FACULTAD ECONOMÍA Y NEGOCIOS UNIVERSIDAD DE CHILE
DEPARTAMENTO DE ECONOMÍA

SERIE DE DOCUMENTOS DE TRABAJO

SDT 493

A Simple Test for Prejudice in Decision Processes: The Prediction-Based Outcome Test

Autores:
Nicolás Grau
Damián Vergara

Santiago, Abril de 2020

sdt@econ.uchile.cl
econ.uchile.cl/publicaciones

A Simple Test for Prejudice in Decision Processes: The Prediction-Based Outcome Test*

Nicolás Grau[†] Damián Vergara[‡]

First version: January 27, 2020
This version: April 28, 2020

Abstract

We propose a novel implementation of the outcome test for diagnosing prejudice in decision processes, the Prediction-Based Outcome Test (P-BOT). We motivate our approach with a model of prejudice in pretrial detention decisions. The main empirical challenge when implementing outcome tests is the identification of marginal individuals. Our method uses the predicted release status to identify marginally released defendants. Concretely, we provide sufficient conditions under which a ranking of the propensity score among released defendants identifies those who are more likely to be at the margin given their observables, and propose a set of diagnostics to empirically assess the plausibility of the identification assumptions. Some appealing features of the P-BOT are that (i) it does not require instruments nor random assignment of judges, (ii) it is robust to standard omitted variable bias, and (iii) it is very easy to implement. However, its performance depends on the availability of good predictors, something that can be assessed by the econometrician. We use the P-BOT to test for prejudice in pretrial detentions against the main ethnic group in Chile, the Mapuche, using nationwide administrative data. We find strong evidence of prejudice against Mapuche defendants and show that the discrimination patterns are likely to be non-binary. We assess the relative performance of the P-BOT and alternative approaches to test for prejudice, and discuss the test's interpretation in more general versions of the model, sketching a new taxonomy of prejudice.

*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 Hadar Avivi, Aureo De Paula, Marina Dias, Joseph Doyle, Felipe Jordán, Patrick Kline, Attila Lindner, Charles Manski, Pablo Muñoz, Demian Pouzo, James Powell, Jesse Rothstein, Yotam Shem-Tov, Chris Walters, Harry Wheeler, and John Wieselthier, as well as seminar participants at CReAM/UCL, King's College, Toulouse School of Economics, and UC Berkeley, 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. Usual disclaimers apply.

[†]Department of Economics, Faculty of Economics and Business, Universidad de Chile, ngrau@fen.uchile.cl.

[‡]Department of Economics, UC Berkeley, damianvergara@berkeley.edu.

1 Introduction

Consider a situation where an agent (a judge, an employer) decides other individual's status (a defendant's pretrial detention status, a worker's promotion) based on a predicted expected outcome (pretrial misconduct, future productivity). The selection process is said to be prejudiced against a particular group if agents set different selection thresholds for their members because of animus and/or they systematically mispredict their expected outcomes. While testing for this source of discrimination is policy-relevant, it is empirically challenging.[1] Observed disparities that suggest prejudice could be explained by statistical discrimination (i.e., by correlation between observable group traits and unobserved predictors of the expected outcome). More importantly, the potential discriminator could be making decisions based on some information that is not observed by the econometrician, being the empirical analysis affected by standard omitted variable bias. Furthermore, in most cases, it is not possible to have plausible exogenous variation on group membership to infer prejudice from causal estimates.[2]

In this general setting, a prominent approach to test for prejudice in selection processes is the outcome test (Becker, 1957, 1993). Outcome tests are based on the idea that, in the absence of prejudice, success rates at the margin should be equivalent across groups: on average, marginally released defendants should have equal pretrial misconduct rates, and marginally promoted workers should be equally productive. The test for prejudice is then reduced to comparing the average outcome of marginally selected individuals between groups (a simple difference in means), i.e., the econometrician only requires to find a statistically significant correlation (not causal evidence) between pretrial misconduct/productivity and group membership, for those individuals at the margin. While the outcome test is robust to the presence of statistical discrimination and omitted variables, its implementation induces an additional difficulty: the identification of marginally selected individuals. If potential outcome distributions vary between groups, differences in outcomes away from the margin may lead to misleading conclusions regarding prejudice. Since the misspecification of marginal individuals can induce bias in the outcome test, a careful identification of them is central for the proper implementation of the test.

[1] For example, economists have explored to what extent discrimination explains the observed wage gaps between races and genders (for surveys, see Altonji and Blank, 1999; Lang and Lehmann, 2012; and Blau and Kahn, 2017). A similar discussion exists regarding the large and persistent racial disparities in the criminal justice system (Antonovics and Knight, 2009; Abrams et al., 2012; Anwar et al., 2012; Rehavi and Starr, 2014; Chetty et al., 2018; Fryer, 2019; Rose, 2020). For a more general discussion on the empirical literature on discrimination, see Guryan and Charles (2013).

[2] Notable exceptions are the correspondence studies that have been used to test for discrimination in the labor market. See, for example, Bertrand and Mullainathan (2004) and Kline and Walters (2020). For a review, see Bertrand and Duflo (2017).

In this context, this paper proposes a novel implementation of the outcome test –the Prediction-Based Outcome Test (P-BOT). Our approach uses the predicted status (the propensity score) to identify marginally selected individuals. We motivate our approach with a model of prejudice in pretrial detention decisions where judges decide over defendants' pretrial release status based on expected pretrial misconduct (non-appearance in court and/or pretrial recidivism). We provide sufficient conditions under which the released defendants that are more likely to be marginal given their observables also have lower propensity scores. By doing so, our identification strategy reduces the non-trivial challenge of identifying marginal individuals to a standard prediction problem.

Under this result, implementing the outcome test becomes a simple exercise. First, the econometrician has to estimate the propensity score and compute the release probabilities. Then, the econometrician can rank released defendants according to their predicted probabilities and define samples of marginals. With them, the econometrician can compute group-specific pretrial misconduct rates and perform differences in means to test for prejudice. Since what matters for the identification strategy are the predicted values, the structural interpretation of the prediction coefficients is not relevant (what matters is *who* is close to the margin, not *why*). This makes our approach robust to omitted variables. The prediction-based identification argument relies on the availability of good predictors since noise in the estimated ranking can induce bias in the outcome test. The predictive power of the observed covariates can be assessed by the econometrician by looking at the fit of the propensity score. We also propose a perturbation test to assess the potential bias coming from the noise in the estimated ranking.

Our identification strategy relies on two assumptions. First, we impose additive separability between observables and unobservables in the selection equation. Through the lens of the model, this induces strict monotonicity on observables in the risk determination, meaning that the marginal effect of observables on the latent pretrial misconduct does not depend on the value of the unobserved component. Second, while we allow for unrestricted first moments in the joint distribution of observables and unobservables, we restrict the higher moments of the conditional distribution to not depend on observables. While these assumptions are not directly testable, we propose suggestive diagnostics to empirically assess the plausibility of the assumptions.

To assess how restrictive are our assumptions, note that alternative approaches usually rely on either stronger structural assumptions or the availability of instruments that shift the release status in the margin. We discuss the alternative approaches with detail in Section 3. Relative to the instrument-based approach (in our view, the state of the art in this literature, see Arnold et al., 2018), our monotonicity assumption is less restrictive than the equivalent LATE monotonicity assumption in terms of the restrictions imposed on judges' behavior but is more restrictive in terms of the primitives of the risk equation. On the other hand, our restriction on the distribution of unobservables is weaker than standard selection-on-observables assumptions since we allow for unrestricted comovement in the first moment. Instrument-based approaches are preferred in this regard since they are agnostic about the distribution of unobservables. However, in many settings, the instrument-based approach for testing for prejudice is fragile (or even infeasible) for reasons that do not affect our identification strategy. For example, our approach does not need judges to be randomly assigned, a necessary condition for the application of the instrument-based approach. Given this discussion, we claim that our approach, the P-BOT, is a good complement for the available methods. We see our restrictions as a reasonable price for not needing instruments.

As an application of the P-BOT, we test for prejudice against the main ethnic group in Chile, the Mapuche. According to the last census, around 10% of the Chilean population reported being Mapuche. The Mapuche population is an interesting case of analysis for three reasons. First, there exists a conflict between the Mapuche and the Chilean state that dates back to more than a century ago (Cayul et al., 2018). In this context, it is frequently claimed that the institutions that have been established in Chile are biased against the Mapuche. Second, negative stereotypes have been formed about the Mapuche population. Some people in Chile think that Mapuche people are particularly lazy, violent, and alcoholic, although there is no evidence about any systematic difference in behavior between the Mapuche people and the rest of the population (Merino and Quilaqueo, 2003; Merino and Mellor, 2009). Third, Mapuche people are recognizable, both because of physical aspects and their surnames. Then, the exercise of discrimination is feasible in this setting.

We use Chilean administrative data that covers more than 95% of criminal cases in Chile between 2008 and 2017. The data contains detailed information on cases and defendants. We merge the administrative data with a register of Mapuche surnames to create different measures of ethnicity that combine self-reporting and surnames information. We provide suggestive evidence that both assumptions hold in our setting. With this information, we fit different projection models using a wide set of predictors for the release status to identify marginal individuals and perform standard outcome tests. The P-BOT provides strong evidence of prejudice against Mapuche defendants in the dictation of pretrial detention. Depending on the approach used to implement our test and on how we identify Mapuche defendants in the data, our results show that marginal Mapuche defendants are between 4 and 13 percentage points less likely to be engaged in pretrial misconduct relative to marginal non-Mapuche defendants. This result is robust to a perturbation test that induces random noise –proportional to the propensity score fit– in the estimated ranking.

By changing the definition of the margin, we provide evidence of a modest but not problematic potential inframarginality bias in our setting. Therefore, the outcome test using the full sample (à la Knowles et al., 2001) also supports the existence of prejudice against Mapuche defendants, although the implied magnitude is slightly smaller. Also, since the Chilean setting is characterized by quasi-random assignment of judges for arraignment hearings at the court-by-time level, we 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 (and exactly matches the P-BOT estimates of non-Mapuche pretrial misconduct rates at the margin), we show that the estimation is severely underpowered for the Mapuche sample, preventing us from drawing precise conclusions from its application. Moreover, we show that the non-Mapuche marginally released defendants identified by both methods have similar distributions of observables. On one hand, the similarity between the P-BOT and the IV estimation using the non-minority sample is reassuring for our identification strategy. On the other hand, the fragility of the IV estimation in the minority sample illustrates how the P-BOT is an attractive alternative when the instrument-based approach cannot be properly implemented.

We end the paper by discussing the relevance of important but usually overlooked assumptions of the outcome test. These assumptions are (i) that judges make (or, at least, should make) decisions based on expected outcomes, and (ii) that the prejudice patterns are solely based on one characteristic (e.g., race). In addition, we discuss the implications that the assignment rule (of judges to defendants) has on the test's validity and interpretation. We argue that the relaxation of these assumptions does not invalidate the application of the outcome test but affects the interpretation of the results. This leads to a novel taxonomy of prejudice. While the outcome test makes a robust identification of prejudice (understood as systematically facing different effective release thresholds for reasons unrelated with expected pretrial misconduct), the overall bias can be thought of as a combination of four different sources: *pure* prejudice (biased predictions and/or animus), *incentive-driven* prejudice (when judges look at other outcomes to make decisions that ultimately harm specific groups), *correlation-driven* prejudice (when other variables that correlate with group traits matter for the effective threshold definition), and *systemic* prejudice (when the assignment rule systematically imposes stricter conditions to specific groups).

We illustrate how the P-BOT can be used to obtain insights about the underlying sources of prejudice. First, we show that pretrial misconduct rates decrease for all groups as inframarginal defendants are included in the estimation sample. This can be simply computed using the P-BOT by changing the margin definition. This can be thought of as an indirect specification test since it suggests that judges care about potential pretrial misconduct when making release decisions. Second, additional regressors can be included in the outcome equation to test for more complex patterns of prejudice. We present two examples of outcome regressions that group defendants using two categories (and its interaction). In the first one, we group defendants using *Mapuche* and *Low income*, conjecturing that the discrimination patterns could interact with socioeconomic status. In the second one, recognizing that the long-standing conflict between the Chilean state and the Mapuche people has been mainly concentrated in a particular geographic area, we group defendants using *Mapuche* and *Mapuche Region*, conjecturing stronger discrimination patterns in those courts. Results show that prejudice against Mapuche defendants is only relevant for those Mapuche who live in low-income municipalities, with no evidence of prejudice solely based on the place of living. Also, results show that there is prejudice against Mapuche defendants in all courts but, as expected, it is stronger in the Mapuche region. These results suggest that non-binary patterns of discrimination are likely to happen in practice. Finally, we estimate the outcome equations controlling by court-by-time fixed effects (the level at which judges are randomly assigned) and find that almost none of the overall effect can be explained by systemic prejudice. These three analyses using the P-BOT are not only interesting by themselves but also are methodologically appealing since these extensions may induce problems on the alternative approaches.

This paper contributes to the literature of discrimination by proposing a simple methodology to test for prejudice. More specifically, this paper adds to the literature that proposes methods for implementing the outcome test, namely Knowles et al. (2001), Anwar and Fang (2006), Arnold et al. (2018), Marx (2018), and Arnold et al. (2020). Throughout the paper, we argue that our approach constitutes a good complement to the available methods for identifying prejudice.[3] Although we motivate our framework using a model of prejudice in pretrial detention decisions, the usefulness of the outcome test (and, therefore, of the P-BOT) as a diagnostic of prejudice is not restricted to the criminal justice system analysis. Whenever there is a measurable outcome expected to reflect the rationality of the selection equation, the P-BOT can be used to test for prejudice.

Our empirical application also adds to the literature of bias in the criminal justice system. Abrams et al. (2012) document between-judge variation in the racial gap in incarceration rates. Anwar et al. (2012) show that black defendants are more likely to be convicted when the jury pool is exclusively composed of white juries. Rehavi and Starr (2014) find that black defendants receive longer sentences than comparable white defendants for the same crimes. Knowles et al. (2001), Anwar and Fang (2006), Antonovics and Knight (2009), Simoiu et al. (2017), and Marx

[3]An alternative approach is the *threshold test*. This test requires estimating complex Bayesian models. For more details, see Simoiu et al. (2017) and Pierson et al. (2017).

(2018) provide mixed evidence (possibly driven by the heterogeneity in the employed methods) on racial bias in vehicle searches by police officers. Fryer (2019) documents large racial gaps in police use of force, but argues that discrimination is unlikely to be the main driver behind them. Rose (2020) shows that technical rules that induce probation disproportionately affect low-risk black individuals. Arnold et al. (2020) find that more than two-thirds of New York City observed racial disparities in bail decisions is explained by racial discrimination, being the latter driven by both prejudice and statistical discrimination. More related to our paper is Arnold et al. (2018), who find that bail judges are prejudiced against black defendants. Our paper provides evidence of prejudice in pretrial detention decisions against an ethnic minority in Chile in a similar setting than theirs. Understanding racial disparities in the criminal justice system is important because (i) they are large and persistent, and (ii) there is evidence that incarceration negatively affects employment, future crime, and education (Aizer and Doyle Jr, 2015; Muller-Smith, 2015; Cortés et al., 2019) and, more specifically, pretrial detention affects conviction rates, employment, and the use of state benefits (Dobbie et al., 2018; Grau et al., 2019). Therefore, the potential existence of prejudice in judicial decisions is particularly costly both from private and social perspectives.[4]

The rest of the paper is organized as follows. Section 2 sketches the argument of the paper. The goal is to develop intuition and be transparent regarding the scope of our framework and its assumptions. Section 3 presents a formal model of prejudice in pretrial detention decisions, derives the outcome test, and discusses the related empirical challenges. Section 4 describes our approach, the P-BOT. Section 5 describes the institutional setting and the data used in our empirical application, while Section 6 presents our results. Section 7 develops a critical discussion regarding the outcome test's interpretation. Finally, Section 8 concludes.

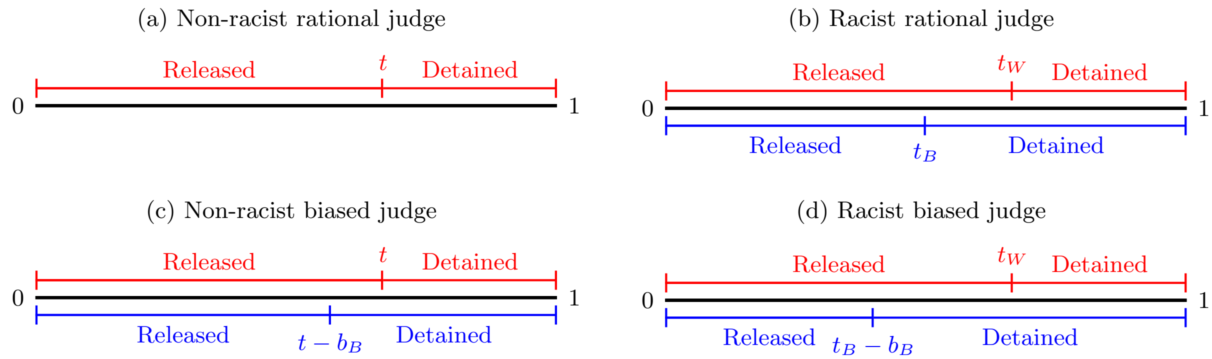
2 Prejudice, the Outcome Test, and the P-BOT

This section sketches the argument of the paper. The goal is to develop intuition and be transparent regarding the scope of our framework and its assumptions before digging into more details. Sections 3 and 4 develop the formal argument.

In this paper, we analyze potential prejudice in decision processes that are based on expected outcomes. To fix ideas, and to relate it to our empirical application, we base the formal analysis

[4]Another strand of the literature has looked at the political ideology of judges. Cohen and Yang (2019) show that Republican-appointed judges determine, on average, 3-months larger sentences for black defendants than similar non-black defendants, compared to Democratic-appointed judges. Anwar et al. (2018) show, using Swedish data, that convictions increase considerably for defendants with Arabic names when the judge is from the far-right party.

Figure 1: Decision rule: examples



on a model of pretrial detention decisions. Consider a situation where a judge decides whether to confer pretrial release or not to a defendant. Each judge has to predict how likely it is that the defendant will be engaged in pretrial misconduct (non-appearance in court and/or pretrial recidivism) during the investigation, compare that to a threshold, and make the decision. Given the legal principle of presumption of innocence, judges should not detain defendants unless the expected risk of pretrial misconduct is very high. The question we address is whether judges are prejudiced against a specific group, e.g., black defendants, when making this decision.

