i, Z_i, j(i)) \leq h(G_i, Z_i, j(i))\}, \tag{1}$$

where i indexes defendants and j judges, $j(i)$ is a function that assigns judges to defendants, R_i takes value 1 if the defendant i is pretrial released, G_i is a group indicator variable (e.g., minority), Z_i is a vector of characteristics of defendant i observed by the judge (e.g., type of crime and criminal record), $m(G_i, Z_i, j(i)) = \mathbb{E}[PM_i|G_i, Z_i, j(i)]$ is the *true* conditional probability of pretrial misconduct if released of defendant i, with $PM_i \in \{0,1\}$, and $h(G_i, Z_i, j(i))$ is the effective threshold that can vary with G_i and Z_i because of preferences or stereotypes (or both), and is potentially heterogeneous across judges.[2] Let the latent release status be given by $R_i^* = h(G_i, Z_i, j(i)) - m(G_i, Z_i, j(i))$, hence $R_i = 1\{R_i^* \geq 0\}$. We say that a released defendant is marginal if $R_i^* = 0$.

Following (1), we focus on an aggregate notion of prejudice. We compare the average effective threshold, $h(G_i, Z_i, j(i))$, between groups $G_i \in \{0,1\}$, across all judges and non-group characteristics, of individuals that are marginal. Formally, defining $\overline{h}(g) = \mathbb{E}[h(G_i, Z_i, j(i))|G_i = g, R_i^* = 0]$ as the average effective threshold faced by marginal defendants with $G_i = g$ motivates the following (contrapositive) definition of prejudice:

DEFINITION 1 (PREJUDICE). *In the absence of prejudice,* $\overline{h}(0) = \overline{h}(1)$.

Under Definition 1, the decision process is prejudiced against defendants of group 1 whenever $\overline{h}(0) > \overline{h}(1)$. Definition 1 relates to Arnold et al. (2018) and Hull (2021) notion of bias based on the *disparate impact* perspective, but differs from Canay et al. (2024) that say a judge j is unbiased if $h(0, j(i), Z_i) = h(1, j(i), Z_i)$, for all Z_i, so non-group characteristics are equalized when defining group-based prejudiced decision-making. The conditioning of $\overline{h}(g)$ on $R_i^* = 0$ is required in the absence of random assignment to make the effective threshold a policy-relevant object, as it is the

presence, the analysis we develop is not informative of accurate statistical discrimination. Also, our definition of prejudice is conditional on true pretrial misconduct probabilities. Then, sources of *systemic discrimination* that affect the signal generating process are also absent from the analysis below (see Bohren et al., 2025b for a discussion).

[2] In some contexts R_i may be non-binary (see Arnold et al., 2022 for a discussion). In our empirical application below the bail decision is effectively binary since the Chilean system does not consider monetary bail.

object that can be identified by the outcome test without further assumptions. Also, defendants at the margin are those for whom bias is likely to change the treatment status.

There are several reasons why differences in average effective thresholds, $\overline{h}(g)$, may not necessarily reflect the intuition discussed above for a single judge. In what follows, we discuss these reasons and the normative implications of these alternative interpretations.

The role of Z_i. Average effective thresholds integrate over non-group characteristics. If different groups have different distributions of Z_i, differences in average thresholds could be recovering prejudice based on other characteristics. Under a disparate impact notion of prejudice (Arnold et al., 2022), Definition 1 remains normatively relevant because group disparities in release rates may emerge for reasons unrelated to the probabilities of pretrial misconduct. Under disparate treatment notions of prejudice, however, this distinction may compromise the validity of the outcome test (Canay et al., 2024). One advantage of the approach we introduce in the next section is that it allows for testing patterns of prejudice that simultaneously depend on G_i and Z_i.

Assignment of judges to defendants. While the approach introduced in the next section does not require random assignment of judges for identification, the nature of the assignment rule $j(i)$ matters for interpreting differences in effective thresholds. To see why, consider two cases. First, all judges are unprejudiced but stricter judges are systematically assigned to minority defendants. Second, all judges are prejudiced against minority defendants, but there is random assignment of judges. In both cases, Definition 1 is violated, but with different interpretations. The second case aligns with the intuition of the single judge, while the first case reflects a situation where $j(i)$ can be said to be prejudiced, evoking the notion of *institutional discrimination* defined in Bohren et al. (2025b). Both cases remain normatively relevant under disparate impact notions of prejudice. Below we show how our approach can be used to decompose both sources of prejudice when judges are quasi-randomly assigned at some intermediate level (e.g., court-by-time).

Alternative objective functions. The analysis assumes that judges *should* make decisions based on predicted pretrial misconduct. However, judges may have different objective functions, echoing the definition of omitted-payoff bias of Kleinberg et al. (2018). In this case, there may be differences in effective thresholds that remain normatively relevant under our disparate impact notion of prejudice because some defendants are discriminated against with respect to the normative standard provided by law. This would not be true under an institutional setting that mandates a different criterion for the use of pretrial detention (e.g., Manski, 2005, 2006). Then, if the mandated selection rule is well defined, individual deviations do not affect the normative relevance of our definition of prejudice. Below we discuss how the approach presented in the next section can be used to assess whether judges care about predicted risk when making the release decisions.

2.2 Outcome test

From Definition 1, testing for prejudice in the release decision is reduced to comparing the average effective thresholds between groups. While this defines an intuitive null hypothesis to be rejected, its application is challenging since effective thresholds are rarely observable.

One approach used to overcome this challenge is the *outcome test* (Becker, 1957, 1993). If there is prejudice in the selection process, for example, against minority defendants, observed pretrial misconduct rates of marginally released minority defendants should be smaller than those observed for non-minority defendants. That is, testing for prejudice is reduced to a difference in means: the econometrician needs to find a statistically significant correlation between pretrial misconduct and group for the defendants at the margin. The next proposition, similar to results presented in Hull (2021), establishes that the observed behavior of marginal individuals of a given group coincides with the average effective threshold.

PROPOSITION 1. *Let PM_i be the observed pretrial misconduct of defendant i. Then*:

$$\mathbb{E}[PM_i|G_i = g, R_i^* = 0] \quad = \quad \overline{h}(g). \tag{2}$$

Proof. See Appendix A.

Putting together Definition 1 and Proposition 1 formalizes the outcome test.

COROLLARY (OUTCOME TEST). *In the absence of prejudice*:

$$\mathbb{E}[PM_i|G_i = 0, R_i^* = 0] \quad = \quad \mathbb{E}[PM_i|G_i = 1, R_i^* = 0]. \tag{3}$$

Identification of marginal individuals. While the outcome test implementation does not require observing effective thresholds, it induces an additional empirical challenge. The difference in means described above can be trivially implemented when knowing which released defendants are marginal. However, R_i^* is usually not observed by the econometrician. This is important because the misspecification of marginal individuals may induce bias in the outcome test: when the risk distributions differ between groups, differences in pretrial misconduct rates computed away from the margin may not be informative about effective thresholds and, therefore, may result in misleading conclusions regarding prejudice. This is called the *inframarginality bias*.[3]

A solution that avoids imposing strong assumptions on judge behavior and the distribution of unobservables is proposed by Arnold et al. (2018). If the econometrician has an instrument for the release status, pretrial misconduct rates at the margin can be recovered by the expected treatment effects at the margin of release. Then, the outcome test can be implemented by comparing group-specific LATEs. Indeed, Hull (2021) shows that the outcome test is equivalent to the difference between group-specific MTE frontiers. By leveraging quasi-random assignment of bail judges, the authors propose to use judge-specific leave-out mean release rates as an instrument (the *judges design*). One problem with this approach is that it is equivalent to running a first-stage on judge fixed effects. This may induce power problems in settings where minority groups represent small shares of the defendants' population. Also, as emphasized by Muller-Smith (2015) and Frandsen et al. (2023), the leave-out mean release rate may fail to meet the LATE monotonicity assumption.[4]

There are many cases, however, where decision-makers are not quasi-randomly assigned and

[3] See Simoiu et al. (2017) and Arnold et al. (2018) for intuitive explanations of the inframarginality bias.

[4] See Arnold et al. (2022) and Chan et al. (2022) for methods that allow for deviations from strict monotonicity.

alternative instruments are unavailable, or when power problems or non-monotone judge behaviors make the judges design infeasible. These situations call for observational approaches to deal with the inframarginality bias. An example is provided by Chandra and Staiger (2010), who derive a test for prejudice that relies on selection-on-observables assumptions. Another example is Knowles et al. (2001). In the context of motor vehicle searches for contraband, the authors model equilibrium conditions under which the marginally searched individuals demonstrate the same behavior as the average ones. Then, linear regressions of the outcome equation using the full sample of selected individuals are enough to test for prejudice. However, Anwar and Fang (2006) argue that Knowles et al. (2001) approach is affected by the inframarginality bias and, as noted by Arnold et al. (2018), the validity of OLS for this problem requires very strong distributional restrictions.

In the context of this discussion, we propose a novel observational approach that seeks to limit the scope of inframarginality bias. Intuitively, if we can identify, within the sample of released defendants, the ones that are more likely to behave as marginals, then we can compute the conditional expectations within samples of defendants that approximate the behavior at the margin and, therefore, reduce concerns for inframarginality bias. Our approach does not require quasi-random assignment of judges for its implementation and allows for non-monotonicities in judge behavior, at the cost of assumptions that we argue are weaker than the implied restrictions of alternative observational approaches. Thus, we believe our approach is an attractive alternative in many settings where the instrument-based approach cannot be properly implemented.

3 The Prediction-Based Outcome Test

In this section, we describe and formalize the Prediction-Based Outcome Test (P-BOT). We discuss identification and estimation, as well as the scope and limits of our method.

Notation. In what follows, we classify all variables that affect the release decision $(G_i, Z_i, j(i))$ into variables that are observed by the econometrician, X_i, and variables that are not, V_i. With this notation we can write $R_i^* = f(X_i, V_i)$, where f is some function, so $R_i = 1\{f(X_i, V_i) \geq 0\}$. The only variable we impose to belong to X_i is G_i since the identification of marginal individuals is, ultimately, an input for testing for prejudice against G_i.

3.1 Intuition

To implement the outcome test, the econometrician needs to identify the defendants that were marginally released. This status is usually not observable. Absent any guidance, the econometrician can compare the average behavior of all released defendants. However, if risk distributions vary by group, differences in averages may be uninformative of the behavior at the margin. In this context, our method helps the econometrician to restrict the sample of released defendants to the ones whose behavior is more likely to approximate the behavior at the margin, so averages computed using these subsamples are less likely to be affected by the inframarginality bias.

Formally, the objects of interest to compute the outcome test (see equation (2)) are:

$$\mathbb{E}[PM_i | G_i = g, R_i^* = 0], \tag{4}$$

for each $G_i \in \{0, 1\}$. If the econometrician does not observe R_i^*, she can predict or estimate which released defendants have $R_i^* = 0$ or, alternatively, which released defendants behave similar to the ones with $R_i^* = 0$ in terms of pretrial misconduct rates. One way to interpret Knowles et al. (2001) results is as if they were answering this very question: under their structural assumptions, the behavior of the average released defendant coincides with the behavior of the marginally released defendant, so conditional expectations using all defendants with $R_i^* \geq 0$ identify (4).

In this paper, we explore an alternative approach to answer the same high-level question. When R_i^* is not observed, how can the econometrician predict the behavior of defendants with $R_i^* = 0$ using observational data? To this end, throughout the analysis we make the following assumption:

ASSUMPTION 0 (A0): *the expectation of PM_i given G_i and R_i^* is continuous in R_i^* at $R_i^* = 0$:*

$$\lim_{\varepsilon \to 0} \mathbb{E}[PM_i | G_i = g, R_i^* \in [0, \varepsilon]] = \mathbb{E}[PM_i | G_i = g, R_i^* = 0]. \tag{5}$$

A0 establishes a smooth relationship between the outcome, PM_i, and the latent risk, R_i^*, for released individuals that are *close* to the margin of release. This regularity assumption is the intuitive basis for our identification argument: outcomes for defendants at the margin can be approximated by outcomes for defendants *close* to the margin. This intuition can offer guidance to select samples of treated individuals that are more likely to behave as marginals and, therefore, are less likely to be subject to inframarginality bias. In simple words, when ε is *small*, then $\mathbb{E}[PM_i | G_i = g, R_i^* \in [0, \varepsilon]]$ constitutes a good approximation of (4). Hence, the econometrician can implement the outcome test by identifying released defendants with $R_i^* \in [0, \varepsilon]$, for *small* values of ε.

Figure 1 illustrates this intuition. Both figures use simulated data with group-specific distributions of R_i^* and PM_i that meet A0 (see Appendix B for details). Panel (a) considers a case where there is prejudice against minority defendants, with a difference in effective thresholds of 0.2. Panel (b) considers a case with no prejudice, so the difference in effective thresholds is zero. The y-axis measures differences in pretrial misconduct between non-minority and minority defendants, while the x-axis specifies the maximum percentile of the distribution of R_i^* among released defendants that is used to compute the conditional expectations. For example, $x = 50$ means that the difference in pretrial misconduct rates is computed using released defendants in the bottom 50% of the distribution of R_i^*. Both figures show that, in this example, using the whole sample of released defendants ($x = 100$) gives wrong conclusions regarding prejudice, but that the differences in pretrial misconduct rates converge to the differences in effective thresholds as the percentile decreases. When the percentile of R_i^* is *small* (which coincides with ε being *small*), the behavior of marginally released defendants can be approximated by the behavior of defendants with $R_i^* \in [0, \varepsilon]$.

The intuition above is not directly applicable when the econometrician does not observe R_i^* because the released defendants with $R_i^* \in [0, \varepsilon]$ cannot be directly identified in the data. Our proposal, then, is to use the following observational object:

$$\mathbb{E}[PM_i|G_i = g, R_i = 1, q\left(p(X_i)\right) \leq \tau], \tag{6}$$

to approximate $\mathbb{E}[PM_i|G_i = g, R_i^* \in [0,\varepsilon]] = \mathbb{E}[PM_i|G_i = g, R_i = 1, R_i^* \leq \varepsilon]$, where $p(X_i) = \mathbb{E}[R_i|X_i]$ is the propensity score, q is the quantile function (conditional on release), and τ is a *small* quantile. In words, (6) is the mean outcome among released individuals with $G_i = g$ and *low* propensity scores conditional on release. Note that $q\left(p(X_i)\right) \leq \tau$ is equivalent to $p(X_i) \leq q^{-1}(\tau) \equiv a$. We call our estimator the Prediction-Based Outcome Test (P-BOT) because the sample selection procedure relies on the predicted release status (i.e., the propensity score). This feature is appealing because it reduces the challenges of the outcome test to estimating $p(X_i)$.

The intuition for using (6) to estimate (4) is the following. If a released defendant's $p(X_i)$ is *large*, it seems unlikely that the defendant was close to being detained. By contrast, if a released defendant's $p(X_i)$ is *small*, it is more likely that the defendant was closer to being detained. This intuition suggests that released individuals with smaller propensity scores, say, $p(X_i) \leq a$, are more likely to have $R_i^* \leq \varepsilon$, for some ε. Implicit in this intuition, however, are restrictions on the conditional distribution of V_i. Intuitively, if conditional on X_i, the variance of V_i is *large*, then individuals with *large* $p(X_i)$ may have draws of V_i such that the true R_i^* is *small*, and vice-versa.

Figure 2 illustrates this intuition. In both panels, the y-axis represents the propensity score $p(X_i)$ and the x-axis the latent release status R_i^*. The seven observations shown are ranked according to their propensity score, with $p_1 < p_2 < \cdots < p_7$, and the blue curves represent the distribution of R_i^* conditional on $p(X_i)$, whose shape is determined by the conditional distribution of V_i given X_i. Panel (a) displays a situation where the conditional variance of V_i is relatively *small*, a case where the P-BOT is expected to work properly. Because of V_i, every X_i has an associated conditional distribution of R_i^*. As some of these distributions overlap, this implies that a ranking based on $p(X_i)$ does not necessarily match a ranking based on R_i^*. However, because these variances are relatively *small*, the realization of R_i^* for observations 1, 2, and 3, will definitely be smaller than the ones of observations 6 and 7. The implication is that, under assumption A0, the average pretrial misconduct of observations 1, 2, and 3 will be better approximations of equation (4) than the average behavior of observations 6 and 7. Therefore, excluding observations 6 and 7 when computing the conditional expectations will attenuate inframarginality concerns relative to a case in which the researcher uses the whole sample to compute the averages.

In practical terms, the researcher needs to make a decision on the observations included in the computation of the conditional expectation, which amounts to define a maximum value for $p(X_i)$ denoted by a. In an ideal world, all observations with $p(X_i) \leq a$ should have $R_i^* \leq \varepsilon$. However, because of the conditional variance of V_i this cannot be guaranteed. In the example displayed in Panel (a), observations 4 and 5 may induce problems. $p_4 \leq a$ is included in the estimation but may have a realization of R_i^* above ε. On the contrary, $p_5 > a$ is not included in the estimation but may have a realization of R_i^* below ε. This potential rank reversal around ε may generate inframarginality bias. However, this bias will be limited when the conditional variance of V_i is small as the potential differences in R_i^* (and per A0 in PM_i) between observations 4 and 5 are presumably small. Moreover, the potential bias from these local rank reversals will be much less problematic than the potential bias incurred by including observations 6 and 7.

On the contrary, Panel (b) displays a situation where the conditional variance of V_i is *large*, a case where the P-BOT is unlikely to work. In this case, all conditional distributions of R_i^* given X_i overlap and, therefore, a ranking based on $p(X_i)$ is not informative of a ranking based on R_i^*. As such, computing the conditional expectation (4) using observations 1 to 4 converges to a noisier version of the standard practice of using the complete sample of released individuals, known to be subject to inframarginality bias. It follows that the conditional variance of V_i given X_i will be a key input for assessing the performance of the P-BOT.

3.2 Assumptions

Now we formalize the argument sketched above. First, we make a separability assumption.

ASSUMPTION 1 (A1). *There are functions d and g such that*:

$$1\{f(X_i,V_i) \geq 0\} = 1\{d(X_i) - g(V_i) \geq 0\} \equiv 1\{d(X_i) - W_i \geq 0\}. \tag{7}$$

A1 says that there is an additively separable representation of the selection equation, which in turn implies a monotonicity restriction (Vytlacil, 2002). One suggestive assessment of A1 is to regress R_i on X_i in samples of defendants with (presumably) different V_i and compare the sign of the estimated coefficients. We discuss this suggestive test with more detail in Appendix D and illustrate it in our empirical application.[5]

The following assumption restricts the conditional distribution of W_i given X_i.

ASSUMPTION 2 (A2). *The structure of W_i given X_i is given by*:

$$W_i = \mu(X_i) + \sigma\zeta_i, \tag{8}$$

where $\sigma > 0$ and ζ_i is a scalar random variable independent from X_i with $\mathbb{E}[\zeta_i] = \overline{\zeta} < \infty$ and strictly increasing CDF.

A2 says that the correlation between X_i and W_i is unrestricted, since $\mu(X_i)$ is a flexible function. However, A2 imposes homoskedasticity in the conditional variance of W_i. This is a strong assumption. The homoskedasticity restriction makes the identification argument and its underlying intuition more transparent. Below we characterize patterns of heteroskedastic conditional variance, $\sigma(X_i)$, under which the analysis remains valid.[6]

[5] Recall from equation (1) that some structure in the judge decision leads to $f(X_i,V_i) = h(X_i,V_i) - p(X_i,V_i)$. Then, through the lens of the model, sufficient conditions for meeting A1 are given by (i) $h(X_i,V_i) = h_X(X_i) + h_V(V_i)$, and (ii) $m(X_i,V_i) = m_X(X_i) + m_V(V_i)$. While (i) is not testable, (ii) implies monotonicity on observables in the expected risk equation. Then, an additional suggestive assessment of A1 is to regress PM_i on X_i in samples of released defendants with (presumably) different V_i and compare the sign of the estimated coefficients, with the caveat that PM_i is possibly contaminated by selection in unobservables. We illustate this exercise in our empirical application.

[6] Importantly, A2 does not impose that the variance in risk is constant across groups, as the homoskedasticity restriction is conditional on the vector X_i that includes G_i and presumably several other covariates. To the extent that the distribution of X_i varies with G_i, it is still the case that both groups may have very different risk distributions with heterogeneous variances even if the conditional variance of V_i given X_i is homoskedastic. Then, this model still allows for heterogeneous variances by group as in Aigner and Cain (1977) and Bartoš et al. (2016). For the same reason, A1

Discussion. To assess A1 and A2, it is illustrative to compare them to the assumptions required by other methods. Arnold et al. (2018) show that the Knowles et al. (2001) recommendation of using the average behavior of selected individuals implies strong restrictions on the conditional distribution of unobservables to avoid inframarginality bias (equal risk distributions across groups or constant treatment effects across the risk distribution). We view our assumptions as an improvement relative to the observational literature, especially, regarding the flexibility of $\mu(X_i)$.

On the other hand, under the assumption that a valid instrument is available, there is a tradeoff between our assumptions and the necessary conditions of the IV approach. To see this, assume that judges are randomly assigned and that the only X_i the econometrician observes on top of G_i is judge leniency. In this instance, A1 coincides with the LATE monotonicity assumption. Moreover, random assignment implies that A2 is trivially met and, more generally, the IV approach does not restrict conditional variances. Yet, if, for example, judges' behavior is non-monotone, the LATE monotonicity assumption is likely to be violated. A1 becomes more flexible in that regard since $d(X_i)$ is unrestricted and, therefore, can accommodate non-monotone prejudice patterns at the judge-level if X_i contains additional observables. Within the IV framework, one solution is to compute the instrument for finer groups (as in the conditional monotonicity argument of Muller-Smith, 2015). This, however, is likely to induce power problems. This flexibility in A1 comes at two specific costs. First, adding variables to X_i that are not as good as randomly assigned means that A2 is potentially more restrictive. Second, through the lens of our model, A1 induces conditions on the risk generating process that are absent in the instrument-based approach.[7]

3.3 Identification

Proposition II summarizes the identification argument.

and A2 do not mechanically imply the absence of inframarginality bias.

[7] It is natural to conjecture that, conditional on A1, A2 holds without loss of generality. If $1\{d(X_i) \geq W_i\}$, then there is a representation of the selection rule where $1\{p(X_i) \geq U_i\}$, with $U_i \sim \mathscr{U}[0,1]$. This representation meets A2 by construction. However, this representation is not valid for the perturbed model, $1\{d(X_i) \geq W_i + \varepsilon\}$, unless additional restrictions on the conditional distribution of W_i given X_i are imposed. Then, we need to impose restrictions on the *structural* representation of the selection rule. See Appendix A for more details.

PROPOSITION II. *Assume A1 and A2 hold, and let $a > 0$ be fixed. Then:*

$$\mathbb{E}\left[PM_i | G_i = g, R_i = 1, p(X_i) \leq a\right] = \mathbb{E}\left[PM_i | G_i = g, R_i = 1, R_i^* \leq \varepsilon(a) + u_i\right], \tag{9}$$

where $p(X_i) = \mathbb{E}[R_i|X_i]$, $\varepsilon(a)$ is constant given a, with $\varepsilon'(a) > 0$, and $u_i = -\sigma\left(\zeta_i - \overline{\zeta}\right)$.

Proof. See Appendix A.

The proposition says that released individuals with low propensity scores are close to the margin up to an uncorrelated mean-zero error. Since $\varepsilon'(a) > 0$, the lower the propensity score threshold, the lower the implied ε-distance to the margin of release. Together with A0, this result implies that when a is *small*, (6) may approximate the behavior of defendants at the margin depending on the behavior of u_i. The result produces five aspects that warrant further discussion.

Choosing a. $\varepsilon(a)$ is a deterministic function. Therefore, there exists some $\overline{a} > 0$ such that $a \to \overline{a}$ implies that $\varepsilon(a) \to 0$.[8] However, as shown in Appendix A, $\overline{a}$ is a function of the distribution of ζ_i that is, in principle, unknown for the econometrician. Therefore, we don't pursue a theoretical characterization of an optimal a and discuss its practical choice below when discussing estimation. For justifying this avenue, we take advantage of the fact that (1) by construction, ε is positive for all selected observations, and (2) $\varepsilon(a)$ is strictly monotone in a. Both points together imply that choosing a smaller a will necessarily lead to better approximations of the margin, the reason why we frame the discussion below in terms of quantiles of $p(X_i)$ instead of levels.

Prediction. The identification argument relies on the predicted release status but not on the specifics of the prediction model. This makes our approach robust to omitted variables in the sense that, as V_i is not observed, it possibly biases the estimated coefficients of the prediction model, but the same bias improves the prediction of the propensity score. Monte Carlo exercises presented in Appendix B confirm this intuition. This argument is a restatement of the fact that u_i is

[8] The requirement of $\overline{a} > 0$ comes from the observation that a vector X_i with $p(X_i) = 0$ corresponds to a unit that is never treated and, therefore, could never be associated with a marginally treated one. Then, the proper thought experiment is to take $a \to \overline{a}$ where $\overline{a} > 0$ is some value that, in the abstract, approximates the selection thresholds that are likely interior in the interval $[0, 1]$, as treated units must have positive propensity scores.

an uncorrelated mean-zero error, which implies that there are no systematic errors when predicting which defendants are closer to the margin of release. The reason is that the econometrician only needs to know *who* is close to the margin, but not *why*.

Inframarginality bias. Proposition II states that our proposed estimator is possibly affected by the inframarginality bias because of the intuition depicted in Figure 2. While $\mathbb{E}[u_i] = 0$ and, therefore, the prediction error is not systematic, the conditional expectation is possibly a non-linear function of R_i^*, so u_i probably has a non-zero impact on the object of interest. To understand this source of bias, consider two defendants, 1 and 2, with realizations $u_1 > 0$ and $u_2 < 0$, with $u_1 + u_2 = 0$. Assume that $R_1^* > \varepsilon$ but $R_1^* \leq \varepsilon + u_1$. Likewise, assume that $R_2^* \leq \varepsilon$ but $R_2^* > \varepsilon + u_2$. In this case, defendant 1 is not within an ε-distance from the margin, but would be included in the estimation sample, while defendant 2 is within an ε-distance from the margin but would be excluded from the estimation sample. These misclassifications can induce inframarginality bias if the outcome distribution differs between the *wrongly included* and the *wrongly excluded* defendants.

The previous argument implies that whether the proposed estimator (equation (6)) constitutes a good approximation of the object of interest (equation (4)) depends on the *magnitude* of the misclassification bias driven by these local rank reversals. The key determinant of this concern is the conditional variance of W_i given X_i, which under the homoskedasticity restriction is equal to $\mathbb{V}(u_i) = \sigma^2\mathbb{V}(\zeta_i)$. If the conditional variance is *small*, then (i) relatively *few* defendants will be misclassified, and (ii) the misclassified defendants will have *similar* values of R_i^* which, under a continuity assumption on the MTE frontier, would imply that their pretrial misconduct behavior is possibly similar. Then, the scope for inframarginality bias would be limited. In fact, as this variance goes to zero, u_i also goes to zero and (6) coincides with (4). However, if the conditional variance is *large*, then the econometrician could be in situations where several defendants with a wide range of R_i^* (and, therefore, large potential differences in potential outcomes) are misclassified. In that situation, the scope for inframarginality bias may be substantial. It is important to acknowledge that, if $\mathbb{V}(u_i) \to 0$, the propensity score becomes a step function and, therefore, the proposed test is not implementable. This feature means that some degree of inframarginality bias is inherent to the implementability of the P-BOT. The potential bias, however, depends on how local the rank reversals are (especially around ε) and, therefore, can be substantially attenuated when deviations are small relative to variation in $p(X_i)$, especially compared to the standard observational practice of computing unconditional averages among the treated.

An alternative framing of the previous intuition is that the bias decreases when the fit of the propensity score (or, alternatively, the predictive power of the available observables, X_i) increases. In Appendix B we present Monte Carlo simulations that show that as the conditional variance increases, our test converges to Knowles et al. (2001)'s test. Intuitively, when the predictive power of X_i is very weak, the propensity score flattens and the sample of defendants with *low* propensity score converges to a random sample of released individuals. Then, the scope for inframarginality bias can be indirectly assessed by evaluating the fit of the propensity score estimation.

The role of homoskedasticity. Proposition II is derived under the assumption that the conditional variance is homoskedastic, which is a restrictive assumption. If we extend A2 to allow for heteroskedasticity through an unrestricted function $\sigma(X_i)$, the classification error is still mean-zero but depends on X_i. More formally, in Appendix A we show that when $\sigma(X_i)$ is unrestricted, $p(X_i) \leq a(\varepsilon)$ implies $R_i^* \leq \varepsilon + \Gamma(X_i, u_i)$, with $\mathbb{E}[\Gamma(X_i, u_i)] = 0$. Hence, while the interpretation of the prediction error depends on whether the particular X_i induces a relatively *large* or *small* $\sigma(X_i)$, the intuition of the bias and the role of the propensity score fit to assess its pervasiveness is as in the discussion above.[9] This robustness can also be inferred from Figure 2: replicating that figure with heterogeneous variances would yield the same conclusions regarding local rank reversals.

Empirical test for inframarginality bias: the perturbation test. One insight from the discussion above is that the scope for inframarginality bias is proportional to the conditional variance of the prediction error even when $\sigma(X_i)$ is heteroskedastic. Since the conditional variance of the prediction error can be estimated by adding additional structure to the unobserved component, we can simulate perturbations to assess the pervasiveness of the inframarginality bias. We for-

[9] The scope for misclassification-driven bias is smaller when the heteroskedasticity patterns follow a monotonicity property, that is, when observations with *low* propensity scores also have smaller $\sigma(X_i)$ (see Appendix A).

malize this idea in the *perturbation test*. Under A1 and A2, the selection rule can be written as $R_i = 1\left\{\frac{d(X_i)-\mu(X_i)}{\sigma(X_i)} \geq \zeta_i\right\}$. Since the econometrician observes R_i and X_i, it is possible to estimate the left-hand-side and the corresponding variance of ζ_i with additional parametric restrictions. The estimated variance of ζ_i can be then used to simulate perturbations that alter the estimated propensity score and, therefore, the defendants that are considered to be close to the margin. By recomputing the outcome test on each of these simulations, the econometrician can check how the result varies with these perturbations to diagnose concerns for potential local rank reversals. Below we provide applications of this test when the prediction is done with a probit estimator.

3.4 Estimation and implementation

The estimator proposed in equation (6) requires the econometrician to estimate the propensity score, use the predicted release probabilities to rank released defendants, and estimate the outcome equation on a sample of defendants at a given margin definition. We propose two approaches for implementing the P-BOT. In what follows, let $\widehat{p}(X_i)$ denote the estimated propensity score.

Simple approach. This approach defines the sample of defendants close to the margin based on the quantiles of the predicted propensity score, that is, a defendant is included in the estimation sample if $q(\widehat{p}(X_i)) \leq \tau$, where τ is *small* and $a = q^{-1}(\tau)$. Then, the outcome test can be implemented estimating a linear regression of PM_i on G_i using the truncated sample. Negative and significant estimates of the coefficient on G_i constitute evidence of prejudice against group $G_i = 1$.

There is a bias-variance tradeoff in the choice of τ: while choosing a larger τ mechanically increases the sample size and therefore improves the precision of the estimation, it also implies that the outcome equation is estimated using a larger share of inframarginal individuals. Formally, the larger τ, the larger the implied ε and, therefore, the less likely that the limit argument based on A0 applies to the estimation. In fact, when $\tau = 100$, the P-BOT is equivalent to Knowles et al. (2001) test. This leads to an additional inframarginality test: the econometrician can assess the pervasiveness of the potential bias by analyzing the sensitivity of the estimation to the choice of τ.

Non-parametric approach. As a refinement, we suggest performing non-parametric local regressions at lower percentiles to better approximate the notion of the limit defined in A0. Formally, the econometrician can estimate $\mathbb{E}[PM_i|G_i = 0, q(\widehat{p}(X_i)) = 1]$ and $\mathbb{E}[PM_i|G_i = 1, q(\widehat{p}(X_i)) = 1]$, and assess the extent of prejudice by taking the difference. Theoretically, the econometrician could condition on $\widehat{p}(X_i) = \min_j\{\widehat{p}(X_j)\}$ given that these expectations have to be estimated when $\varepsilon \rightarrow 0$. We suggest, however, focusing on the 1st percentile to avoid bias due to outliers in the predicted probabilities. An advantage of this approach is that it weights observations according to their relative distance to the margin of release.

Choice of X_i. Since the potential inframarginality bias depends on the residual variance conditional on X_i, the econometrician would like to include as much as possible in X_i to the extent that it improves the quality of the prediction and that these are valid predictors in terms of being predetermined with respect to the treatment decision. In this context, the econometrician could benefit from machine learning techniques that penalize predictors that induce noise to the estimation without improvements in the propensity score fit. Also, the inclusion of multiple variables in X_i increases the support of $p(X_i)$, making the ranking argument more economically meaningful.

Inference. The distributions of the two proposed estimators must consider that the sample definition criterion is estimated, since it is based on the estimated propensity score. We therefore suggest using bootstrap to calculate confidence intervals on the outcome test.[10]

Perturbation test. The *perturbation test* can be implemented as follows. We focus on instances where the propensity score is estimated using a probit model. First, estimate a probit model for the release status. Then, for each released individual, simulate M realizations from a standard normal distribution. This standardized normally distributed random variable corresponds to the (standardized) ζ_i from the previous subsection.[11] Finally, for each of the M realizations, and given

[10] The bootstrap is not always valid in two-step estimations or in RDD and propensity score-based procedures (Abadie and Imbens, 2008; Calonico et al., 2014; Cattaneo and Jansson, 2018; Cattaneo et al., 2019). Our approach, however, is a simple difference in means using a generated regressor that determines the sample of the second step. To the extent that the process that generates the regressor is continuous, the bootstrap is consistent for the P-BOT. This suggests that this inference strategy is valid whenever the propensity score is continuous in X_i.

[11] In a probit model the point estimates are estimations of the regression coefficients divided by the standard devia-

the estimated parameters of the probit model, simulate R_i^* for all released defendants, redefine samples of released defendants close to the margin of release, and re-estimate the group-specific pretrial misconduct rates. Then, the econometrician can assess the bias induced by the classification errors by examining the distribution of the P-BOT estimate across all simulations.

3.5 Discussion

Our approach has four attractive properties. First, the prediction-based argument makes the P-BOT robust to standard omitted variable bias. Second, the P-BOT requires neither instruments nor the random assignment of judges, and allows for non-monotone prejudice patterns. Third, the potential for inframarginality bias can be estimated and, therefore, empirically assessed. Fourth, its implementation is simple. We note two main limitations. First, our identification strategy relies on assumptions that may be restrictive in some settings. Second, the P-BOT's ability to limit the scope of inframarginality bias depends on the availability of good predictors.

4 Empirical Application: Institutional Setting and Data

In the remainder of the paper, we illustrate our approach with an empirical application. We test for prejudice in pretrial detentions against the largest ethnic minority group in Chile, the Mapuche, using nationwide administrative data. This section describes the institutional setting and data.

4.1 Setting

The current criminal justice system in Chile was implemented in 2005 and works uniformly throughout the country. The procedure to define pretrial detention for arrested people is as fol-

tion of the error. The size of the conditional variance is then incorporated in the magnitude of the estimated coefficients. Formally, if $\zeta_i \sim \mathscr{N}(\overline{\zeta}, \sigma_\zeta^2)$, we can write $R_i = 1\left\{\frac{1}{\sigma_\zeta}\left(\frac{d(X_i)-\mu(X_i)}{\sigma(X_i)} - \overline{\zeta}\right) \geq \tilde{\zeta}_i\right\}$, where $\tilde{\zeta}_i \sim \mathscr{N}(0,1)$. Then, the probit model estimates the left-hand-side and simulations of $\tilde{\zeta}_i$ can be used to perturb the estimated ranking.

lows. During the 24 hours after the initial detention, there is an in-person arraignment hearing in which a detention judge determines if the defendant will be incarcerated during the investigation. Since monetary bail is not an option in the Chilean system, the judge's decision is effectively binary. Following the legal principle of presumption of innocence, judges should not incarcerate defendants unless there is a high probability of failing to appear in court, there is a high probability of committing a new crime during the investigation, or imprisonment aids the investigation of the criminal case. The law does not emphasize a particular type of pretrial misconduct that judges should prioritize, implicitly putting equal weight on the different outcomes. In general, the arraignment hearing is very brief (lasting about 15 minutes) and is carried out by quasi-randomly assigned judges at the court-by-time level.[12] The information set for judges includes administrative information available in the criminal justice system – including detailed criminal records – as well as an arrest report that specifies the crimes that are being charged and some preliminary evidence.

We test for prejudice against the largest ethnic minority group in Chile, the Mapuche. Around 10% of the Chilean population identifies as Mapuche according to the Chilean Census. The Mapuche population provides an interesting case study for three reasons. First, a long-running conflict exists between the Mapuche and the Chilean state dating back more than a century (Cayul et al., 2022). In this context, it is frequently claimed that the Chilean institutions are biased against the Mapuche. Second, the Mapuche people are subject to numerous negative stereotypes, such as tendencies towards laziness, violence and alcoholism, from some quarters of Chilean society. Third, Mapuche people are identifiable, mainly because of their surnames but also to some extent due to their physical appearance. Thus, prejudice against members of this group is feasible.

4.2 Data

We use administrative records from the Public Defender's Office (PDO). The PDO is a centralized public service under the oversight of the Ministry of Justice. It offers criminal defense services

[12] In every court at the beginning of each month, judges are non-systematically assigned to different time slots to lead arraignment hearings for no reason other than splitting the duty among the court's judges.

to all individuals accused of or charged with a crime; as such, it ensures the right to a defense by a lawyer and due process in criminal trials. Our estimation sample covers more than 95% of the criminal cases for the period between 2008 and 2017, and contains detailed case and defendant characteristics. In addition, we can identify the judges and attorneys assigned to each case at the beginning of the criminal process (i.e., when the determination of pretrial detention occurs).

We observe defendants' self-reported ethnicity. However, since self-reported ethnicity is subject to measurement error because of potential under-reporting, we merge the administrative data with a register of Mapuche surnames to build more robust measures of ethnicity. Since Chilean citizens are identified by both their father and mother's surnames, we define the following Mapuche indicators: defendants are identified as Mapuche if they (i) have at least one Mapuche surname, (ii) have two Mapuche surnames, (iii) self-report as being Mapuche, or (iv) have at least one Mapuche surname or self-report as being Mapuche (our preferred and most comprehensive definition). On the other hand, defendants are identified as non-Mapuche if condition (iv) fails to hold.[13]

To build the estimation sample, we consider all detention hearings for adult defendants who were arrested between 2008 and 2017. We exclude hearings due to legal summons, since the information set available to the judge may be different in those cases. To focus on arraignment hearings in which pretrial detention is a plausible outcome, we only consider types of crimes with at least a 5% probability of pretrial detention. For the same reason, when defendants are accused of more than one crime during the same arraignment hearing, we only retain the information related to the most severe crime (with severity measured as the probability of pretrial detention). Finally, we exclude cases assigned to judges or attorneys with fewer than 10 cases. A more detailed description of the data, the sample restrictions, and the variables is presented in Appendix C.

Descriptive statistics. Table 1 presents descriptive statistics from our estimation sample. Mapuche defendants represent 7.4% of the total sample when we consider our most comprehensive definition of Mapuche ($52,001/699,730$). Release occurs in about 84% of the cases, with a minor

[13] We exclude defendants that self-report as belonging to other ethnic groups (0.4% of the cases).

difference in favor of Mapuche defendants. In terms of the outcomes that pretrial detention seeks to avoid, conditional on being released, between 23% and 30% of the defendants (depending on the group) engage in at least one type of pretrial misconduct, either non-appearance in court or pretrial recidivism. Across all measures of pretrial misconduct, released Mapuche defendants demonstrate better conduct during prosecution than released non-Mapuche defendants. On average, the criminal records of Mapuche defendants are less severe, measured as both the number of previous cases and their severity. The current cases of Mapuche defendants are also slightly less severe.

5 Empirical Application: Results

This section presents the results of our empirical application. First, we discuss the prediction model and present some diagnostics. Then, we perform the P-BOT and the perturbation test. Then, we perform alternative tests for prejudice and compare the results. Finally, we develop extensions to the basic model to discuss the interpretation of the outcome test.

5.1 Prediction model

We estimate the propensity score using a probit model and consider the following covariates defined in Section 4: a Mapuche indicator, a male indicator, whether the individual has previous prosecutions, the number of previous prosecutions, the severity of previous prosecutions, whether the individual engaged in pretrial misconduct during a previous prosecution, whether the individual has been convicted in the past, the severity of the current prosecution, the number of cases seen in the court during the year of the prosecution, the number of judges working at the court during the year of the prosecution, the assigned public attorney's quality and its square, the assigned judge's leniency and its square, and year of prosecution fixed effects.[14] While the probit model does not return out-of-bounds predictions, it may be limited in the number of fixed effects that

[14] Following Dobbie et al. (2018), judge leniency is the leave-out mean release rate, after adjusting for court-by-year fixed effects. We follow the same procedure to compute the quality of the attorney.

can be included in the estimation. Then, we also compute the release probabilities using a linear probability model adding court fixed effects. We also use Lasso to select regressors considering all interactions and squared terms, and judge fixed effects. Finally, we also fit a heteroskedastic probit model. Since results are consistent between models, we restrict our discussion to the baseline probit case. Results using alternative prediction models can be found in Appendices E and G.