Figure 1 illustrates the decision process. Panel (a) shows how the decision rule looks for a non-prejudiced judge. The judge predicts the probability of pretrial misconduct using all the available information and release defendants whenever that predicted probability is smaller than t. Now suppose that the judge is racist. In this case, because of animus, the judge sets a smaller threshold for black defendants. Panel (b) shows how the decision rule looks for a racist judge, with t_W and t_B the thresholds set for white and black defendants, respectively. Only the white defendants are released when pretrial misconduct probability is between t_B and t_W. Then, this type of decision rule is discriminatory against black defendants. Now suppose that the judge is non-racist, but systematically overestimates risk for black defendants. In particular, when the true probability of pretrial misconduct is p, the judge predicts $p + b_B$ if the defendant is black. This implies that the effective threshold is smaller for black defendants. This is illustrated in panel (c). This decision rule is also discriminatory against black defendants since defendants with pretrial misconduct probability between $t - b_B$ and t will be released or not depending on their race. Finally, panel (d) illustrates a judge that is racist and makes biased predictions against black defendants. The definition of prejudice we use in this paper is the composite effect of racism and bias in predictions. The framework we develop, as it is standard in the outcome test literature, is not able to separately identify between both sources (Arnold et al., 2018).

Figure 2: Sample of Marginals

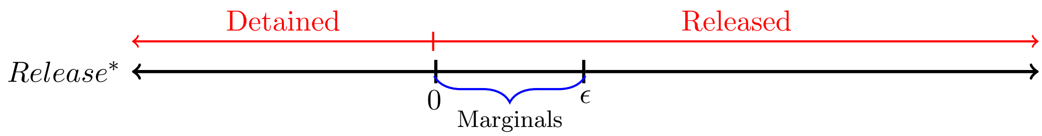


Figure 1 suggests that testing for prejudice against a certain group is equivalent to test for differences in the effective thresholds: there is prejudice against black defendants if $t_W > t_B - b_B$. This is challenging because effective thresholds are rarely observable. However, since the decision process is based on expected outcomes, we can use realized outcomes of released defendants to infer effective thresholds. This insight is the basis of the outcome test, first proposed by Becker (1957).

To understand why, consider the decision rule illustrated in panel (a). Define the marginally released defendants as the ones with probability of pretrial misconduct equal to t. Those defendants were just released. In expectation, t% of defendants that were just released should be engaged on some type of pretrial misconduct. Then, pretrial misconduct rates of marginally released defendants recover the effective threshold. Now consider panel (d). Using the same logic, $(t_B - b_B)$% and t_W% of marginally released black and white defendants, respectively, should be engaged on some type of pretrial misconduct. Then, if there is prejudice in the decision process, observed pretrial misconduct rates of marginally released black defendants should be smaller than the ones observed for white defendants. That is, testing for prejudice is reduced to a difference in means: the econometrician only requires to find a statistically significant correlation (not causal evidence) between pretrial misconduct and race, for those defendants at the margin. This difference in means can be trivially implemented if the econometrician knows, among released defendants, who is marginal. However, as we discuss in Section 3, the identification of marginal individuals is a difficult challenge since (i) the structure of the decision rule is unknown, and (ii) some of the variables that affect the release decision are not observed by the econometrician.

The contribution of this paper is to propose a novel way to identify marginal individuals. To illustrate our method, let $Release^*$ be the latent release status, so $Release = 1\{Release^* \geq 0\}$. $Release^*$ is the difference between the effective threshold and the true pretrial misconduct probability. Whenever the threshold is larger than the predicted probability ($Release^* \geq 0$), the defendant is released. Let $\epsilon > 0$ be a (small) distance from the margin. Then, we can define marginally released defendants as defendants with $Release^* \in [0, \epsilon]$. Figure 2 illustrates this definition of marginal defendants. Since for released defendants, $Release^*$ is truncated at 0, identifying released defendants with $Release^* \in [0, \epsilon]$ is equivalent to identifying the released defendants with the smaller latent indexes. Then, identifying a ranking of $Release^*$ among released defendants allows creating samples of marginal defendants to perform outcome tests.

To the extent that there are variables that judges use to make the release decisions that the econometrician does not observe, $Release^*$ is not observable nor estimable. However, the econometrician can ask, given their observables, which defendants are more likely to be marginals according to the previous definition. The main contribution of this paper is to provide sufficient conditions under which a ranking of the propensity score among released defendants identifies a ranking of these conditional probabilities. Concretely, under our assumptions, the released defendants who are more likely to have $Release^* \in [0, \epsilon]$ given their observables also have smaller propensity scores.

Under this result, implementing the outcome test becomes a simple exercise. First, the econometrician has to estimate the propensity score and compute the predicted probabilities. Then, the econometrician can rank released defendants according to their predicted probabilities and define samples of marginals to compute group-specific pretrial misconduct rates and perform differences in means to test for prejudice. Since what matters for the identification strategy are the predicted values, the structural interpretation of the prediction coefficients is not relevant (what matters is *who* is close to the margin, not *why*). This makes our approach robust to omitted variables.

To prove the ranking equivalence, we make two assumptions. The first imposes additive separability between observables and unobservables in the selection equation. Through the lens of the model, this induces strict monotonicity on observables in the potential risk equation. As we discuss in Section 4, this is less restrictive than similar monotonicity assumptions imposed in alternative methods for testing for prejudice in terms of the restrictions imposed over judges' behavior but more restrictive in terms of the primitives of the risk equation. Regarding the joint distribution of observables and unobservables, we allow for unrestricted first moments but restrict the higher conditional moments. Both assumptions may be restrictive in some scenarios. However, the alternative is to proceed either with stronger structural assumptions or with instruments which, as we argue in Section 3, is not always feasible nor desirable in this setting. Then, we claim that our approach, the P-BOT, is a good complement for the available methods. We see our restrictions as a reasonable price for not needing instruments.

3 Model

Based on the intuition developed in the previous section, this section presents a formal model of prejudice in pretrial detention decisions that closely follows Arnold et al. (2018).

3.1 Model

Preliminaries Judges are indexed by j and defendants by i. Judges are assigned to defendants according to $j(i)$. Judges use all available information to compute defendant-specific probabilities of pretrial misconduct and release defendants whenever that probability is smaller than a judge-specific threshold. Let G_i be an indicator variable that takes the value 1 if defendant i belongs to group G. The question we address is whether judges are prejudiced against defendants of group G in the release decision. On top of G_i, judges observe other characteristics of the individual, namely X_i and V_i. We assume that the econometrician observes G_i and X_i, but does not observe V_i.

Pretrial misconduct Let PM_i be an indicator variable that takes the value 1 if defendant i is engaged in pretrial misconduct. Let PM_{i0} and PM_{i1} denote pretrial misconduct if detained and released, respectively. Let $Release_i$ be an indicator variable that takes the value 1 if defendant i is released. Then, $PM_i = Release_i PM_{i1} + (1 - Release_i)PM_{i0}$. Note that $PM_{i0} = 0$, $\forall i$, given that detained defendants cannot be engaged in pretrial misconduct. We assume PM_{i1} is given by

$$PM_{i1} = 1\{PM_i^* \geq 0\} = 1\{m\left(X_i, V_i, \nu_i\right) \geq 0\}, \tag{1}$$

where ν_i are variables that affect pretrial misconduct that are not observed by the judge, and m is some function.[5] The information set is the same for all judges: all judges observe X_i and V_i and do not observe ν_i.

Selection process For making the release decision, judges use all the available information to predict PM_{i1} and compare their prediction to a threshold. Formally

$$Release_i \quad = \quad 1\left\{p(G_i, X_i, V_i, j(i)) \leq t(G_i, j(i))\right\}, \tag{2}$$

where p is a function that computes the prediction of PM_{i1}, and t is the release threshold that judges set depending on G_i. Note that $j(i)$ enters in both functions because judges are allowed to be heterogeneous in the way they make predictions and in the way they set thresholds.

We assume that the judge-specific prediction can be decomposed into two components: a rational prediction (the true conditional expectation) and a judge-specific deviation (bias). Formally

$$p(G_i, X_i, V_i, j(i)) \quad = \quad \mathbb{E}_\nu[PM_{i1}|G_i, X_i, V_i] + b(G_i, j(i)), \tag{3}$$

[5] To simplify notation, we assume that the assigned judge, $j(i)$, doesn't affect PM_i^*.

where b is a function that accounts for the judge-specific bias in the risk prediction. Putting (2) and (3) together, we can write

$$\begin{aligned} Release_i &= 1\left\{\mathbb{E}_\nu[PM_{i1}|G_i, X_i, V_i] \leq t(G_i, j(i)) - b(G_i, j(i))\right\}, \\ &\equiv 1\left\{\mathbb{E}_\nu[PM_{i1}|G_i, X_i, V_i] \leq h(G_i, j(i))\right\}. \end{aligned} \tag{4}$$

We denote the function $h(G_i, j(i)) = t(G_i, j(i)) - b(G_i, j(i))$ as the *effective threshold*.

Before proceeding, we make explicit two important assumptions of the model. First, we assume that the selection process is only based on the expected risk of pretrial misconduct. Second, we assume that G_i is the only variable that induces heterogeneity in effective thresholds at the judge-level. To keep things simple, we stick to those assumptions in what follows. However, there are valid reasons to think that judges may be optimizing different objectives functions and/or that the effective thresholds could depend on other variables as well. In Section 7 we discuss the implications of these assumptions (and their relaxation) on the test's application and interpretation.

Notions of discrimination Although we assumed (without loss of generality) that there is no direct impact of G_i on PM_i^*, there are two channels through which G_i affects the release decision. First, judges may use G_i to rationally predict unobservables that directly affect PM_i^*, namely ν_i. This is usually referred to as *statistical discrimination*. Second, judges may be prejudiced against defendants of group G, i.e., the effective threshold may depend on G_i for reasons unrelated with defendants' true risk of pretrial misconduct. This could be the case if (i) judges make biased predictions about the correlation between G_i and ν_i, and/or (ii) judges base their subjective thresholds on G_i because of animus. In the literature, (i) and (ii) are usually referred to as *bias in predictions* and *taste-based discrimination*, respectively. Consistent with the outcome test literature, the framework we develop in this paper identifies the composite effect of bias in predictions and taste-based discrimination. We denote this composite source of discrimination as *prejudice*. While it is robust to its presence, our framework does not identify statistical discrimination.[6]

The benchmark test To illustrate that the difficulties of testing for prejudice arise even in very simple frameworks, consider the following simplified version of the model. Let PM_{i1} be given by $PM_{i1} = \alpha_X X_i + \alpha_V V_i + \nu_i$, and let the true conditional expectation of ν_i be given by $\mathbb{E}[\nu_i|G_i, X_i, V_i] = \delta_G G_i + \delta_X X_i + \delta_V V_i$. In this specification, $\delta_G > 0$ accounts for statistical

[6] This does not mean that identifying statistical discrimination is not relevant. In fact, in the context of employment discrimination, both sources of discrimination are illegal (Kleinberg et al., 2018; Kline and Walters, 2020). For a discussion about the role of statistical discrimination in the criminal justice system, see Yang and Dobbie (2020).

discrimination against defendants of group G. Also assume that judges make unbiased predictions, so $b(G_i, j(i)) = 0$, $\forall j(i)$, and that they are homogeneous with release thresholds given by $t(G_i, j(i)) = \beta_0 - \beta_G G_i$, $\forall j(i)$. In this specification, $\beta_G > 0$ accounts for prejudice against defendants of group G. The decision rule that determines defendant i's pretrial status is given by

$$Release_i \quad = \quad 1\left\{\beta_0 + (\delta_G + \beta_G)G_i + (\alpha_X + \delta_X)X_i + (\alpha_V + \delta_V)V_i \leq 0\right\}. \tag{5}$$

An econometrician could estimate (5) using data on $Release_i$, G_i, and X_i, and test for prejudice by looking at the marginal effect of G_i. This is called the *benchmark test*. This approach has two potential problems that affect the identification of β_G. First, statistical discrimination and prejudice are not separately identified since the estimated parameter is a function of $\delta_G + \beta_G$. This is not a problem if the researcher cares about total discrimination. Second, and more important, if there is correlation between G_i and V_i, the estimation of the effect of G_i on the release probability is biased. Then, this approach only identifies prejudice if (i) there is no statistical discrimination, and (ii) there is no correlation between group membership and the unobserved variables.[7]

3.2 The outcome test

Let's return to the general model. Testing for prejudice in the release decision is reduced to comparing the average effective thresholds, $h(G_i, j(i))$, between groups. Define $\overline{h}(g) = \mathbb{E}[h(G_i, j(i))|G_i = g]$ as the average effective threshold faced by defendants with $G_i = g \in \{0, 1\}$. This motivates the following definition of prejudice.

Definition 1 (Prejudice): *In the absence of prejudice*

$$\overline{h}(0) \quad = \quad \overline{h}(1). \tag{6}$$

It follows that the decision rule is prejudiced against defendants of group G whenever $\overline{h}(0) > \overline{h}(1)$. While this defines an intuitive null hypothesis to be rejected, its application is challenging since effective thresholds are not observable. Moreover, omitted variable bias and the presence of statistical discrimination prevent basic observational approaches to be informative about them.

As discussed in Section 2, one prominent approach to identify effective thresholds without observing them is the *outcome test* (Becker, 1957, 1993). This approach is based on the success rates of the selection process, and it is robust to omitted variables and the presence of statistical dis-

[7] Arnold et al. (2020) develop a clever weighting methodology for recovering total discrimination from benchmark regressions in the presence of relevant omitted variables under random assignment of judges.

crimination. Let the latent release status be given by $Release_i^* = h(G_i, j(i)) - \mathbb{E}_\nu[PM_{i1}|G_i, X_i, V_i]$, hence $Release_i = 1\{Release_i^* \geq 0\}$. We say that a released defendant is marginal if $Release_i^* = 0$. The following proposition formalizes the outcome test.

Proposition I:

$$\mathbb{E}[PM_i|G_i = g, Release_i^* = 0] \quad = \quad \overline{h}(g). \tag{7}$$

Proof. See Appendix A.

Corollary (Outcome Test): *In the absence of prejudice*

$$\mathbb{E}[PM_i|G_i = 0, Release_i^* = 0] \quad = \quad \mathbb{E}[PM_i|G_i = 1, Release_i^* = 0]. \tag{8}$$

If the econometrician rejects the null hypothesis in favor of $\mathbb{E}[PM_i|G_i = 0, Release_i^* = 0] > \mathbb{E}[PM_i|G_i = 1, Release_i^* = 0]$, then the decision process is prejudiced against group G. Note that to properly perform this test, the econometrician does not need to identify the causal parameter for the impact of group membership on pretrial misconduct. To reject the null hypothesis of non-discrimination the test only requires finding a statistically significant correlation between pretrial misconduct and group membership, for those defendants at the margin.[8]

The intuition of the outcome test is as follows. In the absence of prejudice, effective thresholds do not vary with G_i. Since effective thresholds are equal to the expected pretrial misconduct if released at the margin, marginally released defendants of different groups should have the same expected potential pretrial misconduct. Then, at the margin, pretrial misconduct rates (potentially observable) are informative about effective thresholds (not observable).[9]

The testable implications of the outcome test are the same if potential prejudice is driven by biased predictions or taste-based discrimination. It is not possible to separately identify both sources of discrimination without additional assumptions or complementary tests. See Arnold et al. (2018) for a discussion on tests for identifying between the two sources of prejudice.

Identification of marginal individuals Conditional on identifying marginally released defendants, the implementation of the outcome test is straightforward. However, the identification of

[8] In Appendix B we derive the outcome test in models in which (i) the outcome is binary but judges predict the (continuous) latent risk, and (ii) the outcome is continuous.

[9] An implicit assumption for the applicability of the outcome test is full-support, i.e., there is a mass of defendants with $G_i = 0$ and $G_i = 1$ as we approach (from the right) to $Release_i^* = 0$.

marginal individuals is not trivial. Since $Release_i^*$ depends on V_i and the structure of the release rule, in most settings the ranking judges make when deciding who to release is not observed by the econometrician. This is important because the misspecification of marginal individuals may induce bias in the outcome test. In particular, 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 lead to misleading conclusions regarding prejudice. This is called the *inframarginality bias*.[10]

The empirical literature has taken three different paths to deal with this issue. The first is to assume that the observables available to the econometrician are rich enough to invoke selection-on-observables assumptions. An example of this approach is Chandra and Staiger (2010).

The second path is to provide additional structural assumptions to infer the behavior of marginal defendants. An influential example is Knowles et al. (2001). In the context of motor vehicle searches for contraband, the authors model conditions under which the marginally searched individuals have the same behavior than the average ones, so linear regressions of the outcome equation using the full sample of selected individuals are enough to test for prejudice. However, as Anwar and Fang (2006) note, some key assumptions in Knowles et al. (2001)'s model are problematic, making their approach potentially affected by the inframarginality bias.[11]

To avoid strong structural assumptions, the third path consists on using instruments that shift the release status at the margin to identify the local conditional expectation of the outcome. Introduced by Arnold et al. (2018), this is the state of the art in the empirical literature of prejudice in decision processes. The context of the paper is pretrial detention decisions and its logic is as follows. Suppose the econometrician has an instrument for the release status whose compliers are the marginal defendants. Then, performing 2SLS estimations where the second stage regress PM_i on $Release_i$ identifies the average behavior of marginally released defendants.[12] The outcome test is then reduced to compare group-specific LATEs. Exploiting a setting in which bail judges are randomly assigned to defendants, the authors propose to use the judge-specific leave-out mean release rate as an instrument (as in Dobbie et al., 2018). This is known as the *judges design*.[13]

[10]Section 2.2. of Simoiu et al. (2017) and Online Appendix C of Arnold et al. (2018) provide intuitive explanations of the inframarginality bias.

[11]Among other assumptions, Knowles et al. (2001) assume that police officers are *monolithic*, which means that officers of different races use the same search criteria for a given motorist race. Anwar and Fang (2006) provide evidence against this assumption.

[12]The instrument can also be used to estimate MTEs at the margin of release (Heckman and Vytlacil, 1999, 2005).

[13]As it is discussed in Arnold et al. (2018), the judges design approach needs the instrument to be continuous. Marx (2018) proposes an instrument-based test for absolute prejudice that allows for discrete instruments.

While this approach is appealing, its assumptions may be restrictive in some settings. First, to meet the exogeneity condition, judges have to be randomly assigned to defendants.[14] Second, because this instrument is equivalent to running a first-stage on judge fixed effects, it may under-powered in some settings.[15] Finally, as emphasized by Muller-Smith (2015) and Frandsen et al. (2019), under plausible judge heterogeneity, the leave-out mean release rate may fail to meet the LATE monotonicity assumption (Imbens and Angrist, 1994).[16] In particular, if effective thresholds depend on additional observable variables, the leave-out mean release rate may not shift the release status in the same direction for all defendants within a group.[17]

In the context of this discussion, we propose a novel and simple approach to identify marginally released defendants to implement standard outcome tests. Our approach does not need a valid instrument nor random assignment for its implementation, at the cost of imposing additional assumptions. In addition, our approach relies on the availability of good predictors for the release status, a feature that is not needed in the instrument-based approach. Given that, we argue that our proposed approach is a suitable complement to the available methods.

4 The Prediction-Based Outcome Test

Based on the model presented in Section 3, in this section we propose a novel approach to identify marginal individuals to perform outcome tests. We formally discuss identification and estimation, and the virtues and weaknesses of our method relative to the available alternatives in the literature.