Propensity score fit. Appendix E shows the results of the probit model. We perform different diagnostics for the propensity score fit. Considering 0.5 as the probability threshold, 85% or more of the cases are correctly classified by the prediction model (86% for Mapuche and 85% for non-Mapuche defendants). We also perform an out-of-sample cross-validation exercise that gives similar conclusions.[15] Finally, we apply the methods of inference for rankings set out in Mogstad et al. (2022) and conclude that more than 80% of the released defendants that are in the bottom 5% of the predicted propensity score distribution have true propensity scores in the bottom 5% of the distribution, with 95% confidence.[16]

Which observables matter the most? To assess the relative role of the different covariates in the propensity score fit, we test whether excluding elements from X_i affects the predicted propensity score ranking and, therefore, the samples of released defendants that are considered to be *close to the margin*. In other words, we test which individual covariates help the most to reduce the conditional variance by assessing the stability of the estimation sample to restrictions on the propensity score model. Appendix D discusses this exercise in more detail and presents the results. Results suggest that the propensity score ranking and the samples of marginal defendants are very stable across permutations. This follows from the fact that the elements in X_i are likely to be correlated and, therefore, the included variables do a good job of capturing the omitted variation. The only variable that, when excluded, induces substantially different estimation samples and noisier

[15] We randomly select 90% of the sample, estimate the probit model, and compute the correct classified cases in the remaining 10%. We repeat the exercise 50 times. In all repetitions, 85% of the cases are correctly classified.

[16] We calculate standard errors for the predictions based on the probit model and compute the joint (simultaneous) confidence sets for the ranks. Then, we count how many individuals in the estimation sample have ranking upper bounds within the bottom 5% (as in their τ-worst suggested procedure). For computational feasibility, we consider the 40,000 observations of released individuals with lower estimated propensity scores, use three decimals for the predicted probabilities and their standard errors, and derive critical values using 100 bootstrap repetitions.

propensity score rankings is the severity of the current case, which suggests that the imputed type of crime is an important variable for judges when determining pretrial release that is not fully informed by criminal records, demographics, or court characteristics.

Testing A1. Proposition II relies on A1, which implies monotonicity in observables in the selection equation. Based on equation (1), A1 also implies monotonicity in observables in the risk equation. Appendix D shows that the coefficients of regressions of R_i and PM_i on observables are very stable (in terms of sign and magnitude) when they are estimated using subsamples with presumably different unobservables. We include several observables in each regression and consider eight different criteria for splitting the sample. In 97.5% (78/80) of the cases, the sign of the coefficient is consistent between subsamples. We interpret this as suggestive evidence in favor of A1.

5.2 Outcome equation

To formally test for prejudice, we use the predicted propensity score to rank released defendants and build estimation samples following the procedure described in Section 3. We present results for the most comprehensive definition of Mapuche (self-reported or at least one surname). Results for the other definitions are presented in Appendix F.

As a first exploratory analysis, we analyze how the P-BOT varies as we increase the estimation sample. We first calculate the Mapuche and non-Mapuche averages of pretrial misconduct only considering the first quintile of the distribution of the predicted release probability among released defendants, then the first and the second quintiles, and so on until we consider the entire sample. Figure 3 shows the result. The outcome is defined as any pretrial misconduct. We highlight three implications from the figure. First, the Mapuche defendants' pretrial misconduct rate is below the non-Mapuche defendants' rate in the first quintile of the predicted probability distribution, which suggests prejudice against Mapuche defendants. Second, for both groups, the rates of pretrial misconduct decrease as we add defendants with a higher probability of release. This result suggests that defendants that are more likely to be released are also less likely to be engaged in pretrial misconduct, which in turn suggests that judges care about expected outcomes when making pretrial detention decisions. Finally, the two lines are mostly parallel with a slightly wider gap in the first quintile. This suggests that in our setting the potential inframarginality bias exists but is modest.

Going beyond the graphical evidence, Table 2 presents the results of the implementation of the P-BOT. In Panel A, we implement the simple approach, where the point estimate is obtained from a linear regression of pretrial misconduct on a Mapuche indicator in a sample of released defendants with *low* predicted propensity scores. We consider two thresholds for the propensity score distribution: the bottom 5% and bottom 10%. In Panel B, we implement the non-parametric version, where the point estimate is obtained by subtracting the Mapuche and non-Mapuche conditional expectations for pretrial misconduct, which are non-parametrically calculated at the first percentile of the estimated release probability distribution. Point estimates are negative and statistically significant, suggesting evidence of prejudice against Mapuche defendants. Marginally released Mapuche defendants are between 3 and 4 percentage points less likely to be engaged in pretrial misconduct relative to marginal non-Mapuche defendants.[17]

Perturbation test. To assess the scope for inframarginality bias driven by local rank reversals, we perform the perturbation test described in Section 3. We implement the test using the coefficients of the probit model (see Appendix G for similar results using the heteroskedastic probit model). For each individual in our sample of released defendants, we simulate 500 realizations from a standardized normal distribution to simulate R_i^*, recompute the ranking, and redefine the estimation sample. Then, in each simulation, we re-estimate the outcome test and plot the distribution of the outcome test. Figure 4 shows the results. The perturbation test suggests limited scope for inframarginality bias since the distribution of the perturbed outcome test does not include zero. Even in the worst-case scenario induced by this test, the conclusion of prejudice is not reversed.

[17] Results are robust to considering non-appearance in court and pretrial recidivism as separate outcomes (see Appendix G). Also, prejudice is more than three times larger when we identify Mapuche defendants using both surnames (see Appendix F), which we conjecture is explained by the salience of the ethnicity measure.

5.3 Alternative tests

To assess the relative performance between the P-BOT and other approaches, we also test for prejudice using alternative methods. We consider the outcome test using the full sample (Knowles et al., 2001) and the instrument-based approach (Arnold et al., 2018). For the latter, we exploit the quasi-random assignment of bail judges that characterizes the Chilean setting.[18]

Table 3 presents the results. The outcome test using the full sample provides evidence of prejudice. Consistent with Figure 3, we note that the inframarginality bias biases the estimation downwards. Perhaps the most interesting analysis relates to the application of the IV approach. While the LATE for the non-Mapuche defendants is precisely estimated, the Mapuche estimation is severely underpowered. Hence, standard errors are large enough to prevent the test from making conclusions about prejudice. The case is even more problematic for the less comprehensive indicators and does not improve when adding controls for precision purposes (see Appendix F). In Appendix H we report the first-stage F-tests, which corroborates the lack of power of the instrument in the minority sample. Therefore, our setting is one in which the instrument-based approach is not well-behaved because of power problems.[19]

Moreover, note from Table 2 and Appendix G that the P-BOT's estimate of the pretrial misconduct rate of marginally released non-Mapuche defendants is between 37.4% and 42.4%, while the corresponding estimated LATE using the IV approach is between 36.3% and 37.4%. Therefore, the estimated pretrial misconduct rates of non-Mapuche marginal defendants are similar between both methods. Also, Appendix I performs a complier analysis and show that the non-Mapuche defendants identified as marginals by both methods have comparable observables. This suggest

[18] Appendix H presents the results of the randomization test suggested by Arnold et al. (2018).

[19] We also perform the test proposed by Frandsen et al. (2023) and reject the null hypothesis of strict monotonicity. While their procedure jointly tests for exclusion and monotonicity, the institutional setting suggests exclusion holds and, therefore, we interpret rejections of the null as deviations from strict monotonicity. We parametrize the test following the recommendations of the authors. For computational feasibility, we compute the test for random subsamples. We generate random samples considering (i) 25% of court-by-year cells, and (ii) 25% of bail judges. For each criterion, we build 10 random subsamples. In all subsamples, the composite p-value is 0.000. We only consider non-Mapuche defendants. Since courts, years, and judges vary in their caseloads, random samples have different sizes. Among the 20 samples used, the average sample contains 159,069 observations.

that both methods may yield similar results in cases where they are expected to work properly.

5.4 Extensions

In Section 2 we argued that the interpretation of the outcome test depends on some features of the selection process. While our notion of prejudice remains relevant from a disparate impact perspective under different interpretations, it may be of interest to disentangle between cases. In the reminder of the section we illustrate how the P-BOT can be used to explore these distinctions.

Determinants of judges' thresholds. Table 2 only tests for differences in effective thresholds between Mapuche and non-Mapuche defendants. However, prejudice patterns can be more complex, meaning that effective thresholds can also be influenced by other variables. We can use the P-BOT to test for the relevance of additional covariates in the determination of effective thresholds by adding observables to the linear regression that characterizes the outcome equation.

To illustrate the latter, Table 4 presents two examples of this extension. In Panel A, we group defendants by two categories: *Mapuche* and *low income*. The latter is calculated using the Chilean national household survey (CASEN), with *low income* equal to one if the defendant lives in a municipality whose average income is below the sample median. In Panel B, we group defendants using *Mapuche* and *Mapuche region*, which is an indicator variable that takes the value of one if the defendant lives in the Araucanía Region, the administrative region historically associated with the Mapuche conflict. We show the results from the simple version of the P-BOT for our most comprehensive Mapuche definition and using the 10% margin definition.

Table 4 shows that prejudice patterns are possibly multi-dimensional. This becomes clear when looking at the differences in the four conditional means. In Panel A, the results show that prejudice against Mapuche defendants is mainly relevant for those Mapuche who live in low-income municipalities. This suggests that the relevant prejudice is against low-income Mapuche defendants. In Panel B, the results suggest that Mapuche defendants are slightly more afflicted by prejudice in the conflict region; however, the interaction term is not statistically significant. These results suggest that non-monotone patterns of discrimination are likely to occur in practice.

Assignment rule for judges. The assignment rule matters for interpreting whether the estimated prejudice is driven by judges who are, on average, prejudiced, or by Mapuche defendants visiting courts that are, on average, less lenient. When judges are randomly assigned at the court-by-time level, implementing our simple P-BOT regression while controlling for court-by-year fixed effects will yield an estimate for prejudice net of the role of the assignment rule. We present these results in Table 5. Point estimates are about half as large as the baseline results. This suggests that institutional prejudice driven by the assignment rule is an important force driving the results.

6 Conclusion

While economists have been aware of the outcome test since Becker (1957, 1993), its implementation is not straightforward. The need to identify marginal individuals is a significant challenge.

In this paper, we propose a novel observational method – the Prediction-Based Outcome Test (P-BOT) – for identifying samples of released defendants that are more likely to be close to the margin of release, thus attenuating concerns about inframarginality bias in the outcome test. Our main result provides conditions under which released defendants that are more likely to be close to the margin of release also have smaller propensity scores. We develop a detailed discussion about the assumptions needed for identification and propose empirical diagnostics to assess the scope of inframarginality bias in our proposed implementation. We argue that the P-BOT is an attractive methodology in the absence of well-behaved instruments.

Our contribution is mostly practical, as the P-BOT increases the applicability of the outcome test while minimizing concerns for inframarginality bias. Moreover, its implementation is straightforward. The econometrician can proceed by fitting projection models for the release status, ranking released defendants according to their predicted probabilities, defining estimation samples of released defendants, and estimating simple outcome equations. The P-BOT relies on the availability of good predictors for the release status. The increasing availability of rich administrative datasets suggests that, in some cases, this may be a feasible requirement, especially relative to the requirement of having well-behaved instrument variation.

We use the P-BOT to test for prejudice in pretrial detentions against the Mapuche, the largest ethnic minority in Chile. We find strong evidence of prejudice using different outcomes, Mapuche definitions, and estimation methods, both in the projection and outcome equations. We provide evidence of modest inframarginality bias and show that the instrument-based approach has implementation issues in our setting. We show that prejudice patterns are likely to be multidimensional, and that the assignment rule of judges to defendants partly explains the overall estimated effect.

We conclude the discussion by stressing two additional features of the P-BOT that can be useful for researchers interested in the estimation of prejudice.

First, when the instrument-based approach can be properly implemented, the P-BOT can be used to perform complementary analyses that may not be pursued with the judges design. Consider a researcher that has an IV estimate of prejudice. If, for example, the researcher is interested in testing for heterogeneity by time, geography, or type of crime, the judges design may be underpowered for such analyses. Then, the researcher could test whether the P-BOT yields aggregate estimates comparable to the quasi-experimental estimates and then use the P-BOT to proceed with these heterogeneities. This is one of many examples on how the P-BOT may allow researchers to extend quasi-experimental analyses to get additional insights about prejudice patterns.

Second, the underlying model and the outcome test are useful frameworks for analyzing prejudice in a variety of contexts. In fact, Gary Becker's original ideas that gave rise to the outcome test were formalized in the context of discrimination in the labor market. In general, the outcome test is applicable to any setting where the selection process is expected to be based on a predicted (and ex-post measurable) outcome. The fact that the P-BOT does not require instruments for its implementation may foster the application of the outcome test in a broader range of settings where testing for prejudice is important but, usually, decision-makers are not randomly assigned.

References

Abadie, Alberto and Guido W Imbens, "On the failure of the bootstrap for matching estimators," *Econometrica*, 2008, *76* (6), 1537–1557.

Abrams, David S, Marianne Bertrand, and Sendhil Mullainathan, "Do judges vary in their treatment of race?," *The Journal of Legal Studies*, 2012, *41* (2), 347–383.

Aigner, Dennis J and Glen G Cain, "Statistical theories of discrimination in labor markets," *Industrial and Labor Relations Review*, 1977, *30* (2), 175–187.

Antonovics, Kate and Brian G Knight, "A new look at racial profiling: Evidence from the Boston Police Department," *Review of Economics and Statistics*, 2009, *91* (1), 163–177.

Anwar, S. and H. Fang, "An alternative test of racial prejudice in motor vehicle searches: Theory and evidence," *American Economic Review*, 2006, *96* (1), 127–151.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson, "The impact of jury race in criminal trials," *Quarterly Journal of Economics*, 2012, *127* (2), 1017–1055.

_ , _ , and _ , "Politics in the courtroom: Political ideology and jury decision making," *Journal of the European Economic Association*, 2018, *17* (3), 834–875.

Arnold, D., W. Dobbie, and C. Yang, "Racial bias in bail decisions," *Quarterly Journal of Economics*, 2018, *133* (4), 1885–1932.

Arnold, David, Will Dobbie, and Peter Hull, "Measuring racial discrimination in bail decisions," *American Economic Review*, 2022, *112* (9), 2992–3038.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka, "Attention discrimination: Theory and field experiments with monitoring information acquisition," *American Economic Review*, 2016, *106* (6), 1437–1475.

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

— , "Nobel Lecture: The economic way of looking at behavior," *Journal of Political Economy*, 1993, *101*, 385–409.

Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope, "Inaccurate statistical discrimination: An identification problem," *Review of Economics and Statistics*, 2025, *107* (3), 605–620.

— , Peter Hull, and Alex Imas, "Systemic discrimination: Theory and measurement," *The Quarterly Journal of Economics*, 2025, *140* (3), 1743–1799.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik, "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 2014, *82* (6), 2295–2326.

Canay, Ivan A, Magne Mogstad, and Jack Mountjoy, "On the Use of Outcome Tests for Detecting Bias in Decision Making," *The Review of Economic Studies*, 2024, *91* (4), 2135–2167.

Cattaneo, Matias D and Michael Jansson, "Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency," *Econometrica*, 2018, *86* (3), 955–995.

— , — , and Xinwei Ma, "Two-step estimation and inference with possibly many included covariates," *Review of Economic Studies*, 2019, *86* (3), 1095–1122.

Cayul, Pedro, Alejandro Corvalan, Dany Jaimovich, and Matteo Pazzona, "Introducing MACEDA: New micro-data on an indigenous self-determination conflict," *Journal of Peace Research*, 2022, *59* (6), 903–912.

Chan, David C, Matthew Gentzkow, and Chuan Yu, "Selection with variation in diagnostic skill: Evidence from radiologists," *The Quarterly Journal of Economics*, 2022, *137* (2), 729–783.

Chandra, Amitabh and Douglas O Staiger, "Identifying provider prejudice in healthcare," *Working Paper*, 2010.

Cohen, Alma and Crystal S Yang, "Judicial politics and sentencing decisions," *American Economic Journal: Economic Policy*, 2019, *11* (1), 160–91.

Dobbie, Will and Crystal Yang, "The Economic Costs of Pretrial Detention," *Brookings Papers on Economic Activity*, 2021.

_ and _ , "The US pretrial system: Balancing individual rights and public interests," *Journal of Economic Perspectives*, 2021, *35* (4), 49–70.

_ , Jacob Goldin, and Crystal S Yang, "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 2018, *108* (2), 201–240.

Feigenberg, Benjamin and Conrad Miller, "Would eliminating racial disparities in motor vehicle searches have efficiency costs?," *The Quarterly Journal of Economics*, 2022, *137* (1), 49–113.

Frandsen, Brigham R, Lars J Lefgren, and Emily C Leslie, "Judging judge fixed effects: Testing the identifying assumptions in judge fixed-effects designs," *American Economic Review*, 2023, *113* (1), 253–277.

Gelbach, J, "Testing Economic Models of Discrimination in Criminal Justice," *Working Paper*, 2021.

Grau, Nicolás, Gonzalo Marivil, and Jorge Rivera, "The effect of pretrial detention on labor market outcomes," *Journal of Quantitative Criminology*, 2021.

Guryan, Jonathan and Kerwin Kofi Charles, "Taste-based or statistical discrimination: The economics of discrimination returns to its roots," *The Economic Journal*, 2013, *123* (572), F417–F432.

Hull, P., "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making," *Working Paper*, 2021.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

_ , Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein, "Discrimination in the Age of Algorithms," *Journal of Legal Analysis*, 2019, *10.*

Kline, Patrick and Christopher Walters, "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination," *Econometrica*, 2021, *89* (2), 765–792.

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

Lang, K. and A. Kahn-Lang, "Race discrimination: An economic perspective," *Journal of Economic Perspectives*, 2020, *34* (2), 68–89.

Leslie, Emily and Nolan G Pope, "The unintended impact of pretrial detention on case outcomes: Evidence from New York City arraignments," *The Journal of Law and Economics*, 2017, *60* (3), 529–557.

Manski, Charles F, "Optimal search profiling with linear deterrence," *American Economic Review*, 2005, *95* (2), 122–126.

_ , "Search profiling with partial knowledge of deterrence," *The Economic Journal*, 2006, *116* (515), F385–F401.

Marx, Philip, "An absolute test of racial prejudice," *The Journal of Law, Economics, and Organization*, 2022, *38* (1), 42–91.

Merino, M. and Daniel Quilaqueo, "Ethnic prejudice against the Mapuche in Chilean society as a reflection of the racist ideology of the Spanish Conquistadors," *American Indian Culture and Research Journal*, 2003, *27* (4), 105–116.

Merino, Maria Eugenia and David John Mellor, “Perceived discrimination in Mapuche discourse: Contemporary racism in Chilean society,” *Critical Discourse Studies*, 2009, *6* (3), 215–226.

Mogstad, Magne, Joseph P Romano, Azeem Shaikh, and Daniel Wilhelm, “Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries,” *The Review of Economic Studies*, 2022.

Muller-Smith, M., “The criminal and labor market impacts of incarceration,” *Working Paper*, 2015.

Rehavi, M Marit and Sonja B Starr, “Racial disparity in federal criminal sentences,” *Journal of Political Economy*, 2014, *122* (6), 1320–1354.

Rose, Evan K, “Who gets a second chance? Effectiveness and equity in supervision of criminal offenders,” *The Quarterly Journal of Economics*, 2021, *136* (2), 1199–1253.

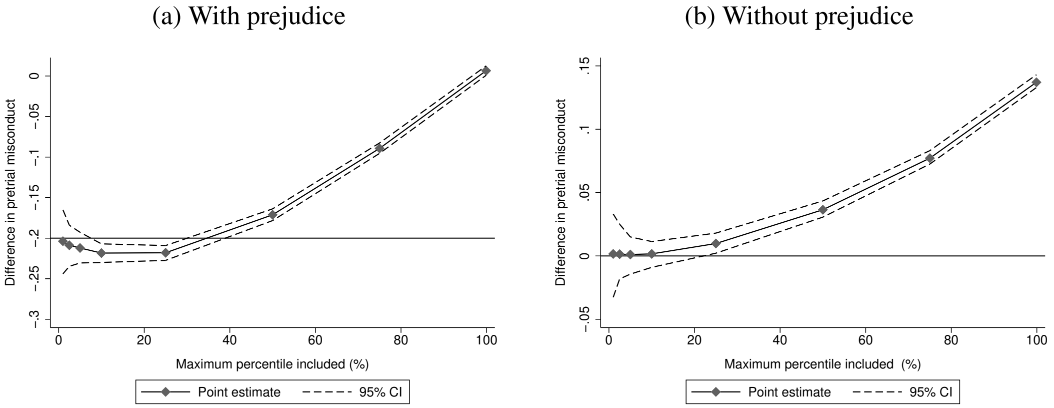
Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel, “The problem of infra-marginality in outcome tests for discrimination,” *The Annals of Applied Statistics*, 2017, *11* (3), 1193–1216.

Small, Mario L and Devah Pager, “Sociological perspectives on racial discrimination,” *Journal of Economic Perspectives*, 2020, *34* (2), 49–67.

Vytlacil, Edward, “Independence, monotonicity, and latent index models: An equivalence result,” *Econometrica*, 2002, *70* (1), 331–341.

Yang, Crystal and Will Dobbie, “Equal Protection Under Algorithms: A New Statistical and Legal Framework,” *Michigan Law Review*, 2020.

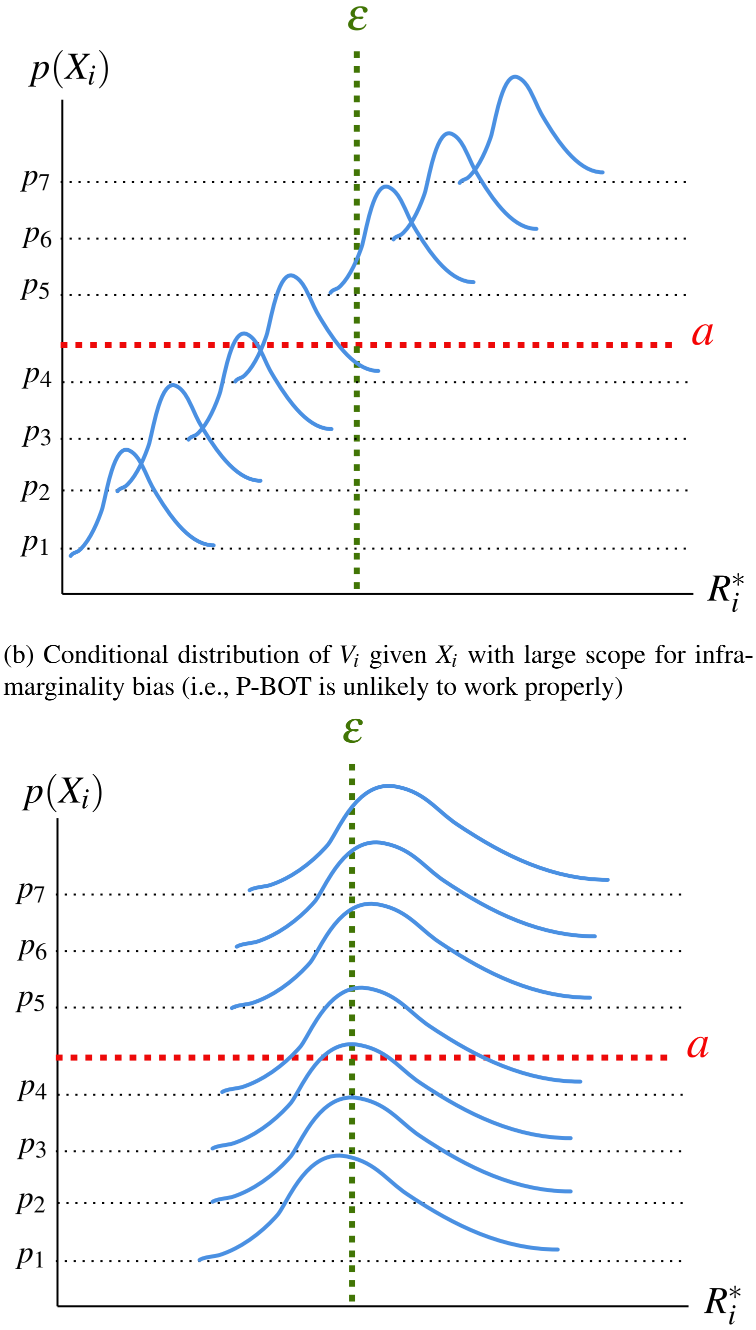
Figure 1: Prediction-Based Outcome Test: Intuition on Thought Experiment


(a) With prejudice

(b) Without prejudice

Note: This figure uses simulated data based on the model presented in Appendix B. Panel (a) considers a case with prejudice (with true difference in effective thresholds of 0.2) and the Panel (b) considers a case with no prejudice (with true difference in effective thresholds of 0). The x-axis measures the maximum percentile of R_i^* considered for computing the difference in pretrial misconduct rates between groups. That is, 100 means that the entire sample of released defendants is considered, 75 that only the 75% with lower R_i^* is considered, etc. The point estimates are the mean estimation across 200 Monte Carlo simulations with sample sizes of 250,000. Confidence intervals correspond to the 2.5 and 97.5 percentiles of the simulations. The plots include points for percentiles 1, 2.5, 5, 10, 25, 50, 75, and 99.9.

Figure 2: Prediction-Based Outcome Test: Intuition on Implementation

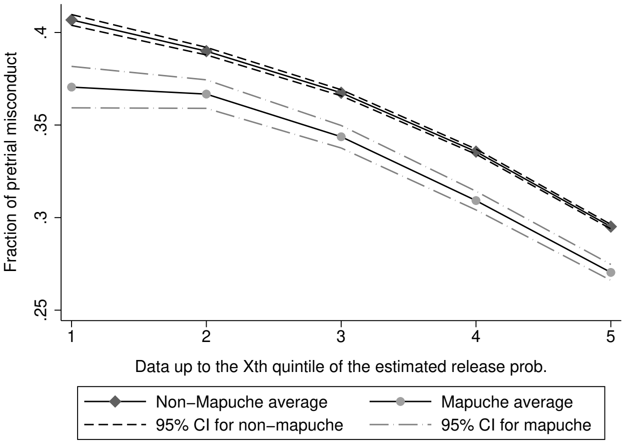


(a) Conditional distribution of V_i given X_i with limited scope for infra-marginality bias (i.e., P-BOT is likely to work properly)

(b) Conditional distribution of V_i given X_i with large scope for infra-marginality bias (i.e., P-BOT is unlikely to work properly)

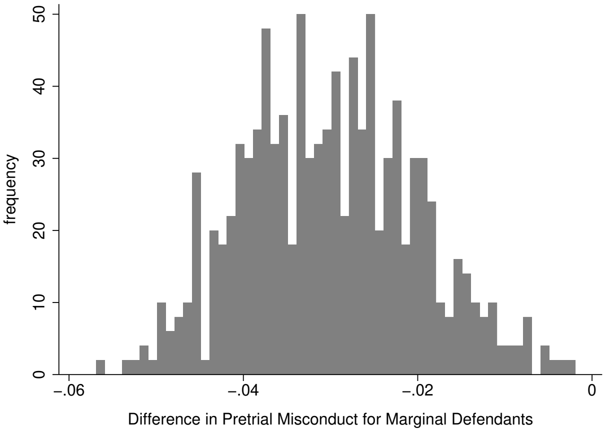
Note: This figure illustrates how the conditional variance of V_i given X_i can affect the performance of the P-BOT. In each panel, the y-axis accounts for the propensity score $p(X_i)$ and the 7 illustrated observations are sorted according the propensity score, with $p_1 < p_2 < \cdots < p_7$. The x-axis accounts for R_i^*, where each blue curve corresponds to the conditional distribution of R_i^* given the propensity score. Panel (a) displays a case where each observation has a relatively *small* conditional variance, a case where the bias in the P-BOT is likely to be limited. Panel (b) displays a case where each observation has a relatively *large* conditional variance, a case where the bias in the P-BOT is likely to be substantial.

Figure 3: Pretrial Misconduct Rates for Different Quintiles of the Predicted Release Probability



Note: This plot presents the Mapuche and non-Mapuche pretrial misconduct rates for different groups of predicted release probability quintiles (1: quintile 1; 2: quintiles 1-2; 3: quintiles 1-3; 4: quintiles 1-4; 5: full sample). Mapuche is defined as self-reported or at least one surname. Predictions are estimated using a probit model. Each plot presents the results for one of the four definitions of Mapuche. Confidence intervals are analytically calculated assuming that quintiles are given. Pretrial misconduct accounts for non-appearance in court and/or pretrial recidivism.

Figure 4: Perturbation Test



Note: This plot presents the perturbation test described in Section 3. Mapuche is defined as self-reported or at least one surname. They are produced in the following steps. First, we estimate the probit model. Then, for each released individual in the sample, we simulate 500 realizations from a standardized normal distribution to simulate R_i^* and redefine the samples of marginal individuals. Within each sample, we estimate the outcome test and plot its distribution across simulations.

Table 1: Descriptive Statistics

		Mapuche			
	Non-Mapuche	At least one surname	Two surnames	Self-Reported	Self-Reported or at least one surname
Released	0.84	0.85	0.87	0.85	0.85
Outcomes (only for released)					
Non-appearance in court	0.17	0.16	0.14	0.16	0.16
Pretrial recidivism	0.19	0.17	0.13	0.16	0.17
Pretrial misconduct	0.30	0.27	0.23	0.27	0.27
Individual Characteristics					
Male	0.88	0.89	0.91	0.92	0.89
At least one previous case	0.68	0.66	0.60	0.65	0.66
At least one previous pretrial misconduct	0.40	0.37	0.29	0.36	0.37
At least one previous conviction	0.65	0.63	0.57	0.62	0.63
No. of previous cases	4.59	4.25	3.47	4.13	4.28
Severity previous case	0.09	0.08	0.06	0.07	0.08
Severity current case	0.18	0.17	0.15	0.16	0.17
Court Characteristics					
Average severity (year/Court)	0.09	0.09	0.08	0.08	0.09
No. of cases (year/Court)	3,053	2,729	2,311	1,802	2,717
No. of judges (year/Court)	46	40	32	20	40
Observations (released)	541,742	42,987	8,455	7,992	43,952
Observations (non-released)	105,987	7,830	1,255	1,431	8,049

Note: This table presents the descriptive statistics of our estimation sample. The sample considers all arraignment hearings for adult defendants who were arrested between 2008 and 2017. We drop hearings due to legal summons and only consider types of crimes with at least a 5% probability of pretrial detention. When defendants are accused of more than one crime, we retain the information related to the most severe crime (with severity measured as the probability of pretrial detention).

Table 2: Prediction-Based Outcome Test, Using Probit to Estimate the Release Probability (Outcome: Pretrial Misconduct)

	A: Simple Version		**B: Non-Parametric**	
Max. percentile considered:	5	10	5	10
Point estimate, (a)-(b):	-0.043	-0.040	-0.030	-0.037
C.I. (95%)	[-0.065, -0.022]	[-0.055, -0.025]	[-0.056, -0.005]	[-0.056, -0.018]
(a) Mapuche expectation	0.365	0.362	0.394	0.373
(b) Non-Mapuche expectation	0.408	0.402	0.424	0.410
No. of Mapuche	1,985	3,911	1,985	3,911
No. of Non-Mapuche	27,300	54,659	27,300	54,659

Note: This table presents the results from the P-BOT using the data described in Table 1, considering two approaches to estimate the outcome equation and two criteria to determine who is the margin. Mapuche is defined as self-reported or at least one surname. Release probabilities are predicted using a probit model. The outcome is any pretrial misconduct. Panel A shows the estimates using the simple approach, considering the individuals whose estimated release probability is lower than or equal to the 5th/10th percentile. Panel B shows the estimates using the non-parametric approach. The margin of release is defined as the 1st percentile of the estimated release probability. The bandwidth is the same for both estimations (for Mapuche and non-Mapuche) and it is defined as the distance between the 1st percentile and the 5th/10th percentile of the estimated release probability. Details of the covariates included in the prediction model can be found in Appendix E. The confidence intervals are calculated using bootstrap with 500 repetitions.

Table 3: Alternative Tests for Prejudice

	Outcome test (full sample)	**IV-Outcome test (Mapuche)**	**IV-Outcome test (non-Mapuche)**
Coeff.	-0.023	0.240	0.363
Robust SE	(0.003)	(0.477)	(0.059)
Observations	699,730	50,801	647,700

Note: This table presents the results from alternative tests for prejudice using the data described in Table 1. Mapuche is defined as self-reported or at least one surname. The outcome is any pretrial misconduct. The outcome test using the full sample reports the estimated coefficient of an OLS regression of pretrial misconduct on a Mapuche indicator. Following Arnold et al. (2018), the IV-outcome test reports the coefficient of a 2SLS regression of pretrial misconduct on release, instrumenting release with the residualized leave-out mean release rate of the assigned judge. In the IV estimation, standard errors are clustered at the year/court level.

Table 4: Prediction-Based Outcome Test for Mapuche and Other Categories, Using Probit to Estimate the Release Probability (Outcome: Pretrial Misconduct)

Panel A: Income		**Panel B: Region**	
Mapuche	-0.014	Mapuche	-0.032
C.I. (95%)	[-0.039, 0.011]	C.I. (95%)	[-0.049, -0.015]
Low income	0.017	Mapuche region	-0.068
C.I. (95%)	[0.009, 0.027]	C.I. (95%)	[-0.094, -0.045]
Mapuche and low income	-0.040	Mapuche and mapuche region	-0.021
C.I. (95%)	[-0.073, -0.007]	C.I. (95%)	[-0.070, 0.028]
Pretrial misconduct expectation for:		**Pretrial misconduct expectation for:**	
Mapuche and low income	0.324	Mapuche and Mapuche region	0.283
Non-Mapuche and low income	0.378	Non-Mapuche and Mapuche region	0.336
Mapuche and high income	0.347	Mapuche and non-Mapuche region	0.373
Non-Mapuche and high income	0.360	Non-Mapuche and non-Mapuche region	0.404
Observations:		**Observations:**	
Mapuche and low income	1,773	Mapuche and Mapuche region	463
Non-Mapuche and low income	22,474	Non-Mapuche and Mapuche region	1,503
Mapuche and high income	1,385	Mapuche and non-Mapuche region	3,448
Non-Mapuche and high income	21,075	Non-Mapuche and non-Mapuche region	53,156

Note: This table presents the results of the P-BOT considering additional categories to group defendants. Mapuche is defined as self-reported or at least one surname. The outcome is any pretrial misconduct. In Panel A, we include indicators for *Mapuche* and *low income*, which is equal to one when defendants live in a municipality whose average income is below the median. In Panel B, we include indicators for *Mapuche* and *Mapuche region*, which is equal to one if the defendant is accused in a court located at the Araucanía Region, the administrative region historically associated with the Mapuche conflict. These models use the data described in Table 1. Release probabilities are predicted using a probit model. The outcome is any pretrial misconduct. We present results for the simple version of the P-BOT and considering the released individuals whose estimated release probability is lower or equal to the 10th percentile. The confidence intervals are calculated using bootstrap with 500 repetitions.

Table 5: Prediction-Based Outcome Test Controlling for Court-by-time Fixed Effects, Using Probit to Estimate the Release Probability (Outcome: Pretrial Misconduct)

Max. percentile considered:	5	10
Point estimate:	-0.022	-0.020
C.I. (95%)	[-0.044, -0.001]	[-0.035, -0.003]
No. of Mapuche	1,985	3,911
No. of Non-Mapuche	27,300	54,659

Note: This table presents the results from the P-BOT controlling by court-by-time fixed effects using the data described in Table 1, and considering two criteria to determine who is the margin. Mapuche is defined as self-reported or at least one surname. The outcome is any pretrial misconduct. Release probabilities are predicted using a probit model. The outcome is any pretrial misconduct. Raw expectations are not reported since conditional levels are not identified given the inclusion of fixed effects. We present results for the simple version of the P-BOT. The confidence intervals are calculated using bootstrap with 500 repetitions.

An Observational Implementation of the Outcome Test with an Application to Ethnic Prejudice in Pretrial Detentions

Appendix for Online Publication

A Proofs and Additional Results

Proof of Proposition I. Let PM_{i1} be pretrial misconduct if released, with $\mathbb{E}[PM_{i1}|G_i, Z_i, j(i)] = m(G_i, Z_i, j(i))$. $R_i^* = 0$ implies that $m(G_i, Z_i, j(i)) = h(G_i, Z_i, j(i))$. Also, $PM_i = PM_{i1}$ when $R_i = 1$. To simplify ther notation below, let $m_i \equiv m(G_i, Z_i, j(i))$ and $h_i = h(G_i, Z_i, j(i))$. Then:

$$\begin{aligned}
\mathbb{E}[PM_i|G_i = g, R_i^* = 0] &= \mathbb{E}[PM_{i1}|G_i = g, R_i^* = 0], \\
&= \mathbb{E}[PM_{i1}|G_i = g, m_i = h_i, R_i^* = 0], \\
&= \mathbb{E}[\mathbb{E}[PM_{i1}|G_i, Z_i, j(i), m_i = h_i, R_i^* = 0]|G_i = g, m_i = h_i, R_i^* = 0], \\
&= \mathbb{E}[m_i|G_i = g, m_i = h_i, R_i^* = 0], \\
&= \mathbb{E}[h_i|G_i = g, R_i^* = 0], \\
&= \overline{h}(g). \quad \square
\end{aligned}$$

Proof of Proposition II. Given A1 and A2, we can define $\Lambda(X_i) = (d(X_i) - \mu(X_i))/\sigma$. Let Θ_ζ be the CDF of ζ_i, with $\Theta_\zeta(\Lambda(X_i)) = p(X_i)$. Then:

$$\begin{aligned}
p(X_i) \leq a &\Leftrightarrow \Theta_\zeta\left(\frac{d(X_i) - \mu(X_i)}{\sigma}\right) \leq a \\
&\Leftrightarrow \frac{d(X_i) - \mu(X_i)}{\sigma} \leq \Theta_\zeta^{-1}(a), \\
&\Leftrightarrow d(X_i) - W_i \leq \sigma\left(\Theta_\zeta^{-1}(a) - \zeta_i\right) \\
&\Leftrightarrow R_i^* \leq \sigma\left(\Theta_\zeta^{-1}(a) - \zeta_i\right). \qquad (10)
\end{aligned}$$

Then, defining $\varepsilon(a) = \sigma\left(\Theta_\zeta^{-1}(a) - \overline{\zeta}\right)$ and $u_i = -\sigma(\zeta_i - \overline{\zeta})$ completes the proof. □

Proposition II with heteroskedastic conditional variance. Consider a modified A2 where $W_i =$

$\mu(X_i)+\sigma(X_i)\zeta_i$, with $\mathbb{E}[\sigma(X_i)]=\overline{\sigma}<\infty$. Then

$$\begin{aligned}
p(X_i)\leq a &\Leftrightarrow \Theta_\zeta\left(\frac{d(X_i)-\mu(X_i)}{\sigma(X_i)}\right)\leq a\\
&\Leftrightarrow \frac{d(X_i)-\mu(X_i)}{\sigma(X_i)}\leq \Theta_\zeta^{-1}(a),\\
&\Leftrightarrow d(X_i)-W_i\leq \sigma(X_i)\left(\Theta_\zeta^{-1}(a)-\zeta_i\right)\\
&\Leftrightarrow R_i^*\leq \overline{\sigma}\left(\Theta_\zeta^{-1}(a)-\zeta_i\right)+(\sigma(X_i)-\overline{\sigma})\left(\Theta_\zeta^{-1}(a)-\zeta_i\right),\\
&\Leftrightarrow R_i^*\leq \overline{\sigma}\left(\Theta_\zeta^{-1}(a)-\overline{\zeta}\right)+\frac{(\sigma(X_i)-\overline{\sigma})}{\overline{\sigma}}\overline{\sigma}\left(\Theta_\zeta^{-1}(a)-\overline{\zeta}\right)-\frac{\sigma(X_i)}{\overline{\sigma}}\overline{\sigma}\left(\zeta_i-\overline{\zeta}\right),\\
&\Leftrightarrow R_i^*\leq \varepsilon+\frac{(\sigma(X_i)-\overline{\sigma})}{\overline{\sigma}}\varepsilon+\frac{\sigma(X_i)}{\overline{\sigma}}u_i, \qquad (11)
\end{aligned}$$

where ε and u_i are defined as above by replacing σ with $\overline{\sigma}$. Note that when $\sigma(X_i)=\sigma$, for all X_i, the expression reduces to the one in Proposition II. Also note that $\mathbb{E}\left[\frac{(\sigma(X_i)-\overline{\sigma})}{\overline{\sigma}}\right]=0$ and $\mathbb{E}\left[\frac{\sigma(X_i)}{\overline{\sigma}}u_i\right]=\frac{\mathbb{E}[\sigma(X_i)]}{\overline{\sigma}}\mathbb{E}[u_i]=0$. So, under this pattern of heteroskedasticity, it is still the case that the classification error has mean zero. That is, $\mathbb{E}[\Gamma(u_i,X_i)]=0$, where $\Gamma(u_i,X_i)=\frac{(\sigma(X_i)-\overline{\sigma})}{\overline{\sigma}}\varepsilon+\frac{\sigma(X_i)}{\overline{\sigma}}u_i$.