4.1 The Prediction-Based Outcome-Test

Thought experiment Our approach tries to mimic the following thought experiment. Suppose that the econometrician observes the latent release status, $Release_i^*$. If that was the case, the econometrician could use the latent variable to rank released individuals and define arbitrary notions

[14] Many settings are characterized by non-random assignment of evaluators. This is, possibly, the reason why there are few applications of the instrument-based approach beyond the specific setting of detention decisions (see Dobbie et al., 2018; and Benson et al., 2019 for other applications).

[15] The reason is that, for not violating monotonicity by construction, the instrument has to be computed separately by group. Then, many/weak concerns arise naturally for groups of observables with a small number of defendants given that several judges may not be exposed to defendants with those characteristics.

[16] Consistent with deviations from the LATE monotonicity assumption, Norris (2020) documents disagreement between judges in the Canadian refugee appeal court.

[17] Arnold et al. (2020) develop a hierarchical MTE model that imposes additional structure to allow for deviations from strict monotonicity. This model is more flexible than the standard IV approach, with the cost of imposing additional structural assumptions.

of the margin. Suppose that the econometrician labels individuals as marginals if $q(Release_i^*) \leq \bar{q}$, where q is the empirical percentile function (defined over the sample of released individuals) and $\bar{q}$ is an arbitrary (small) percentile. The outcome test then would be easily implemented by regressing PM_i on G_i within the sample of marginally released defendants.[18]

Certainly, this approach is infeasible since $Release_i^*$ is unobserved. Moreover, because V_i is also unobserved and the structure of the model is unknown, $Release_i^*$ cannot be easily estimated. However, the econometrician can try to identify which released defendants are more likely to have lower latent release indexes given their observables. This is what our approach does.

The P-BOT To simplify notation, in what follows we suppress the conditioning on $j(i)$ and assume it is part of X_i or V_i, depending on if the econometrician observes the assigned judges or not. Similarly, we suppress the conditioning on G_i and assume it is part of X_i. With this notation we can write $Release_i^* = f(X_i, V_i)$, where f is some function.

The econometrician wants to know, given X_i, which released defendants are more likely to be close to the margin. Let $\epsilon > 0$ be a small distance to the margin. The object the econometrician cares about is $\Pr(Release_i^* < \epsilon | X_i, Release_i = 1)$: the larger its value, the more likely that released defendants with those observables are close to the margin. Then, following the logic of the thought experiment, the econometrician could label as marginals the released defendants with the largest values of $\Pr(Release_i^* < \epsilon | X_i, Release_i = 1)$. Given that the distribution of V_i conditional on X_i is unknown, the econometrician cannot compute the aforementioned conditional probabilities without additional assumptions. However, below we provide sufficient conditions under which a ranking of released defendants based on $\Pr(Release_i^* < \epsilon | X_i, Release_i = 1)$ is identified by a ranking of the predicted release probabilities, $\mathbb{E}[Release_i | X_i]$ (the propensity score). In particular, under our assumptions, observables that induce higher conditional probabilities among released defendants also induce lower propensity scores.

This result is appealing because it reduces the non-trivial challenge of identifying marginal defendants to estimating $\mathbb{E}[Release_i | X_i]$, which can be done by fitting flexible projection models. In a sense, the identification of marginal individuals is reduced to a prediction problem. Therefore, as long as the observables X_i have good predictive power, the strategy will be well-behaved. It is because of this property that we call our method the Prediction-Based Outcome Test: prediction (rather than causal) models can solve the problem of identifying marginal individuals.

[18] Note that in this thought experiment the econometrician can also assess how pervasive is the inframarginality bias by changing $\bar{q}$. In fact, Knowles et al. (2001) approach coincides with setting $\bar{q} = 100$.

4.2 Identification

Assumptions In our setting, identification means that the ranking based on the propensity score is an unbiased estimator of the ranking based on the true conditional probabilities. To prove identification, we make two sets of assumptions.

Assumption 1 (A1): *There are functions d and g such that* $1\{f(X_i, V_i) \geq 0\} = 1\{d(X_i) - g(V_i) \geq 0\} \equiv 1\{d(X_i) - W_i \geq 0\}$.

A1 says that there is an additive separable representation of the selection equation. Recall that, in our model, $f(X_i, V_i) = h(X_i) - \mathbb{E}_\nu[1\{m(X_i, V_i, \nu_i) \geq 0\}|X_i, V_i]$. Then, A1 implies restrictions on the latent risk function, $m(X_i, V_i, \nu_i)$. In particular, a set of possible sufficient conditions are: (i) $m(X_i, V_i, \nu_i)$ is additive separable, i.e., $m(X_i, V_i, \nu_i) = a_X(X_i) + a_V(V_i) + a_\nu(\nu_i)$, (ii) $\mathbb{E}\left[a_\nu(\nu_i)|X_i, V_i\right]$ is additive separable, i.e., $\mathbb{E}\left[a_\nu(\nu_i)|X_i, V_i\right] = c_X(X_i) + c_V(V_i)$, and (iii) the conditional cumulative distribution of $\xi_i = a_\nu(\nu_i) - \mathbb{E}\left[a_\nu(\nu_i)|X_i, V_i\right]$, $F_{\xi|X,V}$, is strictly increasing and independent of V_i. Under these conditions, we can define $d(X_i) = F^{-1}_{\xi|X}(h(X_i)|X_i) - a_X(X_i) - c_X(X_i)$ and $g(V_i) = a_V(V_i) + c_V(V_i)$ to meet A1.[19]

Then, through the lens of our model, A1 implies strict monotonicity on observables in the expected risk equation. Intuitively, this condition states two things. First, it says that, conditional on V_i, changes in X_i move the latent risk in the same direction for every defendant, regardless of the realization of V_i. For example, assume X_i is an indicator variable that takes the value of 1 if the defendant has been prosecuted in the past, and V_i is the defendant's employment status. Both variables are observed by the judge and expected to affect the likelihood of engaging in pretrial misconduct if released, but the employment status is not observed by the econometrician. A1 implies that the marginal effect of the past prosecution indicator on the latent risk is the same for all defendants, regardless of their employment status. If the econometrician has strong beliefs that that is not the case, then A1 is likely to be violated. Second, the conditions imply that while the variables that the judge does not observe but need to be predicted in order to assess the defendant's risk, ν_i, are allowed to have different conditional means depending on X_i and V_i, those conditional means are additive separable and the deviations from the mean are assumed to be independent of the variables that the judge observe but the econometrician does not.

The restrictiveness of A1 is a matter of discussion. In particular, we see the restrictions on the conditional distribution of ν_i as second-order. This is the case if the information set for the judge

[19] Under the assumption that the judges predict the continuous outcome, $m(X_i, V_i, \nu_i)$ rather than the binary realization, A1 imply weaker assumptions on the risk equation. However, the outcome test has a slightly different interpretation. See Appendix B for details.

is rich enough. However, the strict monotonicity on observables in the latent risk equation is less trivial and may be restrictive in some cases. A benchmark for comparing this restriction is the LATE monotonicity assumption of the judges design framework. That assumption states that the instrument shifts the treatment status in the same direction for all individuals. This means that the judge-group-specific leniency measures are monotone. That assumption is likely to be violated if the effective threshold depends on additional observables. While we return to this discussion in Section 8, allowing for thresholds that depend on additional observables does not violate our identification assumptions.[20] On the other hand, the LATE assumption is agnostic about the structure of the latent risk. Then, our monotonicity assumption is weaker than the LATE assumption in terms of judges' behavior but stricter in terms of the structure of the latent risk.

A1 imposes restrictions on the joint effect of X_i and V_i on the decision rule. However it does not impose restrictions on their joint distribution. The needed distributional restrictions are summarized in A2. Recall that, under A1, $1\{f(X_i, V_i) \geq 0\} = 1\{d(X_i) - W_i \geq 0\}$. A2 imposes restrictions on the conditional distribution of W_i.

Assumption 2 (A2): *Let* $W_i = \mathbb{E}[W_i|X_i] + \widetilde{W}_i$. $\widetilde{W}_i$ *is log-concave and independent of* X_i.

While log-concavity is a standard regularity assumption, the independence of $\widetilde{W}_i$ is less trivial. Intuitively, A2 implies that while the conditional expectation of W_i given X_i is unrestricted, higher moments are not allowed to depend on X_i. This restricts, for example, complex patterns of heteroskedasticity or conditional skewness. As it is clear in the identification proof (see Appendix A), A2 is a sufficient but not necessary condition for identification. What the identification argument ultimately needs are restrictions on the conditional densities near the margin, in particular, that the conditional mass near the margin cannot be too different between sets of observables with similar propensity scores. A2 ensures that with a stronger but more transparent and intuitive restriction.[21]

Discussion Both assumptions are not innocuous and may be restrictive in some scenarios. However, relative to the literature, these assumptions are not particularly binding. While imposing restrictions on the distribution of unobservables is not ideal, our restrictions are weaker than independence or selection-on-observables, since we allow for unrestricted comovement in the first moment. On the other hand, while the main advantage of instrument-based approaches is that they are agnostic about the distribution of unobservables, in many settings the instrument-based

[20] The LATE assumption is less restrictive when the leniency measure is constructed for finer bins of observables. However, that exercise is usually infeasible given the power problems of the instrument.

[21] In Appendix C we illustrate with examples that deviations from A2 have to be large enough to invalidate the identification argument.

approach may be fragile (or even infeasible) for reasons that do not affect our identification strategy. Then, we see our approach as complementary to the existing alternatives for testing for prejudice.

Moreover, while both assumptions are not directly testable, we propose empirical diagnostics for assessing their plausibility. These diagnostics are discussed with detail in Appendix E and implemented in our empirical application.

Identification Under A1 and A2, we can prove identification. Proposition II summarizes the main result of the paper.

PROPOSITION II: *Let x_1 and x_2 be two possible realizations of X_i, and $\epsilon > 0$ be a small distance to the margin of release. Under A1 and A2,*

$$\begin{aligned} \Pr\left(Release_i^* \leq \epsilon | X_i = x_1, Release_i = 1\right) \quad &> \quad \Pr\left(Release_i^* \leq \epsilon | X_i = x_2, Release_i = 1\right) \\ \iff \qquad \mathbb{E}\left[Release_i | X_i = x_1\right] \quad &< \quad \mathbb{E}\left[Release_i | X_i = x_2\right]. \end{aligned} \tag{9}$$

Proof. See Appendix A.

Under this result, marginally released defendants can be identified, in expectation, by a ranking of the propensity score. Then, a projection of $Release_i$ on X_i, informs of the relative distance to the threshold in probability. Two things are worth discussing this result.

Prediction The identification argument relies on the predicted release status but not on the specifics of the prediction model. This is what makes our approach robust to omitted variable bias: not observing V_i possibly biases the estimated coefficients of the prediction model, but that bias improves the prediction of the conditional expectation.[22] In other words, omitted variables do not bias the estimation for the expected proximity to the margin, because the econometrician only needs to know *who* is close to the margin, not *why*.

A drawback of this latter feature is that our approach is not informative about the specifics of the selection equation. This makes the P-BOT silent regarding mechanisms: if the econometrician finds evidence of prejudice using the P-BOT, it cannot be assessed whether this discrimination comes from biased predictions or animus. The same is true for a researcher that is interested in the behavioral foundations of the selection process.

[22] Also, note that since the release decision is deterministic from the judge's perspective, not observing V_i makes the estimation of the propensity score feasible.

Conditional variance and inframarginality bias The ranking based on the propensity score identifies the relative distance to the margin among released individuals *in expectation*. That is, the estimation of the ranking is unbiased but it can be noisy. The variance in the estimated ranking is driven by the conditional variance of V_i. Variance in the estimated ranking implies that inframarginal defendants are potentially included in the sample of marginals. Then, the noise in the estimated ranking may generate inframarginality bias. This implies that an implicit assumption in the application of our method is the availability of good predictors. Intuitively, the more predictive power X_i has, the lower the predictive action the econometrician is missing from not observing V_i, and hence the lower its conditional variance.[23]

The predictive power of X_i can be empirically assessed by evaluating the fit of the projection equation. Furthermore, under A1 and A2, it is possible to assess the extent of bias caused by the noise in the estimated ranking. Recall that under our assumptions the selection rule can be written as $Release_i = 1\{Release_i^* \geq 0\} = 1\{d(X_i) - \mathbb{E}[W_i|X_i] - \widetilde{W}_i \geq 0\}$, where W_i is a function of V_i and $\widetilde{W}_i$ is independent of X_i. Since the econometrician observes $Release_i$ and X_i, it is possible to estimate $d(X_i) - E[W_i|X_i]$ and the variance of $\widetilde{W}_i$. The estimated variance of $\widetilde{W}_i$ can be used to simulate perturbations that may alter the estimated ranking and, therefore, the defendants that are considered as marginals. By recomputing the outcome test on each of these simulations, we can check whether there are simulated realizations that revert the conclusion of the test. In the next subsection we describe with more detail how to implement this perturbation test.

4.3 Estimation and implementation

Proceeding as in the thought experiment, the econometrician can use the predicted release probabilities to rank released defendants according to their conditional probability of being close to the margin and then estimate the outcome equation on a sample of defendants at a given margin definition. Following this logic, we propose two approaches for implementing the P-BOT. To simplify notation, let $\hat{R}_i$ denote the estimated propensity score.

Simple approach This approach consists in defining the sample of marginal individuals based on the percentiles of the predicted probabilities, i.e., labeling an individual as marginal if $q(\hat{R}_i) \leq \overline{q}$, where $\overline{q}$ is the arbitrary definition of the margin. Then, the outcome test can be implemented estimating a linear regression of PM_i on G_i using the sample of marginal individuals. Negative

[23]When the predictive power of X_i is very weak, the sample of marginals converges to a random sample of released individuals. In that case, the P-BOT becomes a less precise version of an outcome test using the whole sample.

and significant estimates of the coefficient on G_i, $\hat{\psi}_S$, constitute evidence of prejudice against group G. In our empirical application below we consider $\bar{q} \in \{5, 10\}$. Note that there is a standard bias-variance tradeoff in the choice of $\overline{q}$: while choosing a larger $\overline{q}$ mechanically increases the sample size and, therefore, improves the precision of the estimation, it implies that the outcome equation is estimated using a larger share of inframarginal individuals. This leads to a natural inframarginality test: by analyzing the sensitivity of $\hat{\psi}_S$ to the choice of $\overline{q}$, the econometrician can assess how pervasive is the inframarginality problem.

Non-parametric approach As a refinement, we suggest performing non-parametric local regressions for estimating $\mathbb{E}\left[PM_i|G_i = 0, q(\hat{R}_i) = 1\right]$ and $\mathbb{E}\left[PM_i|G_i = 1, q(\hat{R}_i) = 1\right]$, and assess the extent of prejudice by computing $\hat{\psi}_{NP} \equiv \mathbb{E}\left[PM_i|G_i = 1, q(\hat{R}_i) = 1\right] - \mathbb{E}\left[PM_i|G_i = 0, q(\hat{R}_i) = 1\right]$.[24] An advantage of this approach is that it weights observations according to their relative distance to the margin definition. In our empirical application, we use triangular kernel functions up to the percentiles 5th and 10th to weight observations.

Inference The distributions of the two proposed estimators of prejudice, $\hat{\psi}_S$ and $\hat{\psi}_{NP}$, have to consider that the sample definition criterion is estimated. Thus, we suggest using bootstrap to calculate confidence intervals. In our empirical application below, we consider 200 repetitions.

Perturbation test Recall that the noise in the estimated ranking can generate inframarginality bias. In the previous subsection we described a perturbation test to assess the degree of this source of bias. In this section we describe its implementation.

We focus on the case where the propensity score is estimated using a probit model. This analysis has the following steps. First, estimate a probit model for the release status. Then, for each released individual, simulate K realizations from a standard normal distribution. This standardized normally distributed random variable is the corresponding $\widetilde{W}_i$.[25] Finally, for each of the K realizations, and given the estimated parameters of the probit model, simulate $Release_i^*$ for all the released defendants, define samples of marginally released defendants, and estimate the group-specific pretrial misconduct rates for marginal defendants.

[24]Theoretically, the econometrician could condition on $\hat{R}_i = \min_j\{\hat{R}_j\}$ given that these expectations have to be estimated for those individuals who were closest to not being released. However, we suggest to focus on the 1st percentile to avoid bias due to outliers in the predicted probabilities.

[25]Recall that in a probit model the point estimates are estimations of the regression coefficients divided by the standard deviation of the unobserved component. The size of the conditional variance is therefore implicitly incorporated in the magnitude of the estimated coefficients.

With the estimated pretrial misconduct rates, there are two ways to assess the potential bias induced by the noise in the estimated ranking. The more demanding approach consists on evaluating the degree of overlap between the simulated group-specific distributions of pretrial misconduct rates of marginally released defendants. If the P-BOT provides evidence of prejudice against group G and these simulated distributions do not overlap, then the noise in the estimated ranking is not problematic. The less demanding approach consists on seeing whether the distribution of the P-BOT estimate across all simulations includes the zero value. If the P-BOT provides evidence of prejudice against group G and the distribution does not include the zero value, then the noise in the estimated ranking is not problematic.

4.4 Discussion

We end this section summarizing the properties of the P-BOT.

We think that our approach has four main good properties. First, given that the P-BOT is an application of the outcome test, it is by construction robust to the presence of statistical discrimination and omitted variable bias in the outcome equation. Second, since identification is based on a prediction argument, the identification of marginal individuals is also robust to standard omitted variable bias. Third, the P-BOT does not need random assignment of judges nor valid instruments. This is a very important feature since, as we argued in Section 3, instrument-based approaches are sometimes problematic (or even infeasible). Fourth, its implementation is simple. In our framework, testing for prejudice is reduced to projection models and linear regressions.

On the contrary, we identify two main limitations. First, our identification strategy relies on assumptions that may be restrictive in some settings. Relative to the alternative approaches, our assumptions are, on some regards, weaker, and on others, stronger. In Section 6 and Appendix E we discuss empirical diagnostics to assess the plausibility of the identification assumptions. Second, how well the P-BOT deals with the inframarginality problem depends on the availability of good predictors. If the conditional variability of the unobserved component is relatively large, the noise in the predicted ranking may induce bias in the outcome test. Therefore, our test is expected to work better in settings where rich predictors are available. That said, we propose a simple perturbation test to see the relevance of this potential bias.

Given this analysis, we see the P-BOT as a suitable complement for the available methods offered by the literature. The virtues of our test are tightly related to the usual critiques that affect the other methods. By contrast, the weaknesses of the P-BOT do not apply to its alternatives.

5 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 main ethnic group in Chile, the Mapuche, using nationwide administrative data. This section describes the institutional setting and data.

5.1 Setting

The current Chilean justice system was implemented in 2005 and works equally in all Chilean localities. We focus on pretrial detentions. The procedure for people arrested under probable cause (i.e., red-handed, without a warrant) is as follows. During the 24 hours after the initial detention, there is an 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 judges' decision is effectively binary. Because of the legal principle of presumption of innocence, judges should not incarcerate defendants unless there is clear danger of escape (i.e., high probability of failing to appear in court), the defendant represents a danger to society (i.e., high probability of committing a new crime during the investigation), and/or imprisonment aids the investigation of the criminal case.[26] In general, the arrangement hearing is very short (lasts about 15 minutes) and is carried out by quasi-randomly assigned judges: within a court, at the beginning of a given month, judges are assigned to different time slots to lead arraignment hearings with no reason other than evenly splitting the duty among the court's judges. Moreover, judges are not systematically assigned to the same time slots between months and they only have limited information and time to decide whether incarcerating or not the defendant during the investigation.[27]

We test for prejudice against the main ethnic group in Chile, the Mapuche. According to the last census, around 10% of the Chilean population reported being Mapuche. The Mapuche population is an interesting case of analysis for three reasons. First, there exists a conflict between the Mapuche and the Chilean state that dates back to more than a century ago (Cayul et al., 2018). In this context, it is frequently claimed that the institutions that have been established in Chile are biased against the Mapuche. Second, negative stereotypes have been formed about the Mapuche

[26] As described in Grau et al. (2019), pretrial detention has become more frequent between 2007 (17,891 cases) and 2018 (34,815 cases), which implies that the fraction of cases dictating pretrial detention has increased from 7.3 to 9.6%. Pretrial detainees as a share of total prisoners rose from 21.9% in 2007 to 36% in 2017, an increase of 64.4%.

[27] We only use this exogenous variation in the probability of pretrial detention when we compare our estimation results with other approaches of the literature. As it was described below, the P-BOT does not need instrumental variables nor random assignment of judges to its implementation. As we discuss in Section 7, the assignment rule has implications for the interpretation of the outcome test, but not for its validity.

population. Some people in Chile think that Mapuche people are particularly lazy, violent, and alcoholic, although there is no evidence about any systematic difference in behavior between the Mapuche people and the rest of the population (Merino and Quilaqueo, 2003; Merino and Mellor, 2009). Third, Mapuche people are recognizable, both because of physical aspects and their surnames. Then, the exercise of discrimination is feasible in this setting.

5.2 Data

We use administrative records from the Public Defender Office (PDO). The PDO is a centralized public service depending on the Ministry of Justice that provides criminal defense to all accused individuals who ask for the service, trying to enforce the rightful process in the criminal case. The dataset covers the 2008-2017 period and contains case characteristics (court, type of crime, start and end dates of the case, outcome of the case, if there was pretrial detention and for how long, among others) and defendant characteristics (ID, gender, self-reported ethnicity, municipality of residence, pretrial misconduct, among others). In addition, we are able to identify the judges assigned to each case at the beginning of the criminal process (when the determination of pretrial detention occurs).

Given that self-reported ethnicity is subject to measurement error because of potential under-reporting, we merge the administrative data with a register of Mapuche's surnames to build more robust measures of ethnicity. Since Chilean citizens are identified with two different surnames (father's and mother's), we build several Mapuche indicators. A defendant is identified as Mapuche if i) has at least one Mapuche surname, ii) has two Mapuche surnames, iii) self-reports to be Mapuche, and iv) has at least one Mapuche surname and self-reports to be Mapuche (the most comprehensive definition). On the other hand, a defendant is identified as non-Mapuche if i) and iii) fail to hold.

To build the estimation sample, we consider all adult defendants who were arrested under probable cause between 2008 and 2017. We drop hearings following an ongoing investigation (i.e., with a warrant) since the information set available to the judge may be different in those cases. To focus on arrangement 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 arrangement hearing, we keep the information related to the most severe crime (with severity measured as the probability of pretrial detention).[28]

[28] A more detailed description of the data, the sample restrictions, and the variables, is presented in Appendix D.

Table 1: Descriptive Statistics

	Non-Mapuche	Mapuche			
		At least one surname	Two surnames	Self-Reported	Self-Reported or at least one surname
Released	0.83	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	0.16	0.14	0.10	0.14	0.14
Pretrial misconduct	0.27	0.25	0.21	0.25	0.25
Individual Characteristics					
Male	0.88	0.89	0.91	0.92	0.89
At least one previous case	0.76	0.74	0.71	0.75	0.75
At least one previous pretrial misconduct	0.57	0.54	0.48	0.52	0.54
At least one previous conviction	0.63	0.60	0.54	0.58	0.60
No. of previous cases	3.47	3.15	2.69	3.05	3.17
Severity previous case	0.07	0.06	0.05	0.05	0.06
Severity current case	0.19	0.18	0.16	0.17	0.18
Judge/Court Characteristics					
Judge leniency	-0.00	0.01	0.01	0.01	0.01
Average severity (year/Court)	0.08	0.08	0.07	0.07	0.08
No. of cases (year/Court)	2,899	2,589	2,191	1,689	2,578
No. of judges (year/Court)	46	40	32	20	39
Observations (released)	536,974	43,058	8,429	7,927	44,022
Observations (non-released)	106,233	7,860	1,236	1,400	8,072

Note: This table presents the descriptive statistics of our estimation sample. The sample considers all adult defendants who were arrested under probable cause (i.e., red-handed, without a warrant) between 2008 and 2017. We drop hearings following an ongoing investigation (i.e., with a warrant) 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 keep the information related to the most severe crime (with severity measured as the probability of pretrial detention). Judge leniency is measured as the residualized leave-out mean release rate as in Dobbie et al. (2018).

Descriptive statistics Table 1 presents the descriptive statistics of our estimation sample. Mapuche defendants represent 7.5% of the total sample when we consider our most comprehensive definition ($52,094/695,300$). Release occurs in about 85% of the cases, with a minor difference in favor of the Mapuche. Regarding the outcomes that pretrial detention seeks to avoid, conditional on being released, between 21 and 27% of the individuals (depending on the group) are engaged on at least one type of pretrial misconduct, either non-appearance in court or pretrial recidivism. In all these three ways to measure pretrial misconduct, released Mapuche defendants behave better during prosecution than released non-Mapuche defendants. Also, Mapuche defendants have less severe criminal histories on average, both measured as the number of previous cases and their severity. Their current cases are also slightly less severe. Finally, Mapuche defendants, on average, face judges who are slightly more lenient and courts that handle fewer cases.

6 Empirical Application: Results

This section presents the results of our empirical application. First, we assess the validity of the assumptions of the identification strategy. We then discuss the prediction model for the release status. We then perform the outcome test using our prediction-based method for identifying marginally released defendants and the perturbation test to assess the potential bias due to the noise in the estimated ranking. Finally, we perform alternative tests for prejudice and compare the results. In Section 7 we discuss in depth the interpretation of the outcome test through the lens of the model and provide additional empirical tests to assess the correct interpretation of our results.

6.1 Assumptions' validity

While A1 and A2 are not directly testable, in Appendix E we provide different pieces of evidence that suggest that both assumptions hold in our setting.

Recall that A1 implies monotonicity in observables in the selection equation which, through the lens of the model, implies monotonicity in observables in the potential risk equation. Tables 8 and 9 of Appendix E show that the coefficients of regressions of $Release_i$ and PM_i on observables are very stable (in terms of sign and magnitude) when they are estimated using subsamples with presumably different unobservables. In fact, in 97% of the cases considered, the sign of the coefficient is consistent between subsamples. We interpret this as strong evidence in favor of A1.[29]

A2 is more difficult to test since a formal diagnostic requires stronger structural assumptions.[30] Moreover, as we discussed in Section 4, A2 is sufficient but not necessary. Given that, we propose a second diagnostic that jointly assess the identification assumptions. Noting that the relevant unobservables are variables observed by the judges, we interpret X_i as unobservables that the econometrician happened to see. We then simulate unobservables by excluding covariates and fit prediction models using a restricted set of observables. With those predictions, we can compute rank correlations between the (restricted) propensity scores among released by groups of observables and the conditional probabilities of being marginal recovered from the unrestricted estimation. Table 10 of Appendix E shows, using different rankings, statistics, and excluded variables, that the rank correlations are very large in all cases. We interpret this as a broader support for our necessary conditions.

[29] This test is similar to the monotonicity test presented in Arnold et al. (2018) and Bald et al. (2019).

[30] For example, we could fit an heteroskedastic probit model and formally test for heteroskedasticity. When doing this, we reject the null of homoskedasticity. However, that test is only valid if the model is correctly specified.

6.2 Prediction model

We estimate the propensity score using a probit model, and consider the following covariates: a Mapuche indicator, a male indicator, whether the individual has previous prosecutions, whether the individual was engaged in pretrial misconduct during a previous prosecution, whether the individual was convicted in the past, the number of previous prosecutions, the severity of previous prosecutions, the severity of the current prosecution, the number of cases per court-year, the number of judges working at the court the year of the prosecution, the assigned-judge leniency, the square of the latter, and year of prosecution fixed effects.[31] A more detailed description of the variables can be found in Appendix D. Note that while the probit model does not return out-of-bounds predictions, it may be limited in the number of fixed effects that can be included in the estimation. Then, as a robustness check, we also compute the release probabilities using a linear probability model adding court fixed effects. Finally, we also use Lasso to select regressors both considering (i) all interactions and squared terms, and (ii) judge fixed effects. Since the conclusions of the empirical application are equivalent under the different prediction models, we restrict our discussion to the probit model.[32]

Table 11 of Appendix F shows the results for each of the four definitions of Mapuche. Considering 0.5 as the probability threshold, we find that 85% or more 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 59% of the cases, respectively. We also perform an out-of-sample cross-validation exercise that reinforces our confidence in the prediction model.[33]

6.3 Outcome equation

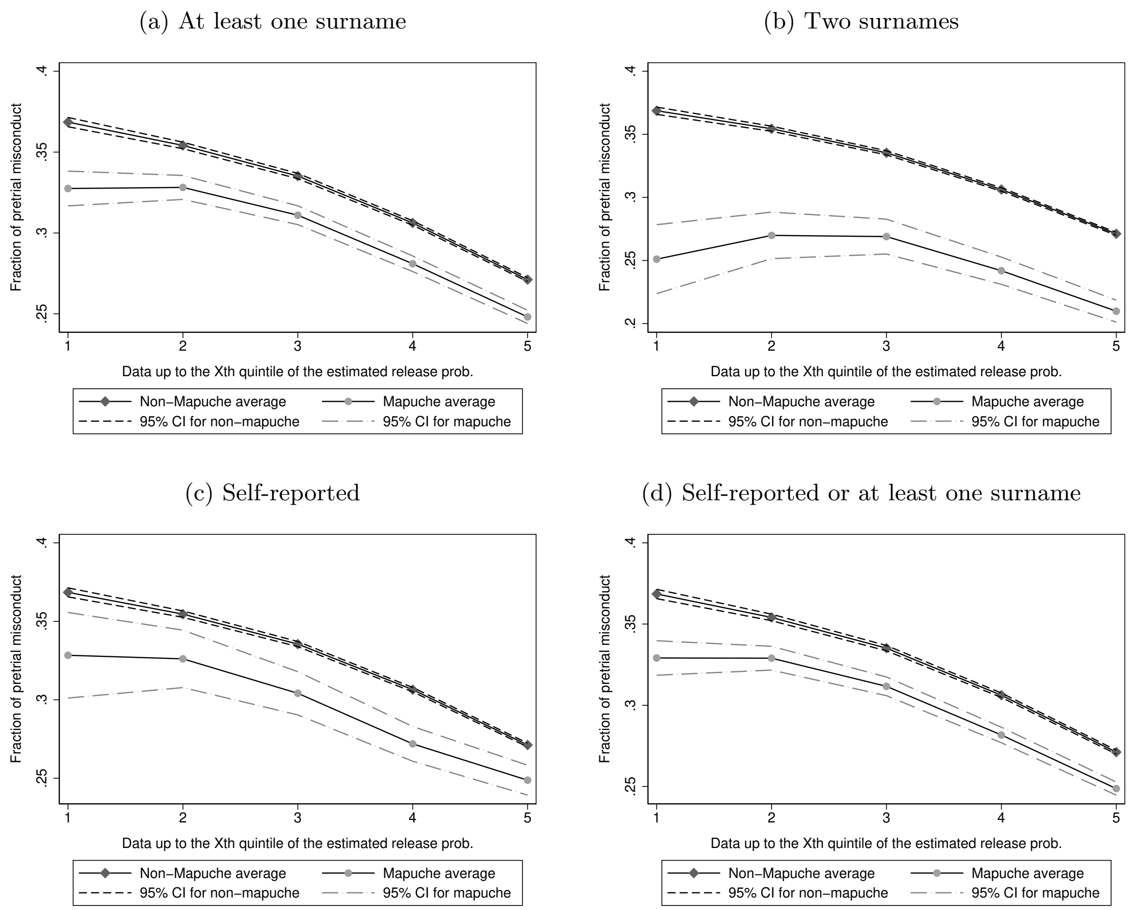
To formally test for prejudice against Mapuche defendants, we use the predicted release probabilities to rank released defendants and build samples of marginal individuals. Recall that under our identification argument, released defendants with lower propensity scores are closer to the margin of release in expectation.

[31] Following Dobbie et al. (2018), the leniency of judge j when reviewing the case of defendant i is calculated by estimating the average pretrial release rate using all cases handled by judge j (except i's), after adjusting for court-by-year fixed effects.

[32] See Table 12 of Appendix F for details on the linear probability model estimation. The P-BOT results using the alternative prediction models can be found in Appendix G.

[33] Specifically, we randomly select 90% of the estimation sample, estimate the probit model, and compute the correct classified cases in the remaining 10%. We repeat the exercise 50 times. On average, 85% of the cases are correctly classified.

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



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 calculated assuming that quintiles are given. Pretrial misconduct accounts for non-appearance in court and/or pretrial recidivism.

As a first exploratory analysis, we analyze how the outcome test varies as we increase the estimation sample. We do this by sequentially adding defendants with a higher predicted probability of being released. 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 (the 20% of released defendants that were closer to the margin of release in probability), then the first and the second quintiles, and so on, until we consider the entire sample. Figure 3 shows the results of this exercise, using the four definitions of Mapuche. The outcome is defined as any pretrial misconduct (non-appearance in court and/or pretrial recidivism).

Three aspects of Figure 3 are worth highlighting. First, the figure provides suggestive evidence of prejudice against the Mapuche. For all Mapuche definitions, the Mapuche defendants' pretrial misconduct rate is below the non-Mapuche defendants' rate in the first quintile of the predicted probability distribution. Second, in all cases, the rates of pretrial misconduct decrease as we add defendants with a higher probability of release. Given that in our model judges decide pretrial detention based on the probability of pretrial misconduct, this result can be thought of as an indirect test of model specification: defendants that are more likely to be released are also less likely to be engaged in pretrial misconduct, which suggests that judges care for expected outcomes when making pretrial detention decisions. We return to this point in Section 7. Finally, within each plot, 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 formal 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 defendants labeled as marginals. 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 in turn non-parametrically calculated at the first percentile of the estimated release probability distribution.[34] We consider two criteria to define the margin (bottom 5% and 10% of the predicted release probability distribution). In both cases, a negative value for the point estimate is evidence of prejudice against Mapuche defendants.

Table 2 shows that all point estimates are negative and statistically significant, providing strong evidence of prejudice against Mapuche defendants. Results are robust to estimating the release probability using a linear model and using Lasso to select covariates, and to considering non-appearance in court and pretrial recidivism as separate outcomes (see Appendix G). Marginally released Mapuche defendants are between 4 and 13 percentage points less likely to be engaged in pretrial misconduct relative to marginal non-Mapuche defendants. Discrimination is larger when we identify Mapuche defendants using both surnames. We conjecture that this is explained by the salience of the ethnicity measure. Finally, consistent with our previous insight of modest potential inframarginality bias, results are similar between the different criteria for defining the margin.

Perturbation test Recall that, depending on the fit of the propensity score, the noise in the ranking estimation may induce bias on the P-BOT. To assess the extent to which this is a threat to the validity of our results, we implement the perturbation test proposed in Section 4.

[34] 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. Observations are weighted using a triangular kernel function.

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

Data up 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):	-0.048	-0.123	-0.087	-0.043
C.I. (95%)	[-0.071, -0.026]	[-0.170, -0.074]	[-0.138, -0.036]	[-0.066, -0.023]
(a) Mapuche expectation	0.305	0.228	0.265	0.309
(b) Non-Mapuche expectation	0.352	0.352	0.352	0.352
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.041	-0.117	-0.079	-0.037
C.I. (95%)	[-0.069, -0.016]	[-0.175, -0.045]	[-0.147, -0.019]	[-0.063, -0.011]
(a) Mapuche expectation	0.322	0.245	0.284	0.326
(b) Non-Mapuche expectation	0.362	0.363	0.363	0.362
No. of Mapuche ($\leq$ 5th pctl.)	1,867	267	310	1,933
No. of Non-Mapuche ($\leq$ 5th pctl.)	27,119	26,989	26,921	27,101
Data up 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):	-0.049	-0.134	-0.058	-0.047
C.I. (95%)	[-0.066, -0.035]	[-0.174, -0.099]	[-0.089, -0.016]	[-0.062, -0.031]
(a) Mapuche expectation	0.307	0.222	0.298	0.309
(b) Non-Mapuche expectation	0.356	0.356	0.356	0.356
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.045	-0.119	-0.071	-0.041
C.I. (95%)	[-0.066, -0.025]	[-0.165, -0.071]	[-0.116, -0.030]	[-0.062, -0.020]
(a) Mapuche expectation	0.312	0.238	0.285	0.315
(b) Non-Mapuche expectation	0.356	0.357	0.356	0.356
No. of Mapuche ($\leq$ 10th pctl.)	3,714	496	587	3,829
No. of Non-Mapuche ($\leq$ 10th pctl.)	54,258	54,015	53,874	54,239

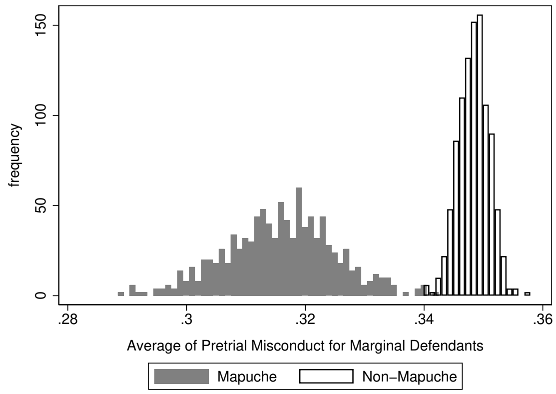
Note: This table presents the results from the P-BOT using the data described in Table 1, considering two approaches to estimate the outcome equation and two criteria to determine who is the margin. Release probabilities are predicted using a probit model. The outcome is any pretrial misconduct. Panel A shows the estimates using a simple difference between the Mapuche and non-Mapuche averages in pretrial misconduct, only considering the individuals whose estimated release probability is lower 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 F. The confidence intervals are calculated using bootstrap with 200 repetitions.