This classification error, however, now depends on X_i. Note that, if $\sigma(X_i)<\overline{\sigma}$:

$$p(X_i)\leq a \quad\Leftrightarrow\quad R_i^*<\varepsilon+\frac{\sigma(X_i)}{\overline{\sigma}}u_i. \qquad (12)$$

Then, for realizations of X_i with *low* conditional variance (relative to $\overline{\sigma}$), the classification error is smaller. By contrast, for realizations of X_i with *large* conditional variance (relative to $\overline{\sigma}$), the two sources of classification error “slack” the margin definition, giving more scope for the inclusion of misclassified observations. This suggests that the P-BOT is expected to have less scope for classification errors when the heteroskedasticity holds a monotonicity property, where observations with *low* propensity scores also have lower variances, and the observations with higher variances have propensity scores *far above* from a.

Why A1 does not imply A2. Under A1, the selection rule takes the form $1\{d(X_i)\geq W_i\}$. Assuming the CDF of W_i given X_i, $\Theta_{W|X}$, is strictly increasing, a well-known result is that $1\{d(X_i)\geq W_i\}$ can be written as $1\left\{\widetilde{d}(X_i) \geq \widetilde{W}_i\right\}$, where $\widetilde{d}(X_i) = \Theta_{W|X}(d(X_i))$, which coincides with the propensity score, and $\widetilde{W}_i = \Theta_{W|X}(W_i)$ is uniformly distributed between 0 and 1 given X_i. Noting that uniform variables satisfy A2 by construction, this suggests that, *in the alternative representation*, A2 is without loss of generality given A1.

While this assessment is correct, A2 in the manuscript restricts the *structural representation* based on $d(X_i)$ and W_i, and not the alternative one based on $\widetilde{d}(X_i)$ and $\widetilde{W}_i$, and this distinction is without loss of generality for the purposes of the identification argument. The reason is that the ε-distance argument requires the relationship to hold under small perturbations of the decision rule, which breaks the isomorphism with the alternative representation if $\Theta_{W|X}$ is unrestricted because ε must be a constant and not dependent from X_i.

To see why, consider the equation $1\{d(X_i) \leq W_i + \varepsilon\}$. In this case, $\Theta_{W|X}(W_i + \varepsilon)$ is not necessarily uniformly distributed and, even in cases where ε is *very small*, the object cannot necessarily be approximated by $\Theta_{W|X}(W_i)$. For example, consider a Taylor approximation around $\varepsilon = 0$. In this case, $\Theta_{W|X}(W_i + \varepsilon) \approx \Theta_{W|X}(W_i) + \theta_{W|X}(W_i)\varepsilon$, where $\theta_{W|X}$ is the conditional density. Then, $\widetilde{W}_i$ no longer satisfies A2 by construction unless we restrict the conditional density. Said differently, the perturbed *structural* model does not necessarily delivers a perturbed *alternative representation* of the type $1\left\{\widetilde{d}(X_i) \leq \widetilde{W}_i + \varepsilon\right\}$, unless, for example, we restrict $\theta_{W|X}$ to be constant.

Note, however, that the more general statement of restricting $\Theta_{W|X}$ and/or $\theta_{W|X}$ can be framed as finding restrictions on the conditional distribution of W_i given X_i, which is exactly what A2 does. While there may be alternative assumptions on $\Theta_{W|X}$ and/or $\theta_{W|X}$ that make the argument work by exploiting the alternative representation, we proceed with A2 which provides explicit and transparent restrictions.

B Monte Carlo Simulations

This section presents the results of different Monte Carlo simulations. The objective of this exercise is twofold. First, it shows that when both assumptions (A1 and A2) are met, the presence of correlation between the observed and unobserved variables does not affect identification. Second, it shows that the P-BOT converges to a less precise version of the full sample outcome test (Knowles et al., 2001) as the conditional variance increases. Intuitively, when the observable component has no predictive power, the P-BOT is essentially an OLS regression of a random subsample of released defendants.

To these purposes, we simulate the model using the following equations:

$$\begin{aligned}
R_i &= 1\{(\alpha_X + \delta_X)X_i + (\alpha_W + \delta_W)W_i + \delta_G G_i \leq \beta_0 - \beta_G G_i\}, \\
PM_i &= 1\{\alpha_X X_i + \alpha_W W_i + \eta_i > \gamma_0\} \cdot R_i, \\
\eta_i &= \delta_X X_i + \delta_W W_i + \delta_G G_i + v_i, \\
W_i &= (X_i + G_i)\beta_W + \zeta_i,
\end{aligned}$$

where X_i and W_i are (scalar) defendant i's characteristics other than group, G_i is defendant i's group, $v_i \sim \mathcal{N}(0, \sigma_v^2)$, and $\zeta_i \sim \mathcal{N}(0, \sigma_\zeta^2)$. We assume that the judge observes both X_i and W_i, but the econometrician only observes X_i. β_0 is a leniency measure, β_G measures prejudice in the form of taste-based discrimination, and δ_G measures (accurate) statistical discrimination. We assume that judges are homogeneous, i.e., β_0 and β_G are not function on $j(i)$. Note that both A1 and A2 hold in this setting. The parameter of interest in the outcome test – in this simplified example – is β_G.

The two set of simulations have the following random structure:

$$\begin{pmatrix} X_i \\ G_i^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0.15 \\ 0.15 & 1 \end{pmatrix}\right),$$

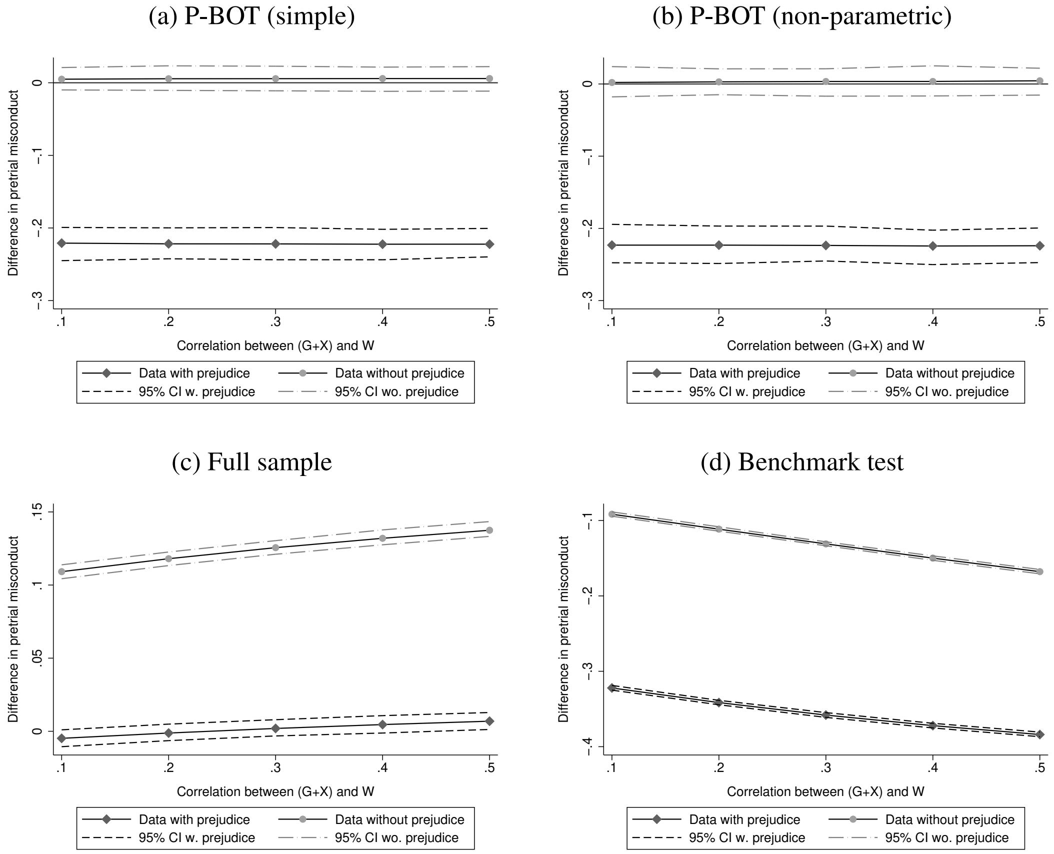
where G_i^* is a latent variable such that $G_i = 1\{G_i^* \geq 0\}$. We simulate the model using $\alpha_X = \alpha_W = \delta_X = \delta_W = \delta_G = 0.1$, $\beta_0 = 0.4$, $\gamma_0 = 0.1$, and $\sigma_v = 0.1$. We provide simulations for a model *without prejudice* (i.e., $\beta_G = 0$) and *with prejudice*, with $\beta_G = 0.2$.

The first set of simulations sets $\sigma_\zeta = 1$ and tests the performance of the P-BOT for different values of β_W, which measures the correlation between the observed and unobserved variables. In particular, we consider $\beta_W \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For estimating the P-BOT, we compute conditional predicted release probabilities and run OLS regressions of PM_i on R_i on samples of released defendants with predicted probabilities up to the 5th percentile. We also perform the non-parametric estimation evaluated at the 1st percentile. The predicted probability is estimated using a probit model of R_i on X_i and G_i. We compare the P-BOT to other models that are likely to be affected by the magnitude of β_W. Specifically, we run OLS regressions of PM_i on G_i using the complete sample of released defendants (outcome test with full sample). We also estimate probit regressions of R_i on G_i and X_i and report the coefficient on G_i (misspecified benchmark test).

Figure B.I shows the results. The point estimates are the mean estimate across 200 Monte Carlo simulations with sample sizes of 250,000, and the confidence intervals are formed using the 2.5 and 97.5 percentiles of the estimated models. The figure shows that the P-BOT correctly identifies β_G regardless of the value of β_W. This is consistent with the discussion in the main text. Importantly, these correlation values are large enough to make both the model subject to substantial inframarginality bias (and, therefore, strongly affecting the performance of the outcome test using the full sample) and to omitted variable bias in the release equation (and, therefore, strongly affecting the performance of the benchmark test).

The second set of simulations sets $\beta_w = 0.5$ and tests the performance of the P-BOT for dif-

Figure B.I: Tests' Performance for Different Values of β_W



(a) P-BOT (simple)

(b) P-BOT (non-parametric)

(c) Full sample

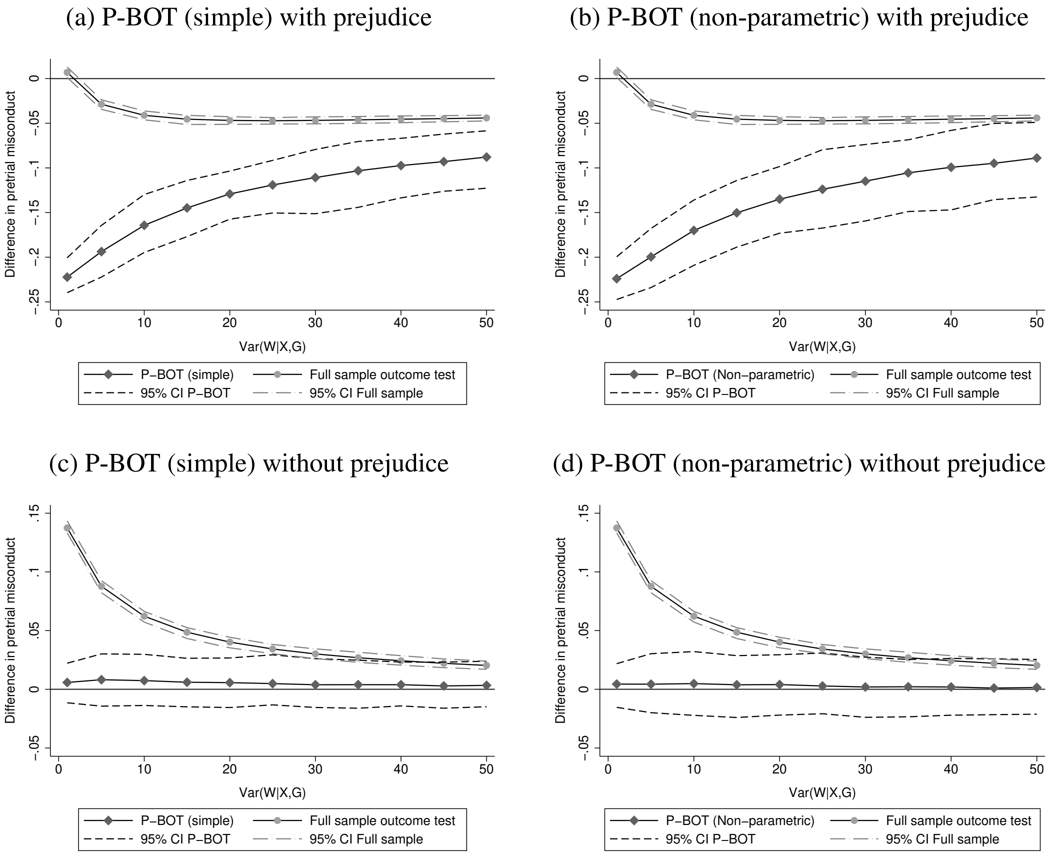
(d) Benchmark test

Notes: These figures plot the performances of P-BOT and alternative tests to test discrimination for different levels of correlation between observables and unobservables. The P-BOT is implemented using both approaches explained in Section 3 of the paper using the 5th percentile of the release probability as the threshold to define marginal defendants. *Full sample* is the outcome test considering the full sample of released defendants. *Benchmark test* reports the G_i coefficient of an OLS regression of R_i on X_i and G_i.

ferent values of σ_ζ, which measures the conditional variance of the unobserved component. In particular, we consider $\sigma_\zeta^2 \in \{1,5,10,15,20,25,30,35,40,45,50\}$. To assess the magnitude of these variances, recall that $\sigma_X = 1$ and $\sigma_{R^*} = 0.1$.

Figure B.II shows the results. As before, the point estimates are the mean estimate across 200 Monte Carlo simulations with sample sizes of 250,000, and the confidence intervals are formed using the 2.5 and 97.5 percentiles of the estimated models. The figures show that when σ_ζ increases, the P-BOT converges to a less precise version of Knowles et al. (2001) test. The reason is that large conditional variances decrease the performance of the prediction model. Then, as the relative predictive power of X_i and G_i decreases, the P-BOT ends essentially selecting a random set of the sample of released individuals.

Figure B.II: P-BOT versus Full Sample Outcome Test, for Different Values of σ_{ζ}^{2}



(a) P-BOT (simple) with prejudice

(b) P-BOT (non-parametric) with prejudice

(c) P-BOT (simple) without prejudice

(d) P-BOT (non-parametric) without prejudice

Notes: These figures plot the performances of the P-BOT and the full sample outcome test for different levels of σ_{ζ}^{2}. The P-BOT is implemented using both approaches explained in Section 4 of the paper using the 5th percentile of the release probability as the threshold to define marginal defendants. *Full sample* is the outcome test considering the full sample of released defendants.

C Data Appendix

This appendix gives a more detailed description of the data, the sample restrictions, and the construction of the variables.

C.1 Sources

We merge three different sources of data to build our database.

PDO administrative records. We use administrative records from the Public Defender Office (PDO, see http://www.dpp.cl/). The PDO is a centralized public service under the oversight of the Ministry of Justice that provides criminal defense services to all individuals accused of or charged with a crime who lack an attorney. The centralized nature of the PDO ensures that the administrative records contain information for all the cases handled only by the PDO or in tandem with a private attorney (as opposed to by only private attorneys), which covers more than 95% of the universe of criminal cases of Chile. The unit of analysis is a criminal case and contains: defendants characteristics (ID, name, gender, self-reported ethnicity, and place of residence, among other characteristics) and case characteristics (case ID, court, public attorney assigned, initial and end dates, different categories for type of crime, pretrial detention status and length, and outcome of the case, among other administrative characteristics). We consider cases whose arraignment hearings occurred between 2008 and 2017.

Registry of judges. In addition, we have access to detention judges and their assigned cases, for hearings that occurred between 2008 and 2017. We merge this registry with the administrative records using the cases' IDs. We do not observe other characteristics of the judges in addition to their names and IDs. This data was shared by the Department of Studies of the Chilean Supreme Court (https://www.pjud.cl/corte-suprema).

Mapuche surnames. The registry of Mapuche surnames was provided by the Mapuche Data

Project (http://mapuchedataproject.cl/). The Mapuche Data Project is an interdisciplinary project that seeks to identify, digitalize, compile, process, and harmonize quantitative information of the Mapuche people for research and policy purposes. The surnames registry, one of the several datasets publicly available in their website, contains 8,627 different Mapuche surnames. The identification is based on the works of Amigo and Bustos (2008) and Painemal (2011). Since we observe names and surnames in the PDO records, we can directly identify defendants with Mapuche surnames.

C.2 Estimation sample

The initial sample contains $3,571,230$ cases and covers all the cases recorded by the PDO whose arraignment hearing occurred between 2008 and 2017. To create our estimation sample, we make the following adjustments.

Basic data cleaning. Due to potential miscoding, we drop observations where the initial date of the case is later than the end date, and observations where the length of pretrial detention is larger than the length of the case. After these adjustments, the sample size reduces to $3,559,019$ (i.e, we drop $12,211$ cases).

Sample restrictions. We then make the following sample restrictions:

- We exclude hearings due to legal summons ($1,233,909$ observations). We do this because the information set of the judges is likely to be different.
- We drop juvenile defendants ($254,243$ observations). We do this because the juvenile criminal system works differently, so the mandated selection rule and the preventive measures differ between systems (see Cortés et al., 2020 for details).
- We drop cases where the defendant hires a private attorney as his exclusive defender ($103,092$ observations). We do this because we do not observe the result of the arraignment hearing (and what happens after in the prosecution) in these cases.

- We drop cases whose length is larger than two years ($55,495$ observations).
- For defendants that are accused of more than one crime in a given case and, therefore, the records provide multiple observations, we consider the most severe crime (see below the severity definition). In this step we drop $193,720$ observations. To be clear, in this step we do not drop defendants, but only cases. We do this to have only at most one case/defendants pair per day of arraignment hearing.
- We drop cases where the detention judge is missing ($67,449$ observations).
- We drop types of crime whose likelihood of pretrial detention is less than 5% ($945,747$ observations). We do this because we want to study judges' decisions in cases where pretrial detention is a plausible outcome.
- We drop cases handled by judges that see less than 10 cases in the whole time-period ($2,845$ observations). We also only consider cases whose public attorney defended at least 10 cases, but we do not drop any data because of this restriction.
- We drop defendants from ethnic groups different than Mapuche ($2,789$ observations).

After all these adjustments the sample size is $699,730$. That matches the numbers of Table 1.

C.3 Variables

Many of the variables used in our empirical application are directly contained in the administrative records. Here we describe how we construct the other variables.

- *Mapuche*: we build four indicators of Mapuche combining self-reporting and surnames information. See Section 4 for details.
- *Severity*: we proxy crime severity by computing the share of cases within the type of crime that use pretrial detention.

- *Criminal record*: we can track all arrests of a given defendant using their IDs. Then, the variables *Previous prosecution*, *Number of previous prosecutions*, *Previous pretrial misconduct*, *Previous conviction*, and *Severity of previous prosecution* are constructed by looking at the characteristics of the cases associated to the defendant's ID that were initiated before the current one. For individuals with no previous prosecutions, these variables are set to zero. For building these variables, we can track cases up to 2005.

- *Pretrial misconduct*: pretrial misconduct is an indicator variable that takes value 1 if the defendant does not return to a scheduled hearing and/or is engaged in pretrial recidivism. Non-appearance in court is recorded in the administrative data. Pretrial recidivism is built by looking at arrests associated to the same defendant's ID whose initial date is between the initial and end dates of the current prosecution.

- *Attorney quality and judge leniency*: as in Dobbie et al. (2018), we use the residualized (against court-by-time fixed effects) leave-out mean release rate.

- *Year of prosecution fixed effects*: we consider the initial date to set the fixed effects.

D Empirical Diagnostics

In this appendix, we provide details of the suggestive test for A1 and the covariate stability exercise of our empirical application.

A1. Recall that A1 says that there are functions d and g such that $1\{f(X_i,V_i) \geq 0\} = 1\{d(X_i) - g(V_i) \geq 0\}$. This implies that the *direction* in which X_i affects the likelihood of being released is not affected by the value of V_i. One way to assess this assumption is to check whether the coefficients of a regression of R_i on X_i are stable (in terms of sign) when considering subsamples with (probably) different unobservables. Likewise, recall that, through the lens of the model, A1 implies monotonicity on observables in the expected risk equation. Then, a similar exercise can be done with the coefficients of a regression of PM_i on X_i among different subsamples of released defendants with (probably) different unobservables, with the caveat that PM_i is presumably selected on unobservables. These tests are similar to the monotonicity tests performed by Arnold et al. (2018) and Bald et al. (2019).