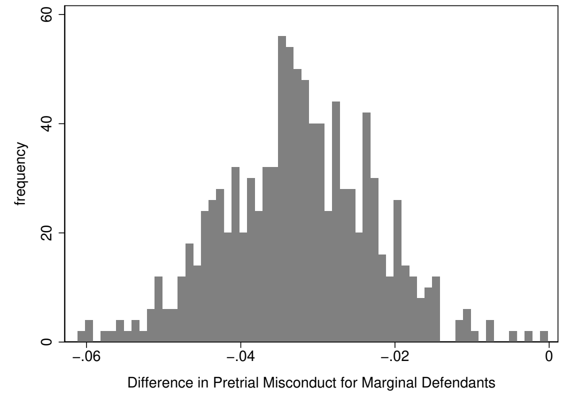
We implement the test using the coefficients of our probit prediction model. For each individual of our sample of released defendants, we simulate 500 realizations from a standardized normal distribution, recompute the ranking, and redefine the sample of marginals. Then, in each of the 500 simulations, we estimate $\mathbb{E}[PM_i|Mapuche_i = 1, Release_i^* = 0]$ and $\mathbb{E}[PM_i|Mapuche_i = 0, Release_i^* = 0]$ by performing a linear regression of pretrial misconduct on a Mapuche indicator in a sample of de-

Figure 4: Perturbation Test (Mapuche: Self-reported or at least one surname)

(a) Mapuche and Non-Mapuche averages

(b) Differences between Mapuche and Non-Mapuche





Note: These plots present the results of the perturbation test described in Section 4. They are produced in the following steps. First, we estimate the probit model (see Table 11). Then, for each released individual in the sample, we simulate 500 realizations from a standardized normal distribution to simulate $Release_i^*$ and redefine the samples of marginal individuals. Within each sample, we estimate $\mathbb{E}[PM_i|Mapuche_i = 1, Release_i^* = 0]$ and $\mathbb{E}[PM_i|Mapuche_i = 0, Release_i^* = 0]$. Panel (a) presents one histogram for each group. Panel (b) presents the histogram for the difference between these two estimated conditional expectations within each simulation.

fendants labeled as marginals. Finally, we plot (i) the group-specific distributions of the conditional expectations, and (ii) the distribution of the difference between the two estimated conditional expectations within each simulation.

Figure 4 shows the results for the most comprehensive definition of Mapuche. Figures 8, 9, and 10 of Appendix G show the results for the other three Mapuche definitions. Reassuringly, both plots suggest that our results are robust to this potential bias. Panel (a) presents the comparison between the group-specific distributions and shows that they do not overlap. Panel (b) presents the distribution of the difference in pretrial misconduct rates between groups and shows that it does not include the zero.[35] This is consistent with (i) the good fit of the propensity score estimation, and (ii) the limited differences in risk distributions suggested in Figure 3.

6.4 Alternative tests

To assess the relative performance between the P-BOT and other approaches, we also test for prejudice using alternative methods. We consider (i) the benchmark test, (ii) the outcome test using the full sample (Knowles et al., 2001), and (iii) the instrument-based approach (Arnold et al.,

[35]Results are consistent across Mapuche definitions. Only the self-reported measure -our least preferred Mapuche indicator- includes the zero in the distribution of the difference. However, the conclusion of the P-BOT is reverted only in 2.2% of the simulations.

Table 3: Alternative Tests for Prejudice

	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Benchmark test:				
Coeff. (dep. variable: release)	-0.005	-0.007	-0.012	-0.005
Robust SE	(0.002)	(0.004)	(0.004)	(0.002)
Observations	693,722	652,502	652,159	694,898
Outcome test (full sample):				
Coeff. (outcome: pretrial misconduct)	-0.023	-0.056	-0.023	-0.022
Robust SE	(0.002)	(0.004)	(0.004)	(0.002)
Observations	694,124	652,500	652,157	694,895
IV-Outcome test:				
Mapuche coeff. (outcome: pretrial misconduct)	0.281	-0.114	1.455	0.091
Mapuche robust SE	(0.404)	(0.268)	(5.303)	(0.338)
Non-Mapuche coeff. (outcome: pretrial misconduct)	0.357	0.357	0.357	0.357
Non-Mapuche robust SE	(0.066)	(0.066)	(0.066)	(0.066)
No. of Mapuche	49,544	7,960	7,733	50,770
No. of non-Mapuche	642,778	642,778	642,778	642,778

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 benchmark test reports the Mapuche coefficient of a probit regression of release on several controls. For details, see Table 11. 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 court level.

2018). For the latter, we exploit the quasi-random assignment of detention judges that characterizes the Chilean setting.[36] Appendix H presents the results of the randomization test suggested by Arnold et al. (2018), validating the random assignment assumption. While Table 18 shows that, after controlling for court-by-time fixed effects, a rich set of observables jointly predicts the release status, Table 19 shows that they do not have joint predictive power for the assigned-judge leniency.

Table 3 presents the results for the alternative methods. The benchmark test provides weak evidence of discrimination. Following the model presented in Section 3, this indicates the presence of statistical discrimination in favor of Mapuche defendants and/or the presence of some relevant omitted variables that are negatively correlated with the Mapuche status. Also, as expected, the outcome test using the full sample provides significant evidence of discrimination. Following Figure 3, the inframarginality bias is biasing the estimation downwards. However, the bias is not large enough to make the test's conclusion misleading.

The most interesting analysis relates to the application of the instrument-based approach. Two things are worth noting from these results. First, while the estimated LATE for the non-Mapuche defendants is precisely estimated, the Mapuche estimations are severely underpowered. For the most

[36]For not violating monotonicity by construction, the instrument is constructed separately for Mapuche and non-Mapuche defendants.

comprehensive indicator of Mapuche, point estimates support the existence of discrimination, but standard errors are large enough to prevent the test from finding significant differences. The case is even more problematic for the less comprehensive indicators: the LATEs are not only extremely noisy but economically meaningless. Therefore, our setting is one in which the instrument-based approach is not well-behaved because of power problems.

Second, recall from Table 2 that the P-BOT estimate of the pretrial misconduct rate of marginally released non-Mapuche defendants is between 35.2 and 36.3%. Notably, the estimated LATE using the instrument-based test in the non-Mapuche sample is 35.7%. Therefore, the estimation of the pretrial misconduct behavior of non-Mapuche marginal defendants is statistically the same between the P-BOT and the instrument-based approach. In addition, in Appendix I we show that the non-Mapuche defendants identified as marginals by both methods have the same distribution of observables. This means that when both approaches for identifying marginal individuals are well behaved both methods provide consistent results. This result is reassuring for the P-BOT, and reinforces our complementarity argument developed throughout the paper.

7 Outcome Test's Assumptions and Interpretation

In Section 4 we introduced a new way of implementing the outcome test, the P-BOT, and in Section 6 we used this approach to test for prejudice in pretrial detention decisions against the main ethnic group in Chile, the Mapuche. For the sake of simplicity, and to focus the attention on what is novel about our method, we have presented this new approach using a stylized model. However, this model has some assumptions that are not always explicitly discussed in the related literature. This section discusses these assumptions and their implications.

In our view, these assumptions are not very restrictive given that the null hypothesis of no prejudice remains valid in more general versions of the model. However, we argue that these assumptions affect the results' interpretation. Intuitively, they are related to the relevant counterfactual when the null hypothesis is rejected. In this section we present additional evidence to discuss different interpretations of our results depending on the assumptions of the underlying model and sketch a taxonomy of prejudice based on the analysis.

Selection process The starting point of the model presented in Section 3 is that judges make (or, at least, should make) decisions based on expected pretrial misconduct. However, it could be the case that judges have different objective functions and, therefore, are not looking at potential outcomes when making decisions.

Regarding this, two things are worth mentioning. First, the P-BOT implementation provides suggestive evidence about the extent to which the potential outcome is being used to make selection decisions. Recall from Figure 3 that, for all groups, pretrial misconduct rates decrease when inframarginal defendants are included in the sample. Given our identification strategy, this means that released defendants with larger release probabilities are less likely to be engaged in pretrial misconduct. This suggests that judges care, on average, about potential pretrial misconduct when deciding who to release.

Second, even in the absence of this suggestive evidence, the nature of the potential deviations from the potential-outcome-based selection process will determine the implications for the validity and interpretation of the test. To see why, consider the following two examples. In the first one, judges are mandated by law to make decisions based on potential pretrial misconduct, but they do not respect the law and base their decisions on different outcomes. For example, they try to please their bosses to increase their likelihood of being promoted. Then, if their bosses are racists, they will release white defendants and detain Mapuche defendants regardless of their predicted risk. In our view, the outcome test is still valid in this case to provide evidence of prejudice, since some defendants may be discriminated against with respect to the normative standard provided by law. The correct interpretation of what prejudice is in this setting, however, will depend on the specific outcome judges are looking at. We define *incentive-driven* prejudice when a group is systematically discriminated against because of the alternative outcomes judges are using to make decisions. Since we do not have additional tests for discerning between alternative objective functions judges may have, and consistent with the indirect evidence displayed in Figure 3, we stick to the standard interpretation provided by the model presented in Section 3.

This is not necessarily true in all settings. In a second example, consider an institutional setting that mandates by law to dictate pretrial detention to all defendants that have prior convictions regardless of the characteristics of the current prosecution. In this case, the unbiased selection process has different implications on observed pretrial misconduct as long as the distribution of prior convictions varies by group. In a case like this, the model should be modified in order to derive the relevant testable equations.[37] Given that in the Chilean setting judges are mandated by law to make decisions based on potential pretrial misconduct, we argue that this is not a threat for the validity of our results.

[37] An interesting application of this deviation is Manski (2005, 2006). The author proposes a model of optimal police profiling where, if the deterrence effects of police search vary by group, then the effective thresholds may be optimally different for reasons unrelated to discrimination.

The bottom line of this discussion is that whenever the mandated selection process is well defined, then individual deviations from it do not affect the validity of the outcome test. An interesting avenue of future research is to provide additional tests to discern between different models of actual judge behavior.

Determinants of judges' thresholds The model of Section 3 assumes that the only variable that can determine judges' thresholds, in addition to judge-specific leniency measures, is group membership. In our empirical application, this implied that we only tested for significant differences in effective thresholds between Mapuche and non-Mapuche defendants.

However, in reality, prejudice patterns can be non-binary, implying that effective thresholds can be also influenced by other variables. Consider, for example, that judges discriminate based on place of living, and that place of living is correlated with being Mapuche. In our opinion, the outcome test remains valid in this context since it is still the case that Mapuche defendants are more frequently imprisoned for reasons that are not related with their probability of pretrial misconduct. We frame this situation as *correlation-driven* prejudice. We come back to this discussion at the end of the section. In addition, there could be complementarities in discrimination patterns between categories. For examples, the prejudice against Mapuche defendants could depend on the socioeconomic status of the defendant.

Two remarks are worth to mention. First, allowing for more determinants of the threshold does not have any consequence in the ability of the P-BOT to predict which released defendants are closer to the margin, as long as potential unobservables affecting the effective threshold (if any) are separable from the observables (i.e., if the more complex effective thresholds do not violate A1). By contrast, this extension induces a complication for the instrument-based approach. More complex patterns of discrimination may induce violations of the LATE monotonicity assumption, even if the additional determinants of the effective threshold are observable. Second, we can use the P-BOT to test for the relevance of additional covariates in the determination of effective thresholds when they are observable. Recall that the outcome test is, essentially, computing conditional expectations within a truncated sample. By adding observables to the linear regression that characterizes the outcome equation, we can test for the existence of more complex patterns of prejudice.

To illustrate that the latter insight can lead to interesting analyses, Table 4 presents two examples of this extension. In Panel A, we consider two categories (and its interaction) to group defendants: *Mapuche* and *Low income*. The latter is calculated using the Chilean national household survey (CASEN), such that *Low income* is equal to one if the defendant lives in a municipality

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	-0.012	Mapuche	-0.034
C.I. (95%)	[-0.045, 0.023]	C.I. (95%)	[-0.056, -0.012]
Low income	0.011	Mapuche region	-0.034
C.I. (95%)	[-0.002, 0.024]	C.I. (95%)	[-0.074, 0.003]
Mapuche and low income	-0.058	Mapuche and mapuche region	-0.047
C.I. (95%)	[-0.098, -0.010]	C.I. (95%)	[-0.116, 0.027]
Pretrial misconduct expectation for:		**Pretrial misconduct expectation for:**	
Mapuche and low income	0.254	Mapuche and mapuche region	0.238
Non-Mapuche and low income	0.324	Non-Mapuche and mapuche region	0.319
Mapuche and high income	0.301	Mapuche and non-mapuche region	0.319
Non-Mapuche and high income	0.313	Non-Mapuche and non-mapuche region	0.353
Observations:		**Observations:**	
Mapuche and low income	867	Mapuche and mapuche region	223
Non-Mapuche and low income	11,434	Non-Mapuche and mapuche region	668
Mapuche and high income	741	Mapuche and non-mapuche region	1,708
Non-Mapuche and high income	10,659	Non-Mapuche and non-mapuche region	26,437

Note: This table presents the results of the P-BOT considering additional categories to group defendants. 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 IX Region (which is the region with more Mapuche individuals and where the political conflict is more salient). These models use the data described in Table 1 (in the case of Panel A without data for 2017 because those defendants do not have information on their place of living). 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 (a linear regression of pretrial misconduct on the indicators and their interaction) and considering the released individuals whose estimated release probability is lower or equal to the 5th percentile. Details of the covariates included in the prediction model can be found in Appendix F. The confidence intervals are calculated using bootstrap with 200 repetitions.

whose average income is below the median.[38] 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 region that has historically hosted the conflict between the Mapuche people and the Chilean government (*IX Región de la Araucanía*). We show results for the simple version of the P-BOT for the most comprehensive Mapuche definition, and using the 5% margin definition.[39]

This table shows that prejudice patterns are more complex than the binary model case. This is clear when looking at the differences in the four conditional means. Regarding the first example, results show that prejudice against Mapuche defendants is mainly relevant for those Mapuche who live in low-income municipalities. Interestingly, there is no significant evidence of prejudice solely based on the place of living. This suggests that the relevant prejudice is against poor Mapuche defendants. Regarding the second example, as expected, results suggest that Mapuche defendants

[38]To have groups of similar sizes, the median is defined considering the sample of defendants.

[39]In addition to some random missing data, we do not observe defendants' place of living in the last year of our sample. That is why the sample size is smaller in the first example.

are more discriminated in the conflict region. These results suggest that more complex patterns of discrimination are likely to happen in practice.

Assignment rule of judges Finally, note that to derive the outcome test and describe the P-BOT, we did not make any assumption on $j(i)$. Usually, the empirical literature focuses on cases where $j(i)$ is characterized by random assignment of judges to defendants since those rules provide useful exogenous variation. However, despite its usefulness, it is clear from Sections 3 and 4 that random assignment of judges is not needed for the validity of the outcome test nor for the identification strategy of our approach. As in the previous discussions, we argue that deviations from random assignment only affect the interpretation of the outcome test results.

To see why, consider the following two polar cases. In the first one, judges are completely unbiased (hence the only variation in the effective thresholds comes from heterogeneity in judges' idiosyncratic leniency) but stricter judges are systematically assigned to Mapuche defendants. In the second one, judges are biased against Mapuche defendants, but they are randomly assigned to defendants. Note that in both cases the outcome test will provide evidence of prejudice against Mapuche defendants, but with different interpretations. In the first case, the evidence can be interpreted as *systemic* prejudice (i.e., $j(i)$ is prejudiced). In the second case, the interpretation is the traditional one, namely, judges are on average prejudiced against Mapuche (*pure* prejudice). This adds a layer to the taxonomy analysis we develop at the end of this section.

When information of judges is available, the relevance of these two types of prejudice can be tested. On one side, regressions of judge leniency on defendants' characteristics are indicative of systematic correlations between judge leniency and observables. This is, in fact, the intuition behind randomization tests performed in studies relying on judges designs (see, for example, Arnold et al., 2018; Dobbie et al., 2018; Cohen and Yang, 2019). On the other side, in settings like ours where judges are randomly assigned at the court-by-time level, implementing our simple P-BOT regression but controlling for court-by-year fixed effects will yield an estimate for prejudice net of systemic discrimination. This is what we present in Table 5. Results are very similar to the baseline regressions, suggesting that systemic prejudice is not the main driver behind our results.

Towards a taxonomy of prejudice Throughout this section we have argued that the validity of the outcome test (and, in particular, of the P-BOT) as a diagnostic of prejudice is not affected by (i) the effective outcomes judges look at when making decisions (as long as they are required to consider potential pretrial misconduct), (ii) how many variables judges consider to set their

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

Data up 5th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Point estimate, (a)-(b):	-0.042	-0.107	-0.080	-0.038
C.I. (95%)	[-0.063, -0.021]	[-0.154, -0.059]	[-0.117, -0.027]	[-0.060, -0.017]
(a) Mapuche expectation	0.313	0.247	0.292	0.317
(b) Non-Mapuche expectation	0.353	0.353	0.353	0.353
No. of Mapuche ($\leq$ 5th pctl.)	1,866	265	312	1,930
No. of Non Mapuche ($\leq$ 5th pctl.)	27,120	26,991	26,923	27,106
Data up 10th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Point estimate, (a)-(b):	-0.045	-0.125	-0.060	-0.043
C.I. (95%)	[-0.061, -0.030]	[-0.160, -0.092]	[-0.095, -0.015]	[-0.057, -0.026]
(a) Mapuche expectation	0.308	0.228	0.313	0.310
(b) Non-Mapuche expectation	0.351	0.351	0.351	0.351
No. of Mapuche ($\leq$ 10th pctl.)	3,700	496	587	3,820
No. of Non Mapuche ($\leq$ 10th pctl.)	54,273	54,015	53,874	54,248

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. Release probabilities are predicted using a probit model. The outcome is any pretrial misconduct. The estimation only considers the simple approach: a regression of pretrial misconduct on dummies for Mapuche. Details of the covariates included in the prediction model can be found in Appendix F. The confidence intervals are calculated using bootstrap with 200 repetitions.

thresholds (as long as more complex effective thresholds involving unobservables do not violate A1), and (iii) how judges are assigned to defendants. However, we have argued that these factors do affect the interpretation of the test.

This leads to a taxonomy of prejudice. While the outcome test makes a robust identification of prejudice (understood as systematically facing different effective thresholds for reasons unrelated with pretrial detention risk), the overall bias can be thought as a combination of four different sources: *pure* prejudice (biased predictions and/or animus), *incentive-driven* prejudice (when judges look at other outcomes to make decisions that ultimately harm specific groups), *correlational-driven* prejudice (when other variables that correlate with group traits matter for the effective threshold definition), and *systemic* prejudice (when the assignment rule systematically imposes stricter conditions to specific groups).

As we discussed in this section, part of this distinction can be empirically tested using the P-BOT. Yet, since the policy implications may vary with the specific sources of the overall-effect, we see the development of additional empirical strategies to discern between sources of prejudice as a relevant avenue of future research. In the absence of that, the researcher should be aware of this taxonomy when interpreting empirical results.

8 Conclusion

Although economists are aware of the virtues of outcome tests to test for prejudice since the theoretical contributions of Becker (1957, 1993), their implementation is not straightforward. In particular, outcome tests need to identify marginal individuals. In most settings, that condition is not observable.

In this paper, we propose a novel method for identifying marginal individuals to implement the outcome test – the Prediction-Based Outcome Test (P-BOT). We motivate our framework with a model of pretrial detentions decisions. Our main result provides sufficient conditions under which released defendants that are more likely to be marginal given their observables also have smaller propensity scores. We develop a detailed discussion about the restrictiveness of our assumptions and argue that they have relative weaknesses and strengths with respect to the literature. Given that, we claim that the P-BOT is a complementary methodology to the already existing approaches.

Our identification strategy considerably simplifies the implementation of the outcome test. The econometrician can proceed by fitting projection models for the release status, ranking released defendants according to their predicted probabilities, defining samples of marginally released defendants, and performing simple outcome equations. The non-trivial challenge of identifying marginally released individuals is reduced to a standard prediction problem. Hence, the P-BOT relies on the availability of good predictors for the release status. The increasing availability of rich administrative datasets suggests that this could not be a strong requirement.

We use the P-BOT to test for prejudice in pretrial detentions against the main ethnic group in Chile, the Mapuche, using nationwide administrative data. We find strong evidence of prejudice using different outcome variables, Mapuche definitions, and estimation methods both in the projection and outcome equations. Also, we illustrate the relative performance of different available diagnostics for prejudice. We provide evidence of modest inframarginality bias and show that the instrument-based approach has implementation issues in our setting. While the estimated LATE for the non-Mapuche sample is precise and exactly matches the P-BOT estimates of pretrial misconduct of marginal defendants, estimations using the Mapuche defendants are severely underpowered.

In the last section of the paper, we discuss the interpretation issues that arise when relaxing some important assumptions of the standard model. In this regard, we show that discrimination patterns are likely to be more complex than commonly assumed. We sketch a novel taxonomy of prejudice and suggest avenues for future research.

We want to end the discussion by stressing that the underlying model and the outcome test are useful frameworks for analyzing prejudice in a variety of contexts. In fact, the original ideas of Gary Becker that gave form 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 outcome (ex-post measurable). 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. To illustrate this point, we provide two examples where random assignment seems very unlikely.

Example 1: Tenure decisions There are documented gender gaps in the economics academia, in particular concerning tenure decisions (Antecol et al., 2018; Lundberg and Stearns, 2019; Sarsons et al., 2020). In this setting, prejudice means that, when making tenure decisions, higher standards are set and/or potential productivity is underestimated, for women. If marginally tenured women have higher post-tenure productivity (e.g., quality-weighted publications after five years) than marginally tenured men, then the outcome test states that the assignment of tenure is prejudiced against women. The identification of marginally tenured individuals can be achieved by fitting a projection model of tenure status on observables (demographics, pre-tenure publication record, graduate program attended, etc.).

Example 2: Credit allocations As in Blanchflower et al. (2003) and Dobbie et al. (2018), a researcher could be interested in testing for discrimination in financial markets. For example, suppose banks are biased against entrepreneurs of a certain age, race, gender, or educational level. In this setting, prejudice means that lenders set higher profitability thresholds and/or underestimate the expected potential profitability of businesses, for entrepreneurs who belong to the analyzed groups. If marginal credit awardees of the analyzed groups are less likely to become bankrupt, then the outcome test states that there is prejudice against that group. The identification of marginal credit awardees can be achieved by fitting a projection model of getting a credit on observables (demographics, financial record, characteristics of the business, etc.).

References

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

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

Aizer, A. and J. J. Doyle Jr (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics 130*(2), 759–803.

Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. *Handbook of Labor Economics 3*, 3143–3259.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political economy 113*(1), 151–184.

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

Antecol, H., K. Bedard, and J. Stearns (2018). Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? *American Economic Review 108*(9), 2420–41.

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

Anwar, S., P. Bayer, and R. Hjalmarsson (2012). The impact of jury race in criminal trials. *The Quarterly Journal of Economics 127*(2), 1017–1055.

Anwar, S., P. Bayer, and R. Hjalmarsson (2018). Politics in the courtroom: Political ideology and jury decision making. *Journal of the European Economic Association 17*(3), 834–875.

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

Arnold, D., W. Dobbie, and P. Hull (2020). Measuring racial discrimination in bail decisions. *Working Paper*.

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, January). The causal impact of removing children from abusive and neglectful homes. *Working paper*.

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

Becker, G. (1993). Nobel Lecture: The economic way of looking at behavior. *Journal of Political Economy 101*, 385–409.

Benson, A., D. Li, and K. Shue (2019). Promotions and the Peter principle. *The Quarterly Journal of Economics 134*(4), 2085–2134.

Bertrand, M. and E. Duflo (2017). Field experiments on discrimination. In *Handbook of Economic Field Experiments*, Volume 1, pp. 309–393. Elsevier.

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

Blanchflower, D. G., P. B. Levine, and D. J. Zimmerman (2003). Discrimination in the small-business credit market. *Review of Economics and Statistics 85*(4), 930–943.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Cayul, P., A. Corvalan, D. Jaimovich, and M. Pazzona (2018). Maceda: A new events data set on the self-determination conflict between the mapuche indigenous group and the chilean state (1990-2016). *Mapuche Data Project - Working paper*.

Chandra, A. and D. O. Staiger (2010). Identifying provider prejudice in healthcare. *Working paper*.

Chetty, R., N. Hendren, M. R. Jones, and S. R. Porter (2018). Race and economic opportunity in the united states: An intergenerational perspective. *Working Paper*.

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

Cortés, T., N. Grau, and J. Rivera (2019). Juvenile incarceration and adult recidivism. *Working paper*.

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014, 08). Family welfare cultures. *The 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.

Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2018). Measuring bias in consumer lending. *Working paper*.

Frandsen, B. R., L. J. Lefgren, and E. C. Leslie (2019). Judging judge fixed effects. *Working paper*.

Fryer, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy 127*(3), 1210–1261.

Grau, N., G. Marivil, and J. Rivera (2019). The effect of pretrial detention on labor market outcomes. *Working paper*.

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

Heckman, J. and E. Vytlacil (1999). Local instrumental variables and latent variable models for identifying bounding treatment effects. *Proceedings for the National Academy of Sciences 96*, 4730–4734.

Heckman, J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*, 669–738.

Imbens, G. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*, 467–475.

Kleinberg, J., J. Ludwig, S. Mullainathan, and C. R. Sunstein (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis 10*.

Kline, P. and C. R. Walters (2016). Evaluating public programs with close substitutes: The case of head start. *The Quarterly Journal of Economics 131*(4), 1795–1848.

Kline, P. and C. R. Walters (2020). Reasonable doubt: Experimental detection of job-level employment discrimination. *Working paper*.

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.

Lang, K. and J.-Y. K. Lehmann (2012). Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature 50*(4), 959–1006.

Lundberg, S. and J. Stearns (2019). Women in economics: Stalled progress. *Journal of Economic Perspectives 33*(1), 3–22.

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

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

Marx, P. (2018). An absolute test of racial prejudice. *Working paper*.

Merino, M. and D. Quilaqueo (2003). 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 27*(4), 105–116.

Merino, M. E. and D. J. Mellor (2009). Perceived discrimination in Mapuche discourse: Contemporary racism in Chilean society. *Critical Discourse Studies 6*(3), 215–226.

Muller-Smith, M. (2015). The criminal and labor market impacts of incarceration. *Working paper*.

Norris, S. (2020). Examiner inconsistency: Evidence from refugee decisions. *Working paper*.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics 37*(2), 187–204.

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

Pierson, E., S. Corbett-Davies, and S. Goel (2017). Fast threshold tests for detecting discrimination. *Working paper*.

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

Rose, E. (2020). Who gets a second chance? effectiveness and equity in supervision of criminal offenders. *Working paper*.

Sarsons, H., K. Gerxhani, E. Reuben, and A. Schram (2020). Gender differences in recognition for group work. *Working paper*.

Simoiu, C., S. Corbett-Davies, S. Goel, et al. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics 11*(3), 1193–1216.

Yang, C. and W. Dobbie (2020). Equal protection under algorithms: A new statistical and legal framework. *Michigan Law Review*.

A Proofs

PROPOSITION I:

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

Proof. $Release_i^* = 0$ implies that $\mathbb{E}[PM_{i1}|G_i = g, X_i, V_i] = h(g, j(i))$. Then

$$\begin{aligned}
\mathbb{E}[PM_i|G_i = g, j(i), Release_i^* = 0] &= \mathbb{E}[PM_{i1}|G_i = g, j(i), Release_i^* = 0], \\
&= \mathbb{E}[\mathbb{E}[PM_{i1}|G_i = g, j(i), X_i, V_i, Release_i^* = 0]|G_i = g, j(i), Release_i^* = 0], \\
&= \mathbb{E}[\mathbb{E}[PM_{i1}|G_i = g, j(i), X_i, V_i, \mathbb{E}[PM_{i1}|G_i = g, X_i, V_i] = h(g, j(i))]|G_i = g, j(i), Release_i^* = 0], \\
&= \mathbb{E}[h(g, j(i))|G_i = g, j(i), Release_i^* = 0], \\
&= \mathbb{E}[h(g, j(i))|G_i = g, j(i)], \\
&= h(g, j(i)).
\end{aligned}$$

The proof is concluded by taking expectation among marginal defendants with $G_i = g$. □

PROPOSITION II: *Let x_1 and x_2 be two possible realizations of X_i, and $\epsilon > 0$ be a small distance to the margin of release. Under A1 and A2,*

$$\begin{aligned}
\Pr\left(Release_i^* \leq \epsilon | X_i = x_1, Release_i = 1\right) &> \Pr\left(Release_i^* \leq \epsilon | X_i = x_2, Release_i = 1\right) \\
\iff \qquad \mathbb{E}\left[Release_i | X_i = x_1\right] &< \mathbb{E}\left[Release_i | X_i = x_2\right].
\end{aligned}$$

Proof. Under A1, $1\{f(X_i, V_i) \geq 0\} = 1\{d(X_i) - g(V_i) \geq 0\} \equiv 1\{d(X_i) - W_i \geq 0\}$. Let $W_i = \mathbb{E}[W_i|X_i] + \widetilde{W}_i$, and $n(X_i) = d(X_i) - \mathbb{E}[W_i|X_i]$. Then

$$\begin{aligned}
\Pr\left(Release_i^* \leq \epsilon | X_i, Release_i = 1\right) &= \Pr\left(d\left(X_i\right) - W_i \leq \epsilon | X_i, W_i \leq d\left(X_i\right)\right), \\
&= \Pr\left(d\left(X_i\right) - \mathbb{E}[W_i|X_i] - \widetilde{W}_i \leq \epsilon | X_i, \mathbb{E}[W_i|X_i] + \widetilde{W}_i \leq d\left(X_i\right)\right), \\
&= \frac{\Pr\left(n\left(X_i\right) - \epsilon \leq \widetilde{W}_i \leq n\left(X_i\right) | X_i\right)}{\Pr\left(\widetilde{W}_i \leq n\left(X_i\right) | X_i\right)}, \\
&= \frac{\Theta_{\widetilde{W}|X}\left(n\left(X_i\right) | X_i\right) - \Theta_{\widetilde{W}|X}\left(n\left(X_i\right) - \epsilon | X_i\right)}{\Theta_{\widetilde{W}|X}\left(n\left(X_i\right) | X_i\right)}, \\
&\approx \frac{\epsilon \cdot \theta_{\widetilde{W}|X}(n(X_i)|X_i)}{\Theta_{\widetilde{W}|X}\left(n\left(X_i\right) | X_i\right)},
\end{aligned}$$

where $\Theta_{\widetilde{W}|X}$ and $\theta_{\widetilde{W}|X}$ are the conditional cdf and density of $\tilde{W}_i$, and the last step comes from taking a first order approximation around $\epsilon = 0$, so $\Theta_{\widetilde{W}|X}\left(n\left(X_i\right) - \epsilon | X_i\right) \approx \Theta_{\widetilde{W}|X}\left(n\left(X_i\right)|X_i\right) - \epsilon \cdot \theta_{\widetilde{W}|X}(n(X_i)|X_i)$. The independence condition in A2 implies that

$$\frac{\epsilon \cdot \theta_{\widetilde{W}|X}(n(X_i)|X_i)}{\Theta_{\widetilde{W}|X}\left(n\left(X_i\right)|X_i\right)} = \frac{\epsilon \cdot \theta_{\widetilde{W}}(n(X_i))}{\Theta_{\widetilde{W}}\left(n\left(X_i\right)\right)},$$

and the log-concavity of the distribution of $\widetilde{W}_i$ implies that $\theta_{\widetilde{W}}(n(X_i))/\Theta_{\widetilde{W}}(n(X_i))$ is a decreasing function of $n(X_i)$. Then, noting that $\Theta_{\widetilde{W}}\left(n\left(X_i\right)\right) = \mathbb{E}[Release_i|X_i]$ is an increasing function of $n(X_i)$, for arbitrary realizations of X_i, x_1 and x_2 we conclude that

$$\begin{aligned} \Pr\left(Release_i^* \leq \epsilon | X_i = x_1, Release_i = 1\right) \quad &> \quad \Pr\left(Release_i^* \leq \epsilon | X_i = x_2, Release_i = 1\right) \\ \iff \qquad \mathbb{E}\left[Release_i | X_i = x_1\right] \quad &< \quad \mathbb{E}\left[Release_i | X_i = x_2\right]. \quad \square \end{aligned}$$

B Model's Extensions

In this appendix we provide two extensions to the model presented in Section 3. First, we maintain the binary outcome, PM_{i1}, but change the decision rule to depend on judge-specific predictions of the latent risk index, $m(X_i, V_i, \nu_i)$. Second, we extend the model to a continuous outcome. This is relevant to assess whether the model presented is useful to analyze alternative settings where the outcome is non-binary.

Binary outcome and latent risk prediction The selection rule is given by

$$Release_i \quad = \quad 1\{\mathbb{E}_{\nu}[m(X_i, V_i, \nu_i)|G_i, X_i, V_i] \leq h(G_i, j(i))\}. \tag{10}$$

In this case, the outcome test has to be slightly modified. As in Section 4, is will be useful to define $m(X_i, V_i, \nu_i) = \mathbb{E}[m(X_i, V_i, \nu_i)|F_i, X_i, V_i] - \xi_i$.

Proposition (Modified Outcome Test): *Let $F_{\xi|G,J,X,V}$ be the conditional cumulative distribution of ξ. Assume $F_{\xi|G,J,X,V} = F_{\xi}$ and that F_{ξ} is strictly increasing. Then*

$$\overline{h}(0) > \overline{h}(1) \quad \Longleftrightarrow \quad \mathbb{E}[PM_i|G_i = 0, Release_i^* = 0] > \mathbb{E}[PM_i|G_i = 1, Release_i^* = 0].$$

Same logic applies if $\overline{h}(0) = \overline{h}(1)$ or $\overline{h}(0) < \overline{h}(1)$.

Proof. We have that

$$\begin{aligned}
\mathbb{E}[PM_i|G_i = g, j(i), Release_i^* = 0] \quad &= \quad \mathbb{E}[PM_i|G_i = g, j(i), Release_i^* = 0], \\
&= \quad \mathbb{E}\left[\mathbb{E}\left[PM_i|G_i = g, j(i), Release_i^* = 0, X_i, V_i\right]|G_i = g, Release_i^* = 0\right], \\
&= \quad \mathbb{E}[F_{\xi}(h(G_i))|G_i = g, Release_i^* = 0], \\
&= \quad F_{\xi}(h(g)).
\end{aligned}$$

Then

$$\mathbb{E}[PM_i|G_i = 1, Release_i^* = 0] > \mathbb{E}[PM_i|G_i = 0, Release_i^* = 0] \Longleftrightarrow F_{\xi}(h(1)) > F_{\xi}(h(0)) \Longleftrightarrow h(1) > h(0). \quad \square$$

Two things are worth highlighting about this result. First, we need additional assumptions to derive the outcome test. However, these assumptions overlap with the sufficient conditions for our identification strategy. Importantly, as it is discussed in the main text, we see the assumptions on ξ_i as second-order. Second, in this case the outcome test no longer has a cardinal interpretation. Recall that in the model presented in Section 3, pretrial misconduct rates of marginally released defendants exactly recover effective thresholds. Then, the difference in means have a structural interpretation. In the modified outcome test proposition, the difference in means identify prejudice, and their relative magnitude informs relative prejudice. However, to map differences in means into actual thresholds the researcher needs to specify F_ξ. Both interpretations coincide when $\xi \sim \mathcal{U}[0, 1]$.

Continuous outcome In this case, the selection rule is given by

$$Release_i \quad = \quad 1\{\mathbb{E}_\nu[Y_{i1}|G_i, X_i, V_i] \leq h(G_i, j(i))\}, \tag{11}$$

where Y_{i1} is the potential continuous outcome if selected. For example, in a model of prejudice in promotion decisions, Y could measure productivity. Assume that $Y_{i1} = m(X_i, V_i, \nu_i)$. In this case, the outcome test is exactly the same as in Proposition I. The proof displayed in Appendix A can be reproduced by changing PM_{i1} by Y_{i1}. While this extension does not add any additional insight, if makes explicit how this model can be used to analyze selection processes that depend on more general outcome definitions.

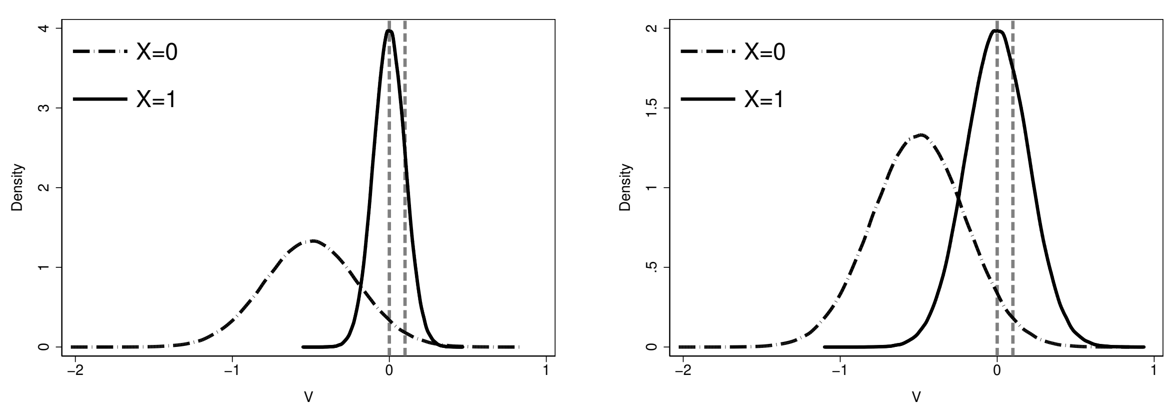
C Understanding A2

As we discuss in the paper, A2 is a sufficient but not necessary condition. In this appendix we provide examples of distributions that violate A2 and illustrate how extreme the deviations have to be in order to invalidate our identification argument.[40]

Example 1 Let $X_i \in \{0, 1\}$ and $V_i | X_i = x \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Assume $Release_i = 1\{V_i \geq 0\}$, so A1 is trivially satisfied. If $\sigma_0 \neq \sigma_1$, A2 is violated. Figure 5 displays the conditional densities for simulated data using two set of parameters. We set $\mu_0 = -0.5$, $\mu_1 = 0$, and $\sigma_0 = 0.3$. Panel (a) sets $\sigma_1 = 0.1$, and panel (b) sets $\sigma_1 = 0.2$. We simulate 1,000,000 of observations and define marginally released individuals as individuals with $V_i \in [0, 0.1]$. Intuitively, heteroskedasticity may affect the propensity score-based ranking since individuals with $X_i = 1$ have a disproportionate mass of marginal individuals. Note that although $Release_i^*$ does not explicitly depends on X_i, it is still the case that X_i is a good predictor given the high correlation with V_i.