Tables D.I and D.II show the results using R_i and PM_i as dependent variables, respectively. Each cell reports the estimated coefficient of the regressor specified in the column, using the sample specified in the first column. Each row represents a different estimation. The first row reports the coefficients using the whole sample, and then rows are paired by mutually exclusive sample categories that are (probably) characterized by different unobservables. For example, row 2 shows results for the Mapuche subsample, while row 3 shows results for the non-Mapuche subsample. Then, rows 4 and 5 split the sample by gender, and so on. Results support the monotonicity assumption. In all but two cases (i.e., 97.5% of cases) the sign of the coefficient is consistent across samples. Moreover, the magnitudes are also similar. This suggests that the direction of the effect of observables is unlikely to be affected by the unobserved variables.

Table D.I: Testing for Monotonicity in Observables (Dep. Variable: Release Status)

Estimation sample	Previous case	Previous pretrial misconduct	Previous conviction	Severity previous case	Severity current case
All	-0.029 [-0.034, -0.024]	-0.027 [-0.029, -0.025]	-0.015 [-0.020, -0.010]	-0.101 [-0.107, -0.095]	-0.757 [-0.761, -0.753]
Mapuche	-0.028 [-0.046, -0.009]	-0.025 [-0.032, -0.017]	-0.014 [-0.031, 0.003]	-0.081 [-0.103, -0.060]	-0.745 [-0.759, -0.731]
Non-Mapuche	-0.029 [-0.034, -0.023]	-0.027 [-0.029, -0.025]	-0.015 [-0.020, -0.010]	-0.103 [-0.109, -0.096]	-0.757 [-0.762, -0.753]
Male	-0.030 [-0.035, -0.024]	-0.029 [-0.032, -0.027]	-0.015 [-0.020, -0.009]	-0.090 [-0.097, -0.084]	-0.766 [-0.771, -0.762]
Female	-0.013 [-0.026, -0.000]	-0.005 [-0.011, 0.001]	-0.022 [-0.034, -0.010]	-0.218 [-0.237, -0.198]	-0.688 [-0.701, -0.675]
Low income	-0.028 [-0.036, -0.020]	-0.024 [-0.027, -0.021]	-0.016 [-0.024, -0.008]	-0.099 [-0.108, -0.089]	-0.764 [-0.771, -0.757]
High income	-0.029 [-0.036, -0.023]	-0.029 [-0.032, -0.026]	-0.015 [-0.021, -0.009]	-0.103 [-0.111, -0.095]	-0.752 [-0.758, -0.747]
Low judge leniency	-0.028 [-0.035, -0.021]	-0.028 [-0.031, -0.025]	-0.017 [-0.024, -0.010]	-0.108 [-0.117, -0.100]	-0.782 [-0.788, -0.776]
High judge leniency	-0.029 [-0.036, -0.022]	-0.025 [-0.028, -0.022]	-0.013 [-0.020, -0.006]	-0.094 [-0.102, -0.086]	-0.732 [-0.738, -0.726]
Low attorney quality	-0.027 [-0.035, -0.020]	-0.027 [-0.031, -0.024]	-0.018 [-0.025, -0.011]	-0.109 [-0.118, -0.100]	-0.824 [-0.831, -0.818]
High attorney quality	-0.030 [-0.037, -0.023]	-0.026 [-0.029, -0.023]	-0.012 [-0.019, -0.006]	-0.093 [-0.101, -0.085]	-0.690 [-0.696, -0.684]
Small Court (No. of cases)	-0.018 [-0.026, -0.011]	-0.025 [-0.028, -0.022]	-0.018 [-0.025, -0.011]	-0.112 [-0.121, -0.103]	-0.794 [-0.800, -0.787]
Big Court (No. of cases)	-0.039 [-0.046, -0.032]	-0.028 [-0.031, -0.025]	-0.013 [-0.020, -0.007]	-0.094 [-0.102, -0.086]	-0.723 [-0.729, -0.717]
Small Court (No. of judges)	-0.020 [-0.027, -0.012]	-0.027 [-0.030, -0.024]	-0.016 [-0.023, -0.009]	-0.114 [-0.123, -0.105]	-0.800 [-0.806, -0.793]
Big Court (No. of judges)	-0.037 [-0.044, -0.030]	-0.026 [-0.029, -0.023]	-0.015 [-0.022, -0.008]	-0.091 [-0.099, -0.083]	-0.716 [-0.722, -0.710]
Low severity court	-0.032 [-0.038, -0.025]	-0.023 [-0.025, -0.020]	-0.011 [-0.017, -0.005]	-0.079 [-0.087, -0.072]	-0.637 [-0.642, -0.631]
High severity court	-0.025 [-0.033, -0.017]	-0.030 [-0.034, -0.027]	-0.020 [-0.028, -0.013]	-0.124 [-0.133, -0.114]	-0.880 [-0.887, -0.874]

Note: This table presents the results of the test for monotonicity in observables. Each reported value is the marginal effect of the variable of the column on the probability of release, estimated using a different sample in each row. The continuous variables were discretized using the respective median as the threshold. The values in parenthesis are 95% confident intervals, estimated using bootstrap with 500 repetitions.

Estimation sample stability To assess which covariates play a more important role in determining the predicted propensity score ranking – and, therefore, the P-BOT estimation sample –, we conduct a stability exercise where we estimate restricted propensity scores that exclude covariates, redefine estimation samples, and assess their differences with respect to the baseline estimation sample. Concretely, we compute (i) the share of the baseline estimation sample also considered for estimation under the restricted estimation when considering the bottom 5% of the propensity score distribution, (ii) the share of the baseline estimation sample also considered for estimation

Table D.II: Testing for Monotonicity in Observables (Dep. Variable: Pretrial Misconduct)

Estimation sample	Previous case	Previous pretrial misconduct	Previous conviction	Severity previous case	Severity current case
All	0.074	0.090	0.035	0.032	0.036
	[0.067, 0.081]	[0.087, 0.093]	[0.028, 0.042]	[0.022, 0.042]	[0.027, 0.045]
Mapuche	0.057	0.082	0.038	0.045	0.047
	[0.031, 0.082]	[0.071, 0.093]	[0.014, 0.063]	[0.009, 0.081]	[0.014, 0.080]
Non-Mapuche	0.075	0.091	0.035	0.030	0.035
	[0.068, 0.083]	[0.088, 0.094]	[0.028, 0.042]	[0.020, 0.041]	[0.025, 0.044]
Male	0.076	0.092	0.034	0.036	0.042
	[0.069, 0.084]	[0.089, 0.095]	[0.027, 0.041]	[0.025, 0.046]	[0.032, 0.052]
Female	0.064	0.076	0.041	-0.023	-0.012
	[0.044, 0.084]	[0.066, 0.085]	[0.022, 0.060]	[-0.060, 0.013]	[-0.039, 0.016]
Low income	0.069	0.083	0.038	0.032	0.077
	[0.058, 0.081]	[0.079, 0.088]	[0.027, 0.048]	[0.016, 0.047]	[0.063, 0.091]
High income	0.075	0.093	0.034	0.033	0.002
	[0.066, 0.084]	[0.089, 0.097]	[0.026, 0.043]	[0.020, 0.046]	[-0.010, 0.014]
Low judge	0.065	0.086	0.044	0.033	0.034
leniency	[0.054, 0.075]	[0.082, 0.090]	[0.034, 0.053]	[0.019, 0.048]	[0.021, 0.047]
High judge	0.083	0.094	0.027	0.030	0.037
leniency	[0.073, 0.093]	[0.090, 0.098]	[0.017, 0.036]	[0.016, 0.044]	[0.024, 0.050]
Low attorney	0.071	0.094	0.041	0.042	0.031
quality	[0.061, 0.081]	[0.089, 0.098]	[0.032, 0.051]	[0.028, 0.057]	[0.018, 0.045]
High attorney	0.077	0.087	0.028	0.021	0.039
quality	[0.067, 0.087]	[0.082, 0.091]	[0.019, 0.038]	[0.007, 0.035]	[0.027, 0.052]
Small Court	0.063	0.087	0.035	0.036	0.094
(No. of cases)	[0.053, 0.073]	[0.083, 0.091]	[0.025, 0.045]	[0.021, 0.051]	[0.081, 0.107]
Big Court	0.084	0.091	0.036	0.029	-0.012
(No. of cases)	[0.074, 0.093]	[0.087, 0.095]	[0.027, 0.046]	[0.015, 0.043]	[-0.025, 0.001]
Small Court	0.076	0.090	0.029	0.038	0.061
(No. of judges)	[0.065, 0.086]	[0.086, 0.095]	[0.019, 0.039]	[0.023, 0.053]	[0.047, 0.074]
Big Court	0.073	0.086	0.041	0.026	0.011
(No. of judges)	[0.064, 0.083]	[0.082, 0.090]	[0.032, 0.050]	[0.013, 0.040]	[-0.001, 0.024]
Low severity	0.074	0.084	0.037	0.033	0.048
court	[0.065, 0.084]	[0.080, 0.088]	[0.028, 0.047]	[0.019, 0.046]	[0.036, 0.061]
High severity	0.073	0.095	0.034	0.029	0.020
court	[0.062, 0.083]	[0.091, 0.100]	[0.024, 0.044]	[0.014, 0.044]	[0.006, 0.034]

Note: This table presents the results of the test for monotonicity in observables. Each reported value is the marginal effect of the variable of the column on pretrial misconduct, estimated using a different sample of released defendants in each row. The continuous variables were discretized using the respective median as the threshold. The values in parenthesis are 95% confident intervals, estimated using bootstrap with 500 repetitions.

under the restricted estimation when considering the bottom 10% of the propensity score distribution, and (iii) the Spearman's-ρ rank correlation statistic for the predicted propensity score across released defendants. We do this for all individual covariates of the propensity score model. We also consider the simultaneous exclusion of all variables that describe, first, criminal records (previous case, previous pretrial misconduct, previous conviction, number of previous cases, and severity of previous case) and, second, court characteristics (average severity at the court by year, number of cases at the court by year, and number of judges at the court by year).

Table D.III presents the results. It can be seen that for almost all excluded covariates, the estimation samples are very stable. In other words, there is a substantial overlap between baseline and restricted estimation samples, which is also consistent with the large rank correlations. The main exception is the severity of the current case, whose exclusion induces noisier rankings and, therefore, estimation samples that differ from the baseline case. The noise induced by the exclusion of the severity of the current case is quantitatively more important that either simultaneously excluding all variables related to criminal records or court characteristics, an insight that suggests that the imputed type of crime of the current prosecution is an important variable for bail judges that is not fully informed by criminal records, demographics, or court characteristics.

Table D.III: Estimation Sample Stability

Excluded predictor(s)	Share (%) in baseline estimation sample (p5)	Share (%) in baseline estimation sample (p10)	Rank correlation Spearman's-ρ
Mapuche	99.76	99.75	1.000
Male	99.83	99.81	1.000
Previous case	98.93	99.17	0.999
Previous pretrial misconduct	95.43	96.69	0.996
Previous conviction	99.40	99.68	1.000
No of previous cases	89.87	87.34	0.986
Severity previous case	93.57	94.85	0.996
Severity current case	24.81	42.74	0.697
Average severity (year/court)	85.35	85.51	0.957
No of cases (year/court)	98.19	98.33	0.999
No of judges (year/court)	96.21	96.68	0.997
Judge leniency	95.61	95.87	0.995
Attorney quality	91.47	91.92	0.982
Criminal records	71.80	74.48	0.854
Court characteristics	83.35	83.47	0.948

Note: This table presents the results of the estimation sample stability exercise. We exclude covariates from the probit model for the propensity score and compute (i) the share of the baseline estimation sample also considered for estimation under the restricted estimation when considering the bottom 5% of the propensity score distribution, (ii) the share of the baseline estimation sample also considered for estimation under the restricted estimation when considering the bottom 10% of the propensity score distribution, and (iii) the Spearman's-ρ rank correlation statistic for the predicted propensity score across released defendants. Each row excludes the covariate specified in the first column. When excluding judge leniency or attorney quality, we also exclude the squared term. The row *Criminal records* means that we simultaneously exclude all variables related to criminal records (previous case, previous pretrial misconduct, previous conviction, number of previous cases, and severity of previous case). The row *Court characteristics* means that we simultaneously exclude all variables related to court characteristics (average severity at the court by year, number of cases at the court by year, and number of judges at the court by year).

E Prediction Models

Table E.I: Determinants of Release Probability Using a Probit Model (Marginal Effects)

	At least one Surname	Two Surnames	Self-Reported	Self-Reported or at least one surname
Mapuche	-0.004	-0.008	-0.012	-0.004
	(0.002)	(0.004)	(0.004)	(0.002)
Male	0.003	0.002	0.003	0.003
	(0.002)	(0.002)	(0.002)	(0.002)
Previous prosecution	-0.028	-0.028	-0.028	-0.028
	(0.003)	(0.003)	(0.003)	(0.003)
Previous pretrial misconduct	-0.027	-0.027	-0.028	-0.027
	(0.001)	(0.001)	(0.001)	(0.001)
Previous conviction	-0.015	-0.015	-0.015	-0.015
	(0.003)	(0.003)	(0.003)	(0.003)
No. of previous Prosecution	-0.007	-0.007	-0.007	-0.007
	(0.000)	(0.000)	(0.000)	(0.000)
Severity (previous prosecution)	-0.101	-0.103	-0.103	-0.101
	(0.004)	(0.004)	(0.004)	(0.004)
Severity (current prosecution)	-0.761	-0.760	-0.761	-0.761
	(0.008)	(0.008)	(0.008)	(0.008)
Average severity of the cases (court/year)	-1.021	-1.028	-1.030	-1.020
	(0.032)	(0.033)	(0.033)	(0.032)
No. of cases per court/year	-0.0000029	-0.0000030	-0.0000029	-0.0000029
	(0.0000007)	(0.0000007)	(0.0000007)	(0.0000007)
No. of judges per court/year	0.00030	0.00030	0.00030	0.00030
	(0.00004)	(0.00004)	(0.00004)	(0.00004)
Judge leniency	0.541	0.541	0.537	0.540
	(0.028)	(0.028)	(0.028)	(0.028)
Judge leniency squared	0.787	0.701	0.738	0.778
	(0.363)	(0.368)	(0.370)	(0.363)
Attorney quality	0.531	0.531	0.528	0.531
	(0.027)	(0.027)	(0.027)	(0.027)
Attorney quality squared	0.607	0.596	0.585	0.606
	(0.118)	(0.117)	(0.118)	(0.118)
Year of Prosecution fixed effects	YES	YES	YES	YES
Court fixed effects	NO	NO	NO	NO
No. of Mapuche	50,817	9,710	9,423	52,001
No. of Non-Mapuche	647,729	647,729	647,729	647,729
Pseudo-R-squared	0.23	0.23	0.23	0.23
Correctly classified (0.5 prob as threshold)	0.85	0.85	0.85	0.85
Correctly classified (prediction: Non-Released)	0.60	0.59	0.59	0.60
Correctly classified (prediction: Released)	0.87	0.87	0.87	0.87

Note: This table presents the marginal effects of a probit model for the determinants of the release status using the data described in Table 1. The standard errors (in parenthesis) are clustered at the year/court level. The four models correspond to the four definitions of Mapuche considered in this paper.

Table E.II: Determinants of Release Probability Using a Linear Probability Model

	At least one Surname	Two Surnames	Self-Reported	Self-Reported or at least one surname
Mapuche	-0.002	-0.007	-0.006	-0.002
	(0.002)	(0.003)	(0.004)	(0.002)
Male	-0.002	-0.003	-0.002	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)
Previous prosecution	-0.002	-0.002	-0.001	-0.002
	(0.002)	(0.003)	(0.003)	(0.002)
Previous pretrial misconduct	-0.024	-0.024	-0.025	-0.024
	(0.001)	(0.001)	(0.001)	(0.001)
Previous conviction	-0.014	-0.014	-0.014	-0.014
	(0.002)	(0.002)	(0.002)	(0.002)
No. of previous prosecution	-0.009	-0.009	-0.009	-0.009
	(0.000)	(0.000)	(0.000)	(0.000)
Severity (previous prosecution)	-0.143	-0.146	-0.145	-0.143
	(0.005)	(0.005)	(0.005)	(0.005)
Severity (current prosecution)	-1.017	-1.013	-1.014	-1.017
	(0.011)	(0.011)	(0.011)	(0.011)
Average severity of the cases (court/year)	-1.060	-1.066	-1.069	-1.061
	(0.028)	(0.029)	(0.029)	(0.028)
No. of cases per court/year	-0.0000048	-0.0000050	-0.0000051	-0.0000048
	(0.0000018)	(0.0000018)	(0.0000018)	(0.0000018)
No. of judges per court/year	-0.00013	-0.00012	-0.00013	-0.00013
	(0.00007)	(0.00008)	(0.00008)	(0.00007)
Judge leniency	0.558	0.559	0.553	0.557
	(0.026)	(0.027)	(0.027)	(0.026)
Judge leniency squared	0.550	0.478	0.527	0.546
	(0.340)	(0.352)	(0.352)	(0.339)
Attorney quality	0.528	0.528	0.525	0.527
	(0.020)	(0.020)	(0.020)	(0.020)
Attorney quality squared	-0.077	-0.087	-0.096	-0.079
	(0.088)	(0.089)	(0.089)	(0.088)
Year of Prosecution fixed effects	YES	YES	YES	YES
Court fixed effects	YES	YES	YES	YES
No. of Mapuche	50,817	9,710	9,423	52,001
No. of Non-Mapuche	647,729	647,729	647,729	647,729
R-squared	0.22	0.21	0.21	0.22
Correctly classified (0.5 prob as threshold)	0.85	0.85	0.85	0.85
Correctly classified (prediction: Non-Released)	0.65	0.65	0.64	0.65
Correctly classified (prediction: Released)	0.86	0.86	0.86	0.86

Note: This table presents the point estimates of a linear probability model for the determinants of the release status using the data described in Table 1. The standard errors (in parenthesis) are clustered at the year/court level. The four models correspond to the four definitions of Mapuche considered in this paper.