Figure 5: Example 1

(a) $\sigma_1 = 0.1$ (b) $\sigma_1 = 0.2$



Since $\mu_0 < \mu_1$, $\mathbb{E}[Release_i | X_i = 0] < \mathbb{E}[Release_i | X_i = 1]$. However, the share of released defendants that are marginal are not necessarily decreasing in the propensity score. This depends on the variances. Table 6 shows that the identification argument is violated in the simulation displayed in panel (a), but not in the one displayed in panel (b). In this unfavorable case (since one normal is located exactly at the margin definition), variances need to be very different to bias the ranking.

[40]We thank Chris Walters for suggesting these examples.

Table 6: Propensity score and share of marginally released defendants

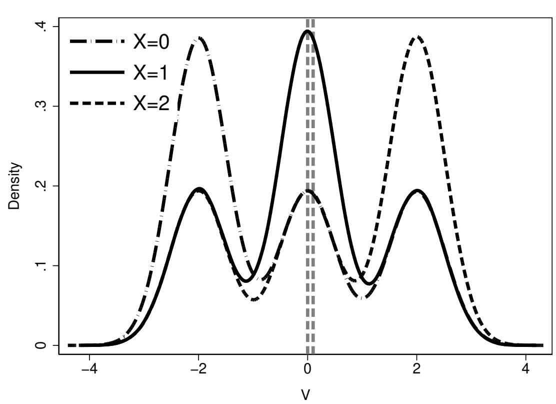
	$\sigma_1 = 0.1$		$\sigma_1 = 0.2$	
X_i	$\mathbb{E}[R_i\|X_i]$	$\Pr(V_i < 0.1\|X_i, R_i = 1)$	$\mathbb{E}[R_i\|X_i]$	$\Pr(V_i < 0.1\|X_i, R_i = 1)$
0	0.05	52.1%	0.05	52.1%
1	0.5	68.2%	0.5	38.1%

Example 2 As a second example, assume that V_i is determined by a mixture of normals. Let $X_i \in \{0, 1, 2\}$ and $V_i = -K \cdot 1\{T_i = 0\} + 0 \cdot 1\{T_i = 1\} + K \cdot 1\{T_i = 2\} + u_i$, where T_i are types, and $u_i \sim \mathcal{N}(0, \sigma^2)$. We consider two conditional distributions of types. In the first one (*Distr. 1*), $\Pr[T_i = 0|X_i = 0] = 0.5$, $\Pr[T_i = 0|X_i = 1] = 0.25$, $\Pr[T_i = 0|X_i = 2] = 0.25$, $\Pr[T_i = 1|X_i = 0] = 0.25$, $\Pr[T_i = 1|X_i = 1] = 0.5$, and $\Pr[T_i = 1|X_i = 2] = 0.25$.[41] In the second one (*Distr. 2*), $\Pr[T_i = 0|X_i = 0] = 0.5$, $\Pr[T_i = 0|X_i = 1] = 0.25$, $\Pr[T_i = 0|X_i = 2] = 0.15$, $\Pr[T_i = 1|X_i = 0] = 0.35$, $\Pr[T_i = 1|X_i = 1] = 0.5$, and $\Pr[T_i = 1|X_i = 2] = 0.35$. The selection rule and the margin definition are as in Example 1.

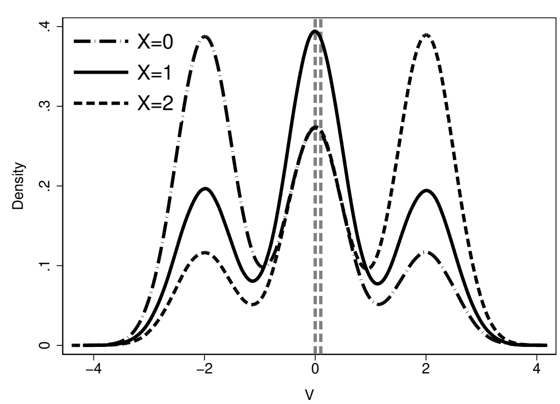
Figure 6 displays the conditional densities of V_i. We set $K = 2$ and $\sigma = 0.5$. Since the distribution of V_i is multi-modal, the conditional densities are concentrated at different values depending on the realization of X_i. The main difference between both distributions is that in the first one the non-modal mass is equally distributed in the other two mass-points while in the second one the mass is proportional to the distance to the mode. As before, these distributions violate A2. They are heteroskedastic and their higher moments vary with X_i.

Figure 6: Example 2

(a) Distr. 1



(b) Distr. 2



[41] Note that $\Pr[T_i = 2|X_i = x] = 1 - \Pr[T_i = 0|X_i = x] - \Pr[T_i = 1|X_i = x]$.

As it is shown in Table 7, in the first distribution the share of marginals is not decreasing in the propensity score, while the second distribution preserves the monotone behavior. Then, the low release probability distribution must to have a significant part of the mass in large unobserved realizations in order to bias the ranking.

Table 7: Propensity score and share of marginally released defendants

	Distr. 1		Distr. 2	
X_i	$\mathbb{E}[R_i\|X_i]$	$\Pr(V_i < 0.1\|X_i, R_i = 1)$	$\mathbb{E}[R_i\|X_i]$	$\Pr(V_i < 0.1\|X_i, R_i = 1)$
0	0.38	5.3%	0.33	8.6%
1	0.5	8.1%	0.5	8.1%
2	0.62	3.2%	0.67	4.2%

Discussion In both examples, the distributions that showed more severe deviations from A2 induced problems in the propensity score-based ranking. This suggests that the differences in the higher moments should be large and somewhat counterintuitive in order to violate the identification argument.

To better see this, note that if we do not impose A2 in the proof of Proposition II, we have that

$$\begin{aligned} \Pr\left(Release_i^* \leq \epsilon | X_i = x_1, Release_i = 1\right) &> \Pr\left(Release_i^* \leq \epsilon | X_i = x_2, Release_i = 1\right) \\ \iff \qquad \mathbb{E}\left[Release_i | X_i = x_1\right] &< \frac{\theta_{\widetilde{W}|X_i=x_1}(n(X_i)|X_i = x_1)}{\theta_{\widetilde{W}|X_i=x_2}(n(X_i)|X_i = x_2)} \mathbb{E}\left[Release_i | X_i = x_2\right], \end{aligned}$$

where $\theta_{\widetilde{W}|X_i}$ is the conditional density of the demeaned unobserved component (see Appendix A for details). Then, we need the conditional densities to be extremely different when the propensity scores are similar (and in one particular direction) in order to bias the ranking. Then, the provided examples are likely to bias the ranking because the mass at the margin varies substantially between realizations of X_i.

D Data Appendix

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

D.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). The PDO is a centralized public service depending on the Ministry of Justice that provides criminal defense to all accused individuals who ask for the service. For more information, see http://www.dpp.cl/. The centralized nature of the PDO ensures that the administrative records contain information for all the cases handled by the PDO (as opposed to by 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 living, among other characteristics) and case characteristics (case ID, court, 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 detention 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 2006 and 2017. We can 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.

D.2 Estimation sample

The initial sample contains $3,571,230$ cases and covers all the cases recorded by the PDO whose detention 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 $3,569,805$ (i.e, we drop 1,425 cases).

Sample restrictions We then make the following sample restrictions:

- We drop cases where detention follows an ongoing investigation ($1,234,304$ observations). This means that we only consider defendants arrested under probable cause. We do this because the information set of the judges is likely to be different for ongoing investigations and the analysis of prejudice/discrimination is more suitable for arrests under probable cause.
- We drop juvenile defendants ($256,013$ 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., 2019 for details).
- We drop cases where the defendant hires a private attorney ($104,445$ observations). We do this because we do not observe the result of the detention hearing (and what happens after in the prosecution) in those cases. Importantly, we do not observe whether the defendant was pretrial incarcerated.
- We drop cases whose length is larger than two years ($55,601$ 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 $224,840$ 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 detention hearing.
- We drop cases where the detention judge cannot be identified ($65,745$ observations).

- We drop types of crime whose likelihood of pretrial detention is less than 5% (930,128 observations). We do this because we want to study judges' decisions in cases where pretrial detention is a plausible outcome.
- We drop defendants who belong to other ethnic group different than Mapuche (3,428 observations).

After all these adjustments the sample size is 695,301. That matches the numbers of Table 1.

D.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 5 for details.
- *Severity*: we proxy crime severity by computing the share of cases within the type of crime that dictate pretrial detention.
- *Criminal history*: 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.
- *Pretrial misconduct*: pretrial misconduct is an indicator variable that takes value 1 if the defendant do 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.
- *Judge leniency*: as in Dobbie et al. (2018), we use the residualized (against court-by-time fixed effects) leave-out mean release rate at the judge level.
- *Year of prosecution fixed effects*: we consider the initial date to set the fixed effect.

E Assessing Assumptions' Validity

In this appendix, we provide suggestive evidence that both assumptions, A1 and A2, hold in our setting. It has to be kept in mind that both A1 and A2 are not directly testable and, therefore, these tests, while reassuring, are only suggestive. We first assess the separability (monotonicty) assumption. We then perform an exercise that provides support of the joint validity of A1 and A2.

Assumption 1 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 $Release_i$ 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. This test is similar to the monotonicity test performed by Arnold et al. (2018) and Bald et al. (2019).

Tables 8 and 9 show the results using $Release_i$ 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 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 strongly support the monotonicity assumption. In all but two cases (i.e., 97% 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.

Joint test The joint test builds on the intuition of Altonji et al. (2005) and Oster (2019).[42] Recall that V_i are variables that the judges observe, so X_i can be interpreted as elements of V_i that the econometrician happened to see. Then, we can use observed variables to simulate unobservables and assess the validity of the identification argument. An application of this logic can be found in Kline and Walters (2016).

[42] Their methodologies are not exactly suitable to our setting since (i) we allow for standard omitted variable bias, and (ii) we do not require the estimated coefficients of the selection equation to have causal interpretation.

Table 8: 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	-0.059	-0.029	0.014	-0.090	-0.768
	[-0.063, -0.056]	[-0.032, -0.026]	[0.010, 0.017]	[-0.097, -0.084]	[-0.772, -0.764]
Mapuche	-0.050	-0.023	0.009	-0.077	-0.753
	[-0.062, -0.038]	[-0.033, -0.013]	[-0.004, 0.022]	[-0.100, -0.053]	[-0.767, -0.739]
Non-Mapuche	-0.060	-0.030	0.014	-0.092	-0.769
	[-0.064, -0.056]	[-0.033, -0.026]	[0.010, 0.018]	[-0.098, -0.085]	[-0.773, -0.764]
Male	-0.060	-0.033	0.014	-0.080	-0.779
	[-0.063, -0.056]	[-0.036, -0.029]	[0.010, 0.018]	[-0.086, -0.073]	[-0.784, -0.775]
Female	-0.053	-0.008	0.014	-0.205	-0.685
	[-0.063, -0.044]	[-0.015, -0.001]	[0.004, 0.024]	[-0.225, -0.184]	[-0.698, -0.672]
Low income	-0.060	-0.031	0.015	-0.085	-0.774
	[-0.065, -0.055]	[-0.036, -0.026]	[0.009, 0.021]	[-0.095, -0.074]	[-0.780, -0.767]
High income	-0.059	-0.028	0.013	-0.096	-0.764
	[-0.063, -0.054]	[-0.032, -0.024]	[0.008, 0.018]	[-0.104, -0.087]	[-0.770, -0.759]
Low judge	-0.066	-0.025	0.014	-0.109	-0.873
leniency	[-0.071, -0.061]	[-0.030, -0.021]	[0.008, 0.019]	[-0.119, -0.099]	[-0.879, -0.866]
High judge	-0.053	-0.033	0.014	-0.074	-0.666
leniency	[-0.057, -0.048]	[-0.037, -0.029]	[0.009, 0.019]	[-0.082, -0.066]	[-0.672, -0.660]
Small Court	-0.057	-0.028	0.015	-0.105	-0.799
(No. of cases)	[-0.062, -0.052]	[-0.032, -0.023]	[0.010, 0.021]	[-0.115, -0.095]	[-0.806, -0.793]
Big Court	-0.061	-0.031	0.012	-0.082	-0.740
(No. of cases)	[-0.066, -0.056]	[-0.036, -0.026]	[0.006, 0.017]	[-0.091, -0.073]	[-0.746, -0.734]
Small Court	-0.059	-0.031	0.020	-0.107	-0.806
(No. of judges)	[-0.064, -0.054]	[-0.035, -0.027]	[0.014, 0.025]	[-0.117, -0.097]	[-0.812, -0.800]
Big Court	-0.059	-0.027	0.007	-0.079	-0.732
(No. of judges)	[-0.063, -0.054]	[-0.031, -0.022]	[0.001, 0.012]	[-0.088, -0.070]	[-0.738, -0.726]
Low severity	-0.049	-0.029	0.009	-0.069	-0.647
court	[-0.054, -0.045]	[-0.033, -0.025]	[0.005, 0.014]	[-0.077, -0.061]	[-0.653, -0.642]
High severity	-0.070	-0.029	0.018	-0.114	-0.891
court	[-0.075, -0.064]	[-0.034, -0.025]	[0.012, 0.024]	[-0.124, -0.104]	[-0.897, -0.884]

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 200 repetitions.

We perform the following exercise. Assume that our set of observed variables, X_i, is a good approximation (up to some small well-behaved noise) of the judges' (complete) information set. Under that assumption, the identification of marginally released defendants using the ranking based on the propensity score is accurate. We fit the propensity score and label as marginal the bottom 5% of the predicted probability distribution. Then, we omit one observable (label it as V_i) and (i) estimate the propensity score with the restricted set of observables and identify marginals using the ranking strategy, and (ii) compute the conditional probabilities of being marginal, namely the shares of marginals identified in the first step for different combinations of the observables used in the restricted estimation. We then compute the rank correlation between (i) the share of marginals using the restricted propensity-score ranking and the conditional probabilities, and (ii) the estimated propensity score using the restricted set of observables and the conditional probabilities of

Table 9: 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	0.006	0.052	0.118	0.040	0.048
	[0.001, 0.010]	[0.048, 0.056]	[0.113, 0.122]	[0.029, 0.051]	[0.039, 0.057]
Mapuche	0.004	0.044	0.108	0.066	0.060
	[-0.012, 0.020]	[0.031, 0.058]	[0.091, 0.125]	[0.026, 0.105]	[0.028, 0.092]
Non-Mapuche	0.006	0.052	0.118	0.038	0.047
	[0.001, 0.010]	[0.048, 0.056]	[0.113, 0.123]	[0.027, 0.049]	[0.037, 0.056]
Male	0.004	0.058	0.118	0.047	0.056
	[-0.001, 0.009]	[0.054, 0.062]	[0.113, 0.124]	[0.036, 0.058]	[0.047, 0.066]
Female	0.020	0.021	0.108	-0.046	-0.022
	[0.007, 0.033]	[0.011, 0.030]	[0.095, 0.121]	[-0.084, -0.008]	[-0.049, 0.004]
Low income	0.005	0.064	0.106	0.033	0.077
	[-0.002, 0.011]	[0.058, 0.070]	[0.099, 0.113]	[0.017, 0.050]	[0.064, 0.091]
High income	0.006	0.044	0.124	0.046	0.022
	[0.000, 0.012]	[0.039, 0.049]	[0.118, 0.131]	[0.032, 0.060]	[0.011, 0.034]
Low judge leniency	0.007	0.033	0.131	0.043	0.049
	[0.000, 0.013]	[0.028, 0.039]	[0.124, 0.138]	[0.027, 0.058]	[0.036, 0.062]
High judge leniency	0.005	0.069	0.105	0.037	0.046
	[-0.001, 0.011]	[0.064, 0.074]	[0.099, 0.111]	[0.023, 0.052]	[0.035, 0.058]
Small Court (No. of cases)	0.003	0.054	0.106	0.046	0.091
	[-0.003, 0.009]	[0.049, 0.060]	[0.100, 0.113]	[0.030, 0.062]	[0.078, 0.103]
Big Court (No. of cases)	0.008	0.049	0.128	0.035	0.014
	[0.002, 0.014]	[0.043, 0.054]	[0.121, 0.135]	[0.020, 0.050]	[0.001, 0.026]
Small Court (No. of judges)	0.006	0.047	0.114	0.057	0.064
	[-0.001, 0.012]	[0.042, 0.053]	[0.107, 0.121]	[0.041, 0.073]	[0.051, 0.077]
Big Court (No. of judges)	0.005	0.057	0.120	0.025	0.030
	[-0.001, 0.011]	[0.051, 0.063]	[0.114, 0.127]	[0.010, 0.039]	[0.018, 0.042]
Low severity court	0.002	0.073	0.101	0.042	0.061
	[-0.004, 0.008]	[0.068, 0.079]	[0.095, 0.108]	[0.028, 0.057]	[0.049, 0.072]
High severity court	0.010	0.028	0.135	0.035	0.030
	[0.003, 0.017]	[0.022, 0.034]	[0.128, 0.142]	[0.019, 0.051]	[0.016, 0.043]

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 200 repetitions.

being marginal. In case (i), the correlation is expected to be positive. In case (ii), the correlation is expected to be negative. If the identification argument holds, we should expect these rank correlations to be large.

We perform this exercise by using each of the 13 observables used in the estimation as V_i.[43] To compute the rank-correlations, we discretize the non-discrete regressors (using the median) to define $2^{(13-1)} = 4,096$ categories of observables. For each of these categories, we compute the average restricted estimated propensity score, the average share of marginals using the restricted propensity score, and the conditional probability of being marginal using the base estimation as

[43]Number of previous cases, severity of previous case, severity of current case, average severity by year-court, number of cases by year-court, judge leniency, jugde leniency squared, Mapuche indicator, gender, previous case indicator, previous pretrial misconduct indicator, previous conviction indicator.