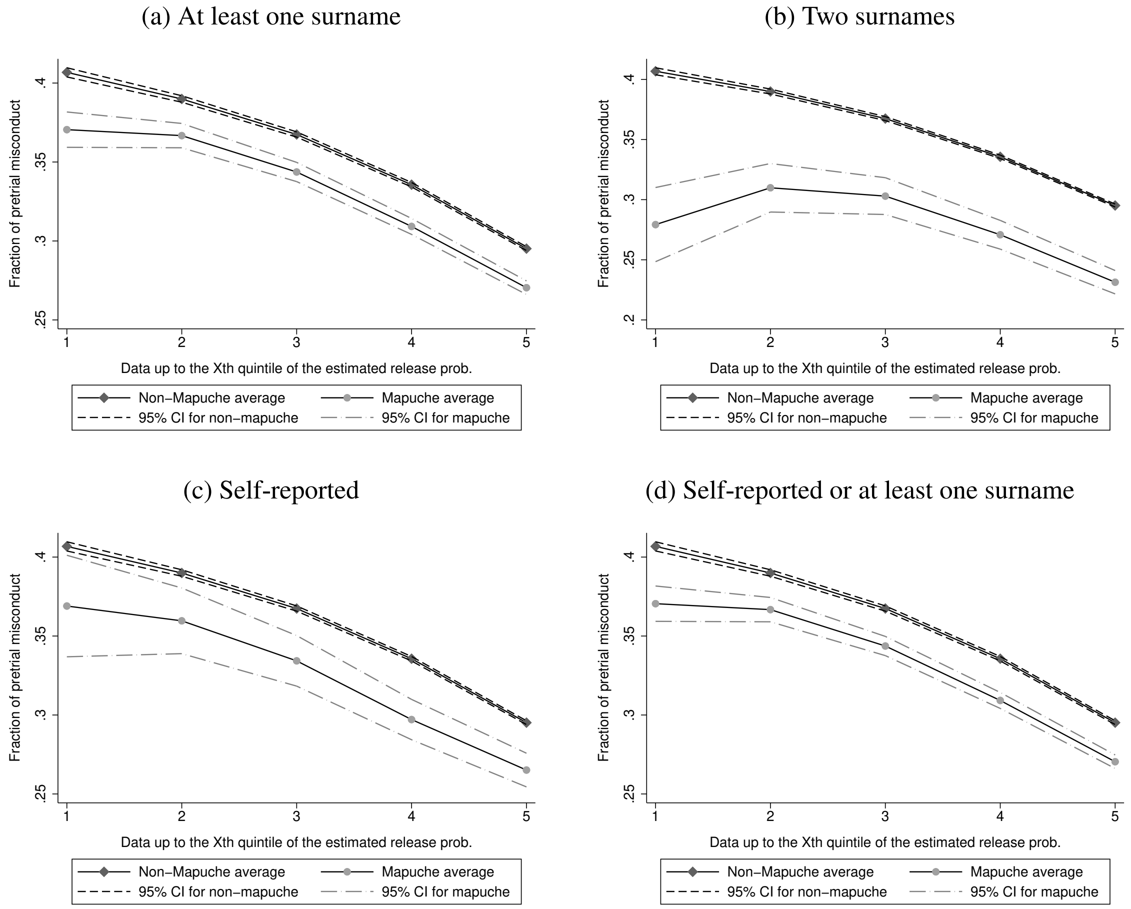
Table E.III: Determinants of Release Probability Using a Heteroskedastic Probit Model

	At least one Surname	Two Surnames	Self-Reported	Self-Reported or at least one surname
Mapuche	-0.042	-0.113	-0.143	-0.043
	(0.014)	(0.029)	(0.031)	(0.014)
Male	0.082	0.073	0.073	0.081
	(0.027)	(0.028)	(0.028)	(0.027)
Previous prosecution	-0.177	-0.185	-0.175	-0.179
	(0.060)	(0.062)	(0.063)	(0.060)
Previous pretrial misconduct	-0.169	-0.167	-0.173	-0.169
	(0.022)	(0.023)	(0.023)	(0.022)
Previous conviction	-0.317	-0.308	-0.320	-0.315
	(0.055)	(0.056)	(0.056)	(0.055)
No. of previous Prosecution	-0.044	-0.044	-0.044	-0.044
	(0.002)	(0.002)	(0.002)	(0.002)
Severity (previous prosecution)	-0.617	-0.625	-0.620	-0.619
	(0.055)	(0.057)	(0.057)	(0.055)
Severity (current prosecution)	-5.581	-5.531	-5.541	-5.588
	(0.222)	(0.227)	(0.228)	(0.222)
Average severity of the cases (court/year)	-7.852	-7.851	-7.850	-7.843
	(0.309)	(0.315)	(0.318)	(0.311)
No. of cases per court/year	-0.000034	-0.000035	-0.000033	-0.000034
	(0.000013)	(0.000013)	(0.000013)	(0.000013)
No. of judges per court/year	0.003290	0.003154	0.003119	0.003293
	(0.000780)	(0.000781)	(0.000782)	(0.000782)
Judge leniency	6.161	6.211	6.135	6.174
	(0.526)	(0.533)	(0.530)	(0.527)
Judge leniency squared	13.778	13.391	13.402	13.775
	(3.415)	(3.461)	(3.458)	(3.416)
Attorney quality	5.475	5.462	5.433	5.482
	(0.358)	(0.367)	(0.364)	(0.358)
Attorney quality squared	6.425	6.356	6.257	6.430
	(0.894)	(0.904)	(0.908)	(0.894)
Conditional variance:				
Male	0.057	0.053	0.052	0.056
	(0.015)	(0.015)	(0.015)	(0.015)
Previous prosecution	0.021	0.018	0.023	0.020
	(0.031)	(0.033)	(0.033)	(0.031)
Previous pretrial misconduct	0.019	0.018	0.016	0.019
	(0.013)	(0.014)	(0.014)	(0.013)
Previous conviction	-0.132	-0.129	-0.134	-0.131
	(0.028)	(0.029)	(0.030)	(0.028)
No. of previous Prosecution	0.017	0.017	0.017	0.017
	(0.001)	(0.001)	(0.001)	(0.001)
Severity (previous prosecution)	0.305	0.306	0.309	0.303
	(0.032)	(0.033)	(0.034)	(0.032)
Severity (current prosecution)	1.355	1.358	1.364	1.356
	(0.054)	(0.056)	(0.056)	(0.054)
Average severity of the cases (court/year)	0.565	0.585	0.595	0.575
	(0.180)	(0.185)	(0.186)	(0.181)
No. of cases per court/year	-0.000002	-0.000002	-0.000001	-0.000002
	(0.000006)	(0.000007)	(0.000007)	(0.000007)
No. of judges per court/year	0.000581	0.000525	0.000520	0.000579
	(0.000331)	(0.000337)	(0.000336)	(0.000331)
Judge leniency	1.157	1.218	1.194	1.168
	(0.249)	(0.253)	(0.252)	(0.248)
Attorney quality	0.773	0.792	0.798	0.777
	(0.113)	(0.113)	(0.113)	(0.113)

Note: This table presents the results of the point estimates of a probit model for the determinants of the release status using the data described in Table 1 and the point estimates for the relationship between covariates and the variance of the unobservable component (modeled as $\exp(X\beta)$). The standard errors (in parenthesis) are clustered at the year/court level. The four models correspond to the four definitions of Mapuche considered in this paper.

F Main results using alternative Mapuche definitions

Figure F.I: Pretrial Misconduct Rates for Different Quintiles of the Predicted Release Probability



(a) At least one surname

(b) Two surnames

(c) Self-reported

(d) Self-reported or at least one surname

Note: These plots present the Mapuche and non-Mapuche pretrial misconduct rates for different groups of predicted release probability quintiles (1: quintile 1; 2: quintiles 1-2; 3: quintiles 1-3; 4: quintiles 1-4; 5: full sample). Predictions are estimated using a probit model. Each plot presents the results for one of the four definitions of Mapuche. Confidence intervals are analytically calculated assuming that quintiles are given. Pretrial misconduct accounts for non-appearance in court and/or pretrial recidivism.

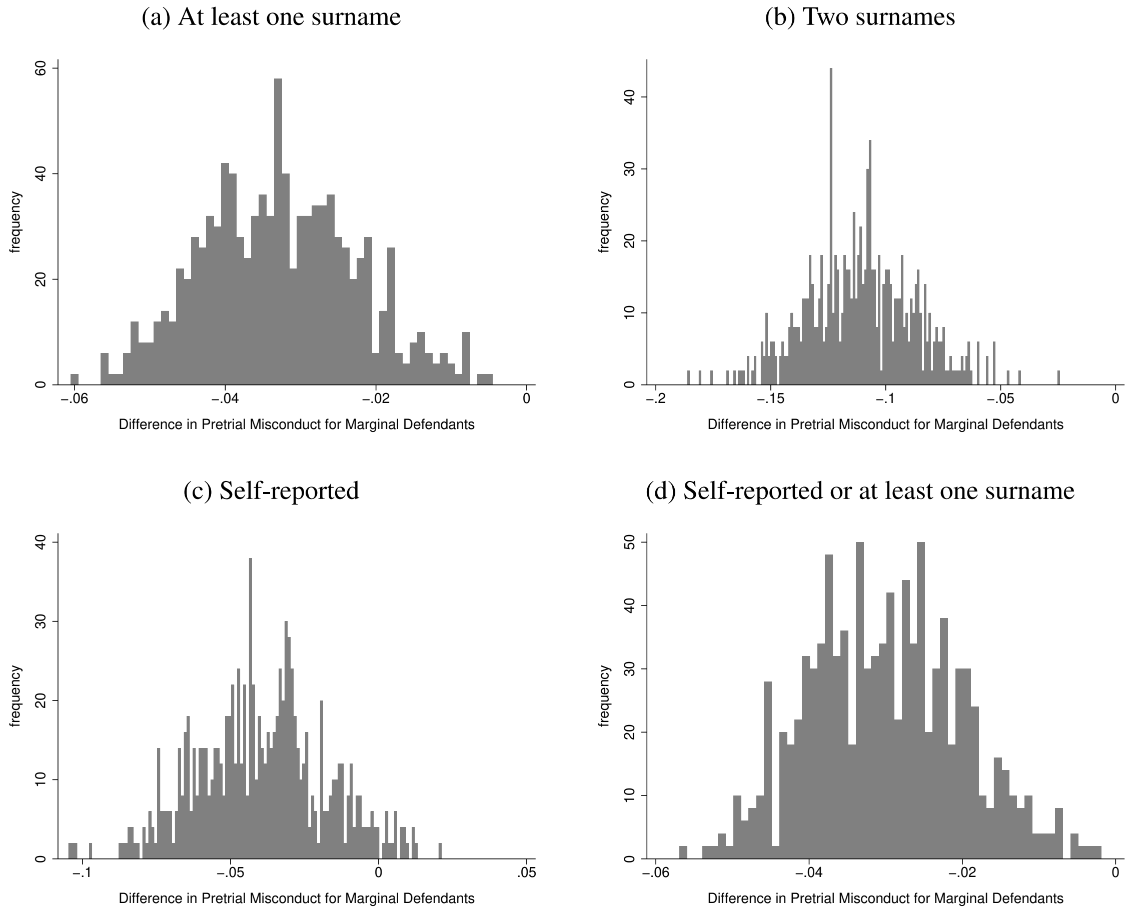
Table F.I: Prediction-Based Outcome Test, Using Probit to Estimate the Release Probability (Outcome: Pretrial Misconduct)

Data up to 5th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.049	-0.144	-0.070	-0.043
C.I. (95%)	[-0.070, -0.027]	[-0.194, -0.090]	[-0.120, -0.021]	[-0.065, -0.022]
(a) Mapuche expectation	0.359	0.264	0.338	0.365
(b) Non-Mapuche expectation	0.407	0.408	0.408	0.408
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.033	-0.143	-0.072	-0.030
C.I. (95%)	[-0.059, -0.008]	[-0.205, -0.077]	[-0.125, -0.006]	[-0.056, -0.005]
(a) Mapuche expectation	0.391	0.281	0.352	0.394
(b) Non-Mapuche expectation	0.424	0.424	0.424	0.424
No. of Mapuche ($\leq$ 5th pctl.)	1,915	269	317	1,985
No. of Non-Mapuche ($\leq$ 5th pctl.)	27,322	27,241	27,171	27,300

Data up to 10th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.043	-0.159	-0.037	-0.040
C.I. (95%)	[-0.058, -0.028]	[-0.195, -0.124]	[-0.078, 0.004]	[-0.055, -0.025]
(a) Mapuche expectation	0.359	0.244	0.365	0.362
(b) Non-Mapuche expectation	0.402	0.403	0.402	0.402
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.041	-0.150	-0.064	-0.037
C.I. (95%)	[-0.059, -0.022]	[-0.197, -0.105]	[-0.107, -0.016]	[-0.056, -0.018]
(a) Mapuche expectation	0.369	0.260	0.346	0.373
(b) Non-Mapuche expectation	0.410	0.410	0.410	0.410
No. of Mapuche ($\leq$ 10th pctl.)	3,784	496	638	3,911
No. of Non-Mapuche ($\leq$ 10th pctl.)	54,689	54,524	54,336	54,659

Note: This table presents the results from the P-BOT using the data described in Table 1, considering two approaches to estimate the outcome equation and two criteria to determine who is the margin. Release probabilities are predicted using a probit model. The outcome is any pretrial misconduct. Panel A shows the estimates using the simple approach, considering the individuals whose estimated release probability is lower than or equal to the 5th/10th percentile. Panel B shows the estimates using the non-parametric approach. The margin of release is defined as the 1st percentile of the estimated release probability. The bandwidth is the same for both estimations (for Mapuche and non-Mapuche) and it is defined as the distance between the 1st percentile and the 5th/10th percentile of the estimated release probability. Details of the covariates included in the prediction model can be found in Appendix E. The confidence intervals are calculated using bootstrap with 500 repetitions.

Figure F.II: Perturbation Test


(a) At least one surname

(b) Two surnames

(c) Self-reported

(d) Self-reported or at least one surname

Note: These plots present the results of the perturbation test described in Section 3. They are produced in the following steps. First, we estimate the probit model. Then, for each released individual in the sample, we simulate 500 realizations from a standardized normal distribution to simulate R_i^* and redefine the samples of marginal individuals. Within each sample, we estimate the outcome test and plot its distribution across simulations.

Table F.II: Alternative Tests for Prejudice

	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Outcome test (full sample):				
Coeff.	-0.023	-0.059	-0.026	-0.023
Robust SE	(0.003)	(0.006)	(0.010)	(0.003)
Observations	698,546	657,439	657,152	699,730
IV-Outcome test (without controls):				
Mapuche coeff.	0.417	-0.153	12.688	0.240
Mapuche robust SE	(0.527)	(0.288)	(141.0)	(0.477)
Non-Mapuche coeff.	0.363	0.363	0.363	0.363
Non-Mapuche robust SE	(0.059)	(0.059)	(0.059)	(0.059)
No. of Mapuche	49,569	8,055	7,853	50,801
No. of non-Mapuche	647,700	647,700	647,700	647,700
IV-Outcome test (with controls):				
Mapuche coeff.	0.867	-0.280	1.982	0.414
Mapuche robust SE	(1.096)	(0.453)	(3.479)	(0.783)
Non-Mapuche coeff.	0.374	0.374	0.374	0.374
Non-Mapuche robust SE	(0.059)	(0.059)	(0.059)	(0.059)
No. of Mapuche	49,569	8,055	7,853	50,801
No. of non-Mapuche	647,700	647,700	647,700	647,700

Note: This table presents the results from alternative tests for prejudice using the data described in Table 1. The outcome is any pretrial misconduct. The outcome test using the full sample reports the estimated coefficient of an OLS regression of pretrial misconduct on a Mapuche indicator. Following Arnold et al. (2018), the IV-outcome test reports the coefficient of a 2SLS regression of pretrial misconduct on release, instrumenting release with the residualized leave-out mean release rate of the assigned judge. In the IV estimation, standard errors are clustered at the year/court level. The version with controls include the covariates used in the randomization test.

G Robustness Checks

Table G.I: Prediction-Based Outcome Test, Using OLS to Estimate the Release Probability (Outcome: Pretrial Misconduct)

Data up to 5th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.037	-0.132	-0.058	-0.031
C.I. (95%)	[-0.056, -0.015]	[-0.182, -0.081]	[-0.102, -0.000]	[-0.051, -0.010]
(a) Mapuche expectation	0.347	0.253	0.326	0.353
(b) Non-Mapuche expectation	0.384	0.385	0.384	0.384
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.033	-0.152	-0.071	-0.029
C.I. (95%)	[-0.060, -0.006]	[-0.210, -0.089]	[-0.127, -0.007]	[-0.056, -0.003]
(a) Mapuche expectation	0.374	0.255	0.336	0.378
(b) Non-Mapuche expectation	0.406	0.406	0.407	0.406
No. of Mapuche (≤ 5th pctl.)	1,981	297	340	2,044
No. of Non-Mapuche (≤ 5th pctl.)	27,256	27,213	27,147	27,241
Data up to 10th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.046	-0.157	-0.032	-0.042
C.I. (95%)	[-0.060, -0.031]	[-0.190, -0.121]	[-0.071, 0.006]	[-0.057, -0.028]
(a) Mapuche expectation	0.328	0.218	0.343	0.332
(b) Non-Mapuche expectation	0.375	0.375	0.375	0.374
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.039	-0.147	-0.057	-0.035
C.I. (95%)	[-0.058, -0.021]	[-0.191, -0.106]	[-0.098, -0.008]	[-0.054, -0.016]
(a) Mapuche expectation	0.346	0.239	0.329	0.350
(b) Non-Mapuche expectation	0.385	0.386	0.385	0.385
No. of Mapuche (≤ 10th pctl.)	3,908	577	627	4,031
No. of Non-Mapuche (≤ 10th pctl.)	54,565	54,443	54,347	54,539

Note: This table presents the results from the P-BOT using the data described in Table 1, considering two approaches to estimate the outcome equation and two criteria to determine who is the margin. Release probabilities are predicted using a linear probability model. The outcome is any pretrial misconduct. Panel A shows the estimates using a simple difference between the Mapuche and non-Mapuche averages in pretrial misconduct, only considering the individuals whose estimated release probability is lower than or equal to the 5th/10th percentile. Panel B shows the estimates using a non-parametric local estimation for the conditional expectation of pretrial misconduct at the margin of release, for Mapuche and non-Mapuche defendants. The point estimate is calculated by subtracting these two estimations. The margin of release is defined as the 1st percentile of the estimated release probability. The bandwidth is the same for both estimations (for Mapuche and non-Mapuche) and it is defined as the distance between the 1st percentile and the 5th/10th percentile of the estimated release probability. Details of the covariates included in the prediction model can be found in Appendix E. The confidence intervals are calculated using bootstrap with 500 repetitions.

Table G.II: Prediction-Based Outcome Test, Using OLS to Estimate the Release Probability and Lasso to Select Predictors (Outcome: Pretrial Misconduct)

Data up to 5th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.052	-0.143	-0.085	-0.044
C.I. (95%)	[-0.073, -0.028]	[-0.193, -0.084]	[-0.125, -0.030]	[-0.067, -0.022]
(a) Mapuche expectation	0.367	0.275	0.335	0.375
(b) Non-Mapuche expectation	0.418	0.419	0.420	0.419
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.040	-0.139	-0.068	-0.035
C.I. (95%)	[-0.068, -0.012]	[-0.204, -0.077]	[-0.125, -0.005]	[-0.063, -0.005]
(a) Mapuche expectation	0.393	0.293	0.365	0.398
(b) Non-Mapuche expectation	0.432	0.432	0.433	0.433
No. of Mapuche (≤ 5th pctl.)	2,065	316	394	2,137
No. of Non-Mapuche (≤ 5th pctl.)	27,172	27,194	27,093	27,148

Data up to 10th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.046	-0.139	-0.048	-0.045
C.I. (95%)	[-0.061, -0.028]	[-0.178, -0.096]	[-0.089, -0.005]	[-0.060, -0.026]
(a) Mapuche expectation	0.372	0.280	0.372	0.374
(b) Non-Mapuche expectation	0.419	0.419	0.420	0.419
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.045	-0.139	-0.064	-0.041
C.I. (95%)	[-0.068, -0.024]	[-0.185, -0.097]	[-0.110, -0.015]	[-0.064, -0.019]
(a) Mapuche expectation	0.379	0.285	0.360	0.383
(b) Non-Mapuche expectation	0.424	0.424	0.425	0.425
No. of Mapuche (≤ 10th pctl.)	3,912	574	667	4,032
No. of Non-Mapuche (≤ 10th pctl.)	54,561	54,446	54,307	54,538

Notes: This table presents the results from the P-BOT with the release probabilities predicted using a linear model. The predictors were selected using Lasso. The original set of covariates included 1,568 variables to be chosen: the predictors considered in Table G.I, their squared terms, their interactions, and judge fixed effects. When Mapuche is defined as *at least one surname*, lasso selected 880 predictors, 878 when it is defined as *two surnames*, 871 when it is defined as *self-reported*, and 877 when it is defined as *self-reported or at least one surname*. In all these models, 85% of the cases are correctly classified by the prediction model. Specifically, those who are predicted as released and detained are correctly classified in 87% and 64% of the cases, respectively. The other characteristics of this table replicates Table G.I. Panel A shows the estimates using a simple difference between the Mapuche and non-Mapuche averages in pretrial misconduct, only considering the individuals whose estimated release probability is lower than or equal to the 5th/10th percentile. Panel B shows the estimates using a non-parametric local estimation for the conditional expectation of pretrial misconduct at the margin of release, for Mapuche and non-Mapuche defendants. The point estimate is calculated by subtracting these two estimations. The margin of release is defined as the 1st percentile of the estimated release probability. The bandwidth is the same for both estimations (for Mapuche and non-Mapuche) and it is defined as the distance between the 1st percentile and the 5th/10th percentile of the estimated release probability. The confidence intervals are calculated using bootstrap with 500 repetitions.

Table G.III: Prediction-Based Outcome Test, Using Heteroskedastic Probit to Estimate the Release Probability (Outcome: Pretrial Misconduct)

Data up to 5th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.049	-0.148	-0.081	-0.043
C.I. (95%)	[-0.070, -0.025]	[-0.191, -0.095]	[-0.126, -0.031]	[-0.066, -0.021]
(a) Mapuche expectation	0.357	0.258	0.325	0.362
(b) Non-Mapuche expectation	0.405	0.406	0.406	0.405
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.033	-0.144	-0.073	-0.029
C.I. (95%)	[-0.062, -0.007]	[-0.200, -0.084]	[-0.132, -0.017]	[-0.057, -0.003]
(a) Mapuche expectation	0.389	0.278	0.350	0.393
(b) Non-Mapuche expectation	0.422	0.422	0.423	0.422
No. of Mapuche ($\leq$ 5th pctl.)	1,965	299	354	2,036
No. of Non-Mapuche ($\leq$ 5th pctl.)	27,272	27,211	27,133	27,249
Data up to 10th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.040	-0.154	-0.028	-0.038
C.I. (95%)	[-0.057, -0.025]	[-0.196, -0.123]	[-0.064, 0.008]	[-0.053, -0.022]
(a) Mapuche expectation	0.357	0.244	0.369	0.360
(b) Non-Mapuche expectation	0.398	0.398	0.398	0.398
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.041	-0.151	-0.059	-0.037
C.I. (95%)	[-0.058, -0.023]	[-0.196, -0.108]	[-0.099, -0.015]	[-0.054, -0.019]
(a) Mapuche expectation	0.364	0.254	0.347	0.369
(b) Non-Mapuche expectation	0.405	0.406	0.406	0.405
No. of Mapuche ($\leq$ 10th pctl.)	3,841	528	658	3,971
No. of Non-Mapuche ($\leq$ 10th pctl.)	54,632	54,492	54,316	54,599

Note: This table presents the results from the P-BOT using the data described in Table 1, considering two approaches to estimate the outcome equation and two criteria to determine who is the margin. Release probabilities are predicted using a heteroscedastic probit model. The outcome is any pretrial misconduct. Panel A shows the estimates using a simple difference between the Mapuche and non-Mapuche averages in pretrial misconduct, only considering the individuals whose estimated release probability is lower than or equal to the 5th/10th percentile. Panel B shows the estimates using a non-parametric local estimation for the conditional expectation of pretrial misconduct at the margin of release, for Mapuche and non-Mapuche defendants. The point estimate is calculated by subtracting these two estimations. The margin of release is defined as the 1st percentile of the estimated release probability. The bandwidth is the same for both estimations (for Mapuche and non-Mapuche) and it is defined as the distance between the 1st percentile and the 5th/10th percentile of the estimated release probability. Details of the covariates included in the prediction model can be found in Appendix E. The confidence intervals are calculated using bootstrap with 500 repetitions.

Table G.IV: Prediction-Based Outcome Test, Using Probit to Estimate the Release Probability (Outcome: Non-Appearance in Court)

Data up to 5th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.021	-0.057	-0.034	-0.019
C.I. (95%)	[-0.038, -0.005]	[-0.091, -0.017]	[-0.073, 0.009]	[-0.035, -0.003]
(a) Mapuche expectation	0.155	0.119	0.142	0.157
(b) Non-Mapuche expectation	0.175	0.176	0.176	0.175
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.010	-0.061	-0.022	-0.010
C.I. (95%)	[-0.029, 0.009]	[-0.102, -0.013]	[-0.067, 0.021]	[-0.030, 0.008]
(a) Mapuche expectation	0.164	0.114	0.152	0.165
(b) Non-Mapuche expectation	0.175	0.175	0.175	0.175
No. of Mapuche ($\leq$ 5th pctl.)	1,915	269	317	1,985
No. of Non-Mapuche ($\leq$ 5th pctl.)	27,322	27,241	27,171	27,300
Data up to 10th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.028	-0.081	-0.028	-0.027
C.I. (95%)	[-0.041, -0.015]	[-0.109, -0.056]	[-0.059, 0.001]	[-0.039, -0.014]
(a) Mapuche expectation	0.165	0.113	0.165	0.166
(b) Non-Mapuche expectation	0.193	0.193	0.193	0.193
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.020	-0.068	-0.029	-0.019
C.I. (95%)	[-0.033, -0.006]	[-0.099, -0.032]	[-0.065, 0.007]	[-0.032, -0.005]
(a) Mapuche expectation	0.160	0.112	0.152	0.162
(b) Non-Mapuche expectation	0.180	0.180	0.180	0.180
No. of Mapuche ($\leq$ 10th pctl.)	3,784	496	638	3,911
No. of Non-Mapuche ($\leq$ 10th pctl.)	54,689	54,524	54,336	54,659

Note: This table presents the results from the P-BOT using the data described in Table 1, considering two approaches to estimate the outcome equation and two criteria to determine who is the margin. Release probabilities are predicted using a probit model. The outcome is non-appearance in court. Panel A shows the estimates using a simple difference between the Mapuche and non-Mapuche averages in non-appearance in court, only considering the individuals whose estimated release probability is lower than or equal to the 5th/10th percentile. Panel B shows the estimates using a non-parametric local estimation for the conditional expectation of non-appearance in court at the margin of release, for Mapuche and non-Mapuche defendants. The point estimate is calculated by subtracting these two estimations. The margin of release is defined as the 1st percentile of the estimated release probability. The bandwidth is the same for both estimations (for Mapuche and non-Mapuche) and it is defined as the distance between the 1st percentile and the 5th/10th percentile of the estimated release probability. Details of the covariates included in the prediction model can be found in Appendix E. The confidence intervals are calculated using bootstrap with 500 repetitions.

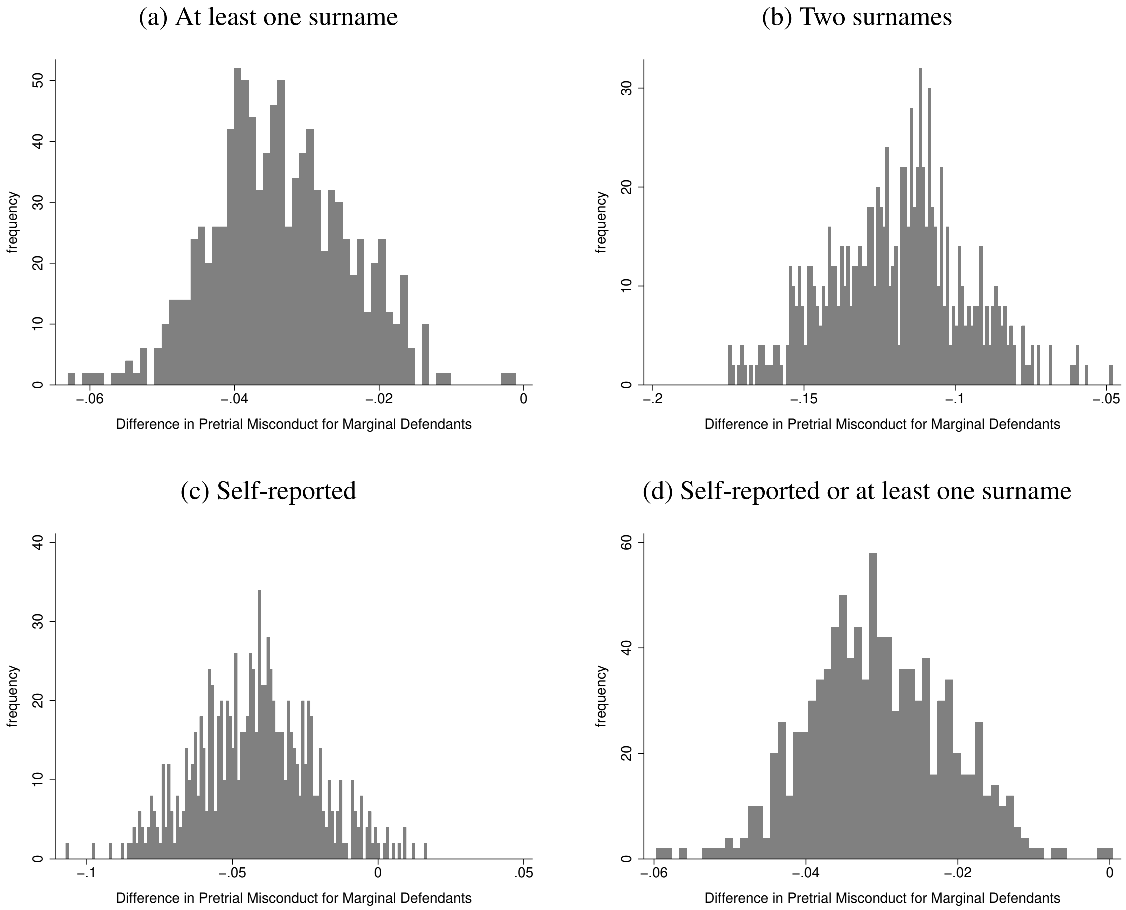
Table G.V: Prediction-Based Outcome Test, Using Probit to Estimate the Release Probability (Outcome: Pretrial Recidivism)

Data up to 5th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.043	-0.121	-0.049	-0.038
C.I. (95%)	[-0.065, -0.019]	[-0.169, -0.071]	[-0.098, 0.000]	[-0.061, -0.015]
(a) Mapuche expectation	0.286	0.208	0.281	0.291
(b) Non-Mapuche expectation	0.329	0.329	0.329	0.329
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.031	-0.119	-0.056	-0.028
C.I. (95%)	[-0.057, -0.003]	[-0.178, -0.058]	[-0.112, 0.006]	[-0.053, -0.001]
(a) Mapuche expectation	0.317	0.228	0.292	0.320
(b) Non-Mapuche expectation	0.348	0.348	0.348	0.348
No. of Mapuche (≤ 5th pctl.)	1,915	269	317	1,985
No. of Non-Mapuche (≤ 5th pctl.)	27,322	27,241	27,171	27,300

Data up to 10th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.033	-0.131	-0.019	-0.031
C.I. (95%)	[-0.047, -0.018]	[-0.164, -0.100]	[-0.055, 0.014]	[-0.046, -0.016]
(a) Mapuche expectation	0.281	0.183	0.295	0.283
(b) Non-Mapuche expectation	0.314	0.314	0.314	0.314
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.036	-0.125	-0.048	-0.033
C.I. (95%)	[-0.055, -0.015]	[-0.168, -0.080]	[-0.092, -0.002]	[-0.052, -0.012]
(a) Mapuche expectation	0.294	0.205	0.282	0.297
(b) Non-Mapuche expectation	0.330	0.330	0.330	0.330
No. of Mapuche (≤ 10th pctl.)	3,784	496	638	3,911
No. of Non-Mapuche (≤ 10th pctl.)	54,689	54,524	54,336	54,659

Note: This table presents the results from the P-BOT using the data described in Table 1, considering two approaches to estimate the outcome equation and two criteria to determine who is the margin. Release probabilities are predicted using a probit model. The outcome is pretrial recidivism. Panel A shows the estimates using a simple difference between the Mapuche and non-Mapuche averages in pretrial recidivism, only considering the individuals whose estimated release probability is lower than or equal to the 5th/10th percentile. Panel B shows the estimates using a non-parametric local estimation for the conditional expectation of pretrial recidivism at the margin of release, for Mapuche and non-Mapuche defendants. The point estimate is calculated by subtracting these two estimations. The margin of release is defined as the 1st percentile of the estimated release probability. The bandwidth is the same for both estimations (for Mapuche and non-Mapuche) and it is defined as the distance between the 1st percentile and the 5th/10th percentile of the estimated release probability. Details of the covariates included in the prediction model can be found in Appendix E. The confidence intervals are calculated using bootstrap with 500 repetitions.

Figure G.I: Perturbation Test with Heteroskedastic Probit Model



Note: These plots present the results of the perturbation test described in Section 3. They are produced in the following steps. First, we estimate the heteroskedastic probit model. Then, for each released individual in the sample, we simulate 500 realizations from a standardized normal distribution to simulate R_i^* and redefine the samples of marginal individuals. Within each sample, we estimate the outcome test and plot its distribution across simulations.

H Randomization Test

Table H.I: Predicting Release Status

		Mapuche			
	Non-Mapuche	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Male	-0.007	-0.000	-0.014	0.008	-0.001
	(0.002)	(0.005)	(0.010)	(0.010)	(0.005)
Previous prosecution	-0.002	-0.005	-0.015	0.000	-0.005
	(0.003)	(0.009)	(0.020)	(0.021)	(0.008)
Previous pretrial misconduct	-0.010	-0.010	-0.016	-0.010	-0.010
	(0.000)	(0.001)	(0.002)	(0.002)	(0.001)
Previous conviction	-0.147	-0.114	-0.150	-0.098	-0.114
	(0.005)	(0.015)	(0.045)	(0.046)	(0.015)
No. of previous prosecutions	-0.029	-0.022	0.011	0.031	-0.023
	(0.005)	(0.008)	(0.023)	(0.022)	(0.008)
Severity (previous prosecution)	-0.023	-0.028	0.000	0.128	-0.020
	(0.009)	(0.022)	(0.047)	(0.049)	(0.022)
Severity (current prosecution)	0.005	0.013	0.012	0.014	0.012
	(0.001)	(0.003)	(0.007)	(0.008)	(0.003)
Court-by-time fixed effects	YES	YES	YES	YES	YES
Observations	647,729	50,817	9,710	9,423	52,001
Joint-F-test	1296.5	475.6	109.7	108.8	472.5
p-value	0.000	0.000	0.000	0.000	0.000
Cragg-Donald F-test (first stage, without controls)	606.0	6.3	18.9	0.0	7.1
Cragg-Donald F-test (first stage, with controls)	698.9	2.1	10.7	0.8	3.3

Note: This table presents the results of an OLS regression of release status on covariates using the data described in Table 1. Drug crime, homicide, and property crime are dummies for the crime types. The null hypothesis in the joint-F-test is that all coefficients are jointly zero. Standard errors are clustered at the year/court level. The Cragg-Donald F-test for the first stage is presented at the bottom of the table.

Table H.II: Predicting Judge Leniency

		Mapuche			
	Non-Mapuche	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Male	0.000	-0.000	-0.001	0.002	-0.000
	(0.000)	(0.001)	(0.002)	(0.002)	(0.001)
Previous prosecution	0.000	0.001	0.002	0.001	0.001
	(0.000)	(0.001)	(0.003)	(0.003)	(0.001)
Previous pretrial misconduct	0.000	-0.000	0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Previous conviction	-0.000	-0.000	-0.005	0.001	0.001
	(0.000)	(0.002)	(0.007)	(0.005)	(0.002)
No. of previous prosecutions	0.000	-0.001	0.001	0.001	-0.001
	(0.000)	(0.001)	(0.003)	(0.003)	(0.001)
Severity (previous prosecution)	0.000	-0.001	-0.003	-0.007	-0.001
	(0.000)	(0.002)	(0.005)	(0.009)	(0.002)
Severity (current prosecution)	-0.000	-0.000	-0.001	-0.001	-0.000
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
Court-by-time fixed effects	YES	YES	YES	YES	YES
Observations	647,700	49,569	8,055	7,853	50,801
Joint-F-test	0.9	0.9	1.5	0.4	0.8
p-value	0.530	0.549	0.134	0.946	0.602
Cragg-Donald F-test (first stage, without controls)	606.0	6.3	18.9	0.0	7.1
Cragg-Donald F-test (first stage, with controls)	698.9	2.1	10.7	0.8	3.3

Note: This table presents the results of an OLS regression of judge leniency on covariates using the data described in Table 1. Judge leniency is measured using the residualized leave-out group-specific release rate, as in Arnold et al. (2018). Drug crime, homicide, and property crime are dummies for the crime types. The null hypothesis in the joint-F-test is that all coefficients are jointly zero. Standard errors are clustered at the year/court level. The Cragg-Donald F-test for the first stage is presented at the bottom of the table.

I Comparing P-BOT and IV Marginal Defendants

This appendix compares, in terms of observed characteristics, the marginal defendants identified by the P-BOT and the instrument-based approach proposed by Arnold et al. (2018). Given that our IV model only has statistical power in the sample of non-Mapuche defendants, we limit the comparison to this group.

The P-BOT explicitly identifies a sample that proxies for marginally released defendants. Then, their distribution of observables is identified by simple descriptive statistics. In the case of the instrument-based approach, under the standard IV assumptions, the marginal defendants are given by the compliers. Then, we characterize the compliers' observables following the method developed by Abadie (2003) and extended to the judges design framework by Dahl et al. (2014), Dobbie et al. (2018), and Bald et al. (2019).

Let $\overline{z}$ and $\underline{z}$ denote the maximum and the minimum value for the judge leniency instrument, respectively. The fraction of compliers is identified by $\Pr(R_i = 1|Z_i = \overline{z}) - \Pr(R_i = 1|Z_i = \underline{z}) = \Pr(R_i(\overline{z}) > R_i(\underline{z}))$. This expression can be estimated using the IV first stage estimation, in particular, by multiplying the estimated coefficient on the instrument by $(\overline{z} - \underline{z})$. In practice, we assign the top and bottom percentile of the distribution of the instrument to $\overline{z}$ and $\underline{z}$, respectively.[1] By repeating the same procedure but restricting the sample to individuals with $X_i = x$, we can estimate the probability of being complier given that $X_i = x$, i.e., $\Pr(R_i(\overline{z}) > R_i(\underline{z})|X_i = x)$. Then, by Bayes rule

$$\Pr(X_i = x|R_i(\overline{z}) > R_i(\underline{z})) = \frac{\Pr(R_i(\overline{z}) > R_i(\underline{z}))}{\Pr(R_i(\overline{z}) > R_i(\underline{z})|X_i = x)} \Pr(X_i = x).$$

Using this equation we can characterize the compliers' distribution of observables.

Tables I.I presents these conditional probabilities for the marginal defendants identified by the

[1] These conditional probabilities can be also estimated by local regressions. Results are similar to the linear case.

P-BOT and the instrument-based approach, defining P-BOT marginal defendants as those released individuals whose propensity score is in the bottom 5% or 10% of the distribution, respectively. As this table shows, in all variables but one (an indicator that takes value 1 if the defendant is accused of a drug crime) when the probability of belonging to some particular group conditional on being IV-complier is higher (lower) than the unconditional one, it is also the case that the conditional probability of being a marginal defendant according to the P-BOT is higher (lower) than the unconditional probability. In the case of gender there is also a change in the direction, but the differences are small in magnitude. In other words, under both methodologies, marginally released defendants are more likely to have previous prosecutions, to have been engaged in pretrial misconduct in the past, to have been convicted in the past, and to be accused of more severe crimes. We interpret this as evidence that the non-Mapuche marginal defendants identified by the P-BOT and the instrument-based approach have similar distribution of observables. Reassuringly, around 6% of non-Mapuche defendants are compliers, while in the P-BOT the share of non-Mapuche defendants identified as marginals are 4% and 8%, when looking at the bottom 5% and 10% of the released defendants propensity score distribution, respectively.

Table I.I: Characteristics of Marginal Defendants

	Pr[$X = x$]	Pr[$X = x$|Marginal] IV	Pr[$X = x$|Marginal] P-BOT (5%)	Pr[$X = x$|Marginal] P-BOT (10%)
Male	0.885	0.885	0.917	0.920
	(0.0004)	(0.0124)	(0.0016)	(0.0011)
Female	0.115	0.117	0.083	0.080
	(0.0004)	(0.0124)	(0.0016)	(0.0011)
At least one previous case	0.680	0.821	0.926	0.874
	(0.0006)	(0.0149)	(0.0019)	(0.0019)
No previous case	0.320	0.173	0.074	0.126
	(0.0006)	(0.0152)	(0.0019)	(0.0019)
At least one previous pretrial misconduct	0.401	0.546	0.678	0.643
	(0.0006)	(0.0185)	(0.0033)	(0.0023)
No previous pretrial misconduct	0.599	0.444	0.322	0.357
	(0.0006)	(0.0191)	(0.0033)	(0.0023)
At least one previous conviction	0.653	0.803	0.901	0.850
	(0.0006)	(0.0157)	(0.0021)	(0.0019)
No previous conviction	0.347	0.193	0.099	0.150
	(0.0006)	(0.0158)	(0.0021)	(0.0019)
High Severity (previous case)	0.570	0.711	0.796	0.748
	(0.0006)	(0.0181)	(0.0025)	(0.0021)
Low Severity (previous case)	0.430	0.287	0.204	0.252
	(0.0006)	(0.0185)	(0.0025)	(0.0021)
High Severity (current case)	0.513	0.808	0.997	0.989
	(0.0006)	(0.0152)	(0.0004)	(0.0006)
Low Severity (current case)	0.487	0.162	0.003	0.011
	(0.0006)	(0.0149)	(0.0004)	(0.0006)
Drug crime	0.124	0.178	0.021	0.066
	(0.0004)	(0.0143)	(0.0009)	(0.0012)
Non-drug crime	0.876	0.819	0.979	0.934
	(0.0004)	(0.0151)	(0.0009)	(0.0012)
Property crime	0.182	0.079	0.002	0.009
	(0.0005)	(0.0114)	(0.0003)	(0.0005)
Non-property crime	0.818	0.919	0.998	0.991
	(0.0005)	(0.0114)	(0.0003)	(0.0005)

Note: This table presents the probability of belonging to different groups of observables (which are binary or were discretized using the respective median as the threshold). The sample is restricted to non-Mapuche defendants. This probability is calculated unconditionally, conditioning on being an IV-complier, and conditioning of being identified as marginal by the P-BOT. The standard errors are calculated by bootstrap (500 repetitions).

Appendix Bibliography

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231 – 263.

Amigo, H. and P. Bustos (2008). Apellidos mapuche historia y significado. *Maigret Ltda*.

Arnold, D., W. Dobbie, and C. Yang (2018). Racial bias in bail decisions. *Quarterly Journal of Economics 133*(4), 1885–1932.

Bald, A., E. Chyn, J. S. Hastings, and M. Machelett (2019). The causal impact of removing children from abusive and neglectful homes. *Working Paper*.

Cortés, T., N. Grau, and J. Rivera (2020). Juvenile incarceration and adult recidivism. *Working Paper*.

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014). Family welfare cultures. *Quarterly Journal of Economics 129*(4), 1711–1752.

Dobbie, W., J. Goldin, and C. S. Yang (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review 108*(2), 201–240.

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

Painemal, N. (2011). Apellidos mapuche vinculados a títulos de merced. *Santiago: Corporación Nacional de Desarrollo Indígena (CONADI)*.