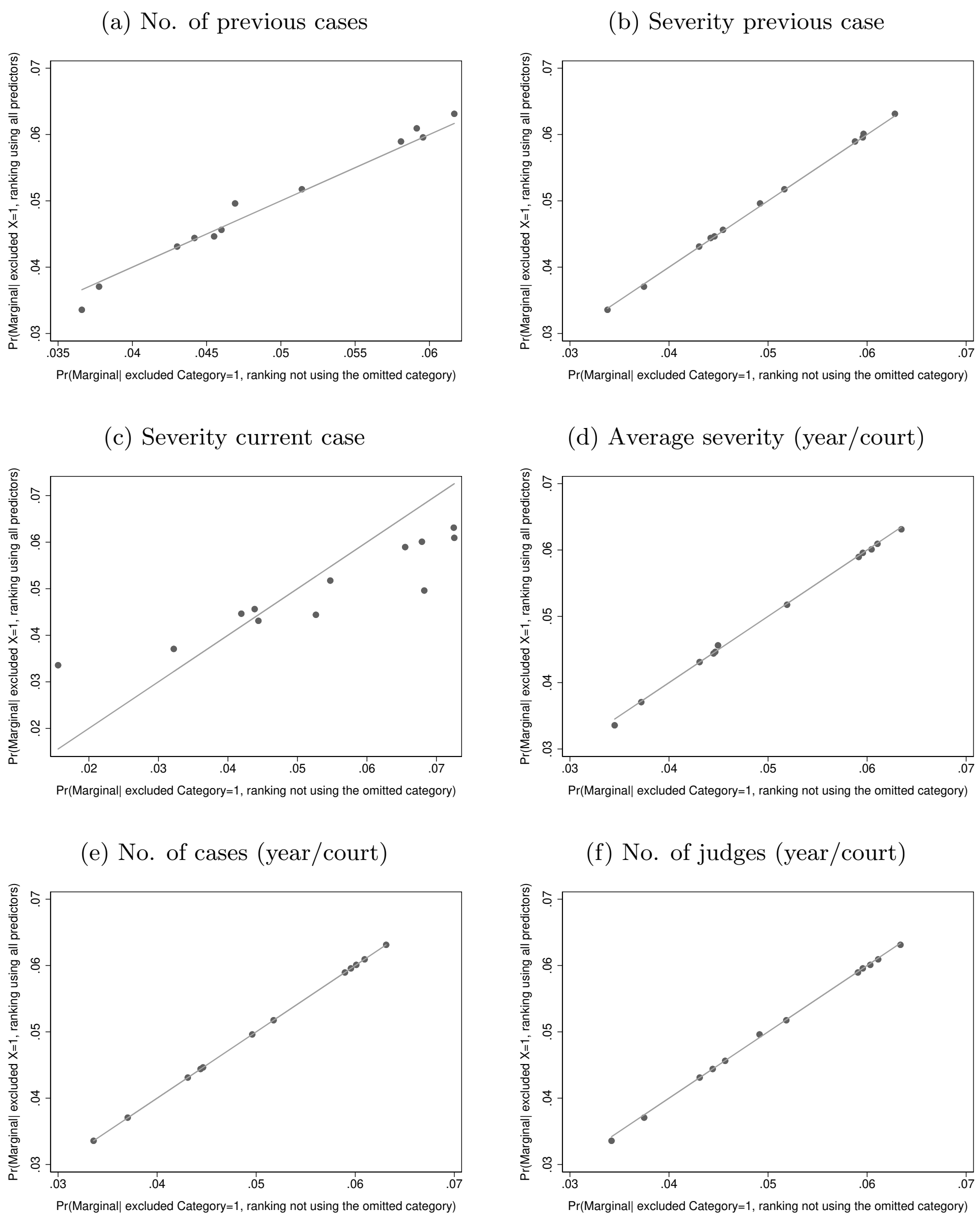
the true share of marginals. Table 10 presents the results. We report both the Spearman's-ρ and Kendall's-τ statistics for rank correlation. It can be seen that in all variables by one (severity of current case), the correlations are very large. Figure 7 plot the share of marginals using the full and restricted set of observables, for each regression excluding one observable. Again, all but severity of current case lie close to the 45 degree line. We interpret this as strong suggestive evidence of the validity of our identification assumption.

Table 10: Rank Correlations

	Corr. btw. $\Pr(Marg \mid X = x, Release = 1)$ and $\mathbb{E}[Marg \mid X = x]$ using restricted p-score		Corr. btw. $\Pr(Marg \mid X = x, Release = 1)$ and $\mathbb{E}[Release \mid X = x]$ using restricted p-score	
Excluded predictor	Spearman	Kendall	Spearman	Kendall
No. of previous cases	0.979	0.966	-0.631	-0.518
Severity previous case	0.969	0.962	-0.633	-0.519
Severity current case	0.544	0.494	-0.400	-0.295
Average severity (year/court)	0.974	0.962	-0.661	-0.544
No. of cases (year/court)	1.000	1.000	-0.655	-0.537
No. of judges (year/court)	0.989	0.985	-0.646	-0.531
Judge leniency	0.979	0.969	-0.660	-0.544
Judge leniency square	0.999	0.996	-0.661	-0.546
Mapuche	1.000	0.999	-0.662	-0.545
Male	0.999	0.998	-0.661	-0.544
Previous case	0.913	0.886	-0.663	-0.540
Previous pretrial misconduct	0.991	0.986	-0.691	-0.565
Previous conviction	0.992	0.988	-0.698	-0.575

Note: This table presents the rank-correlations between the ranking of the conditional probabilities of being marginal and (i) the ranking of the conditional share of marginals using the restricted propensity score estimation, and (ii) the ranking of the predicted propensity score using the restricted estimation. We report the Spearman's-ρ and the Kendall's-τ_b rank correlation statistics. The excluded predictor is specified in the first column. All regressions include year fixed effects. The unit of analysis to build the ranking is the combination of all possible values of the predictors, without considering the excluded category (i.e., 12 predictors), where the continuous predictors were transformed into binary variables by using the median among released as threshold. Then, each combination of predictors defines a cell, where the maximum number of cells is $2^{12} = 4,096$. Since there are cells without released defendants, in practice this number is between $1,829$ and $3,051$, depending on the excluded predictor.

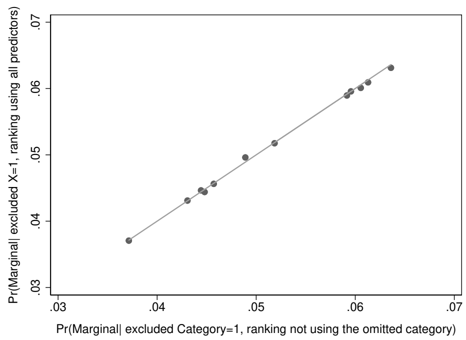
Figure 7: Fraction of Marginals by Group: Unrestricted vs. Restricted Observables (Excluded Predictors in the Titles)



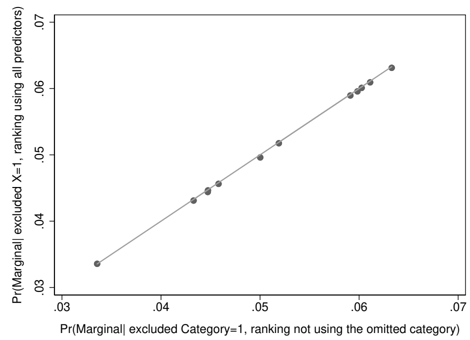
Note: Each figure plots the estimated fraction of marginal defendants using the estimated propensity score using all predictors (y-axis) and all-but-one predictors (x-axis). The excluded regressor is specified in the title of the figure. The fraction of marginally released defendants is estimated by *group*, where a *group* is given by all the released defendants for whom the value of a predictor is equal to one. The continuous predictors were transformed into binary variables by using the median among released as threshold. By construction, these groups overlap. To make easier the comparison between these fractions we also plot a solid 45 degree line.

Figure 7: Fraction of Marginals by Group: Unrestricted vs. Restricted Observables (cont.)
(Excluded Predictors in the Titles)

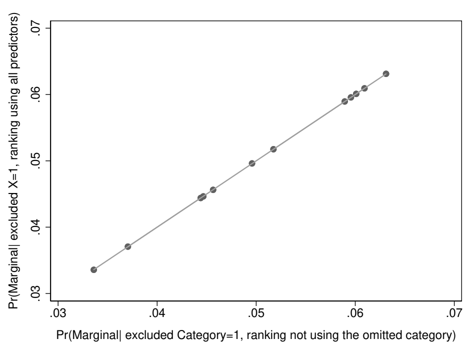
(g) Judge leniency



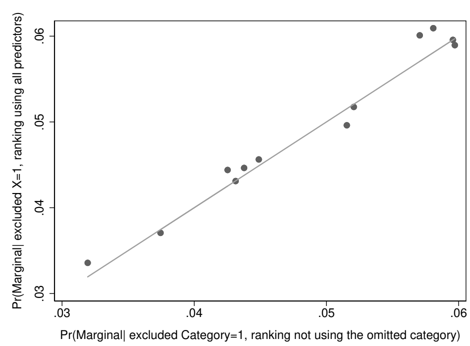
(h) Mapuche



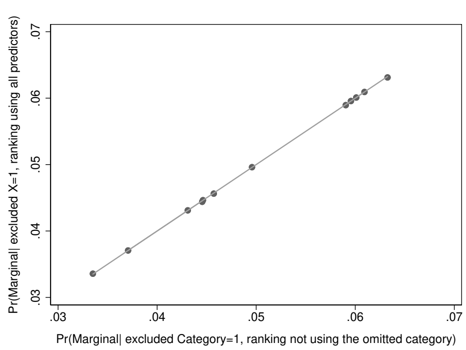
(i) Male



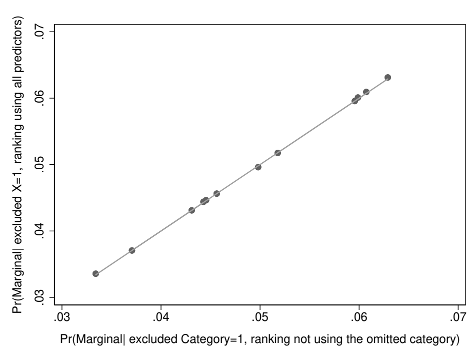
(j) At least one previous case



(k) At least one previous pretrial misconduct



(l) At least one previous conviction



Note: Each figure plots the estimated fraction of marginal defendants using the estimated propensity score using all predictors (y-axis) and all-but-one predictors (x-axis). The excluded regressor is specified in the title of the figure. The fraction of marginally released defendants is estimated by *group*, where a *group* is given by all the released defendants for whom the value of a predictor is equal to one. The continuous predictors were transformed into binary variables by using the median among released as threshold. By construction, these groups overlap. To make easier the comparison between these fractions we also plot a solid 45 degree line.

F Prediction Models

Table 11: 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	-0.005	-0.007	-0.012	-0.005
	(0.002)	(0.004)	(0.004)	(0.002)
Male	-0.004	-0.004	-0.004	-0.004
	(0.001)	(0.001)	(0.001)	(0.001)
Previous prosecution	-0.058	-0.059	-0.059	-0.058
	(0.002)	(0.002)	(0.002)	(0.002)
Previous pretrial misconduct	-0.028	-0.028	-0.028	-0.028
	(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	-0.008	-0.008	-0.008	-0.008
	(0.000)	(0.000)	(0.000)	(0.000)
Severity (previous prosecution)	-0.092	-0.093	-0.093	-0.092
	(0.003)	(0.003)	(0.004)	(0.003)
Severity (current prosecution)	-0.776	-0.775	-0.777	-0.776
	(0.003)	(0.003)	(0.003)	(0.003)
Average severity of the cases (court/year)	-0.754	-0.755	-0.760	-0.755
	(0.015)	(0.016)	(0.016)	(0.015)
No. of cases per court/year	0.0000004	0.0000003	0.0000004	0.0000004
	(0.0000003)	(0.0000003)	(0.0000003)	(0.0000003)
No. of judges per court/year	0.00023	0.00023	0.00023	0.00023
	(0.00001)	(0.00001)	(0.00002)	(0.00001)
Judge leniency	0.390	0.389	0.387	0.389
	(0.012)	(0.013)	(0.013)	(0.012)
Judge leniency squared	0.842	0.787	0.800	0.839
	(0.110)	(0.113)	(0.113)	(0.110)
Year of Prosecution fixed effects	YES	YES	YES	YES
Court fixed effects	NO	NO	NO	NO
No. of Mapuche	50,880	9,658	9,315	52,053
No. of Non-Mapuche	642,843	642,843	642,843	642,843
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)	0.59	0.59	0.59	0.59
Correctly classified (prediction: Released)	0.87	0.87	0.87	0.87
Predictions on release probability:				
Predicted probability (Mapuche)	0.85	0.87	0.85	0.85
Predicted probability (Non-Mapuche)	0.84	0.84	0.84	0.84
Predicted probability (1st percentile)	0.23	0.23	0.23	0.23
Predicted probability (5th percentile)	0.42	0.42	0.42	0.42
Predicted probability (10th percentile)	0.57	0.57	0.57	0.57

Note: This table presents the point estimates and robust standard errors of a probit model for the determinants of the release status using the data described in Table 1. The four models correspond to the four definitions of Mapuche considered in this paper. The predicted probabilities for relevant subgroups are presented at the end of the table.

Table 12: 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	-0.001	-0.002	-0.003	-0.001
	(0.002)	(0.003)	(0.004)	(0.002)
Male	-0.007	-0.007	-0.007	-0.007
	(0.001)	(0.001)	(0.001)	(0.001)
Previous prosecution	-0.039	-0.039	-0.039	-0.039
	(0.002)	(0.002)	(0.002)	(0.002)
Previous pretrial misconduct	-0.019	-0.019	-0.019	-0.019
	(0.001)	(0.001)	(0.001)	(0.001)
Previous conviction	0.013	0.013	0.013	0.013
	(0.002)	(0.002)	(0.002)	(0.002)
No. of previous prosecution	-0.011	-0.011	-0.011	-0.011
	(0.000)	(0.000)	(0.000)	(0.000)
Severity (previous prosecution)	-0.154	-0.155	-0.155	-0.154
	(0.005)	(0.005)	(0.005)	(0.005)
Severity (current prosecution)	-1.023	-1.020	-1.021	-1.023
	(0.004)	(0.004)	(0.004)	(0.003)
Average severity of the cases (court/year)	-1.023	-1.026	-1.033	-1.024
	(0.022)	(0.023)	(0.023)	(0.022)
No. of cases per court/year	-0.0000020	-0.0000022	-0.0000023	-0.0000020
	(0.0000011)	(0.0000011)	(0.0000011)	(0.0000011)
No. of judges per court/year	-0.00014	-0.00013	-0.00014	-0.00014
	(0.00005)	(0.00005)	(0.00005)	(0.00005)
Judge leniency	0.416	0.413	0.411	0.415
	(0.014)	(0.015)	(0.015)	(0.014)
Judge leniency squared	0.627	0.571	0.595	0.625
	(0.140)	(0.144)	(0.144)	(0.140)
Year of Prosecution fixed effects	YES	YES	YES	YES
Court fixed effects	YES	YES	YES	YES
No. of Mapuche	50,880	9,658	9,315	52,053
No. of Non-Mapuche	642,843	642,843	642,843	642,843
R-squared	0.21	0.21	0.21	0.21
Correctly classified (0.5 prob as threshold)	0.85	0.85	0.85	0.85
Correctly classified (prediction: Non-Released)	0.64	0.64	0.64	0.64
Correctly classified (prediction: Released)	0.86	0.86	0.86	0.86
Predictions on release probability:				
Predicted probability (Mapuche)	0.85	0.87	0.85	0.85
Predicted probability (non-Mapuche)	0.83	0.83	0.83	0.83
Predicted probability (1st percentile)	0.36	0.37	0.36	0.36
Predicted probability (5th percentile)	0.50	0.50	0.50	0.49
Predicted probability (10th percentile)	0.58	0.58	0.58	0.58

Note: This table presents the point estimates and robust standard errors of a linear probability model for the determinants of the release status using the data described in Table 1. The four models correspond to the four definitions of Mapuche considered in this paper. The predicted probabilities for relevant subgroups are presented at the end of the table.

G Robustness checks

Table 13: 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):	-0.045	-0.122	-0.076	-0.042
C.I. (95%)	[-0.064, -0.021]	[-0.171, -0.067]	[-0.120, -0.026]	[-0.060, -0.017]
(a) Mapuche expectation	0.291	0.215	0.261	0.295
(b) Non-Mapuche expectation	0.337	0.337	0.337	0.337
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.040	-0.113	-0.076	-0.036
C.I. (95%)	[-0.068, -0.015]	[-0.180, -0.049]	[-0.127, -0.020]	[-0.065, -0.013]
(a) Mapuche expectation	0.308	0.235	0.272	0.312
(b) Non-Mapuche expectation	0.348	0.348	0.348	0.348
No. of Mapuche ($\leq$ 5th pctl.)	1,949	298	333	2,014
No. of Non-Mapuche ($\leq$ 5th pctl.)	27,037	26,958	26,898	27,020
Data up 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):	-0.045	-0.119	-0.050	-0.042
C.I. (95%)	[-0.062, -0.028]	[-0.156, -0.083]	[-0.088, -0.017]	[-0.058, -0.027]
(a) Mapuche expectation	0.284	0.211	0.279	0.287
(b) Non-Mapuche expectation	0.330	0.330	0.330	0.329
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.045	-0.116	-0.069	-0.041
C.I. (95%)	[-0.066, -0.026]	[-0.165, -0.069]	[-0.108, -0.024]	[-0.063, -0.020]
(a) Mapuche expectation	0.289	0.219	0.266	0.293
(b) Non-Mapuche expectation	0.334	0.335	0.335	0.334
No. of Mapuche ($\leq$ 10th pctl.)	3,807	545	580	3,927
No. of Non Mapuche ($\leq$ 10th pctl.)	54,165	53,966	53,881	54,141

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 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 F. The confidence intervals are calculated using bootstrap with 200 repetitions.

Table 14: Prediction-Based Outcome Test, Using OLS to Estimate the Release Probability and Lasso to Select Predictors Using Interactions and Squared Terms (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):	-0.058	-0.120	-0.087	-0.055
C.I. (95%)	[-0.079, -0.040]	[-0.164, -0.074]	[-0.136, -0.032]	[-0.075, -0.036]
(a) Mapuche expectation	0.296	0.236	0.268	0.301
(b) Non-Mapuche expectation	0.354	0.356	0.355	0.356
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.047	-0.119	-0.097	-0.046
C.I. (95%)	[-0.074, -0.025]	[-0.172, -0.065]	[-0.159, -0.028]	[-0.071, -0.023]
(a) Mapuche expectation	0.316	0.245	0.266	0.319
(b) Non-Mapuche expectation	0.363	0.365	0.364	0.365
No. of Mapuche ($\leq$ 5th pctl.)	2,018	318	351	2,084
No. of Non-Mapuche ($\leq$ 5th pctl.)	26,968	26,938	26,880	26,953
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):	-0.051	-0.132	-0.045	-0.049
C.I. (95%)	[-0.066, -0.036]	[-0.167, -0.094]	[-0.088, -0.010]	[-0.063, -0.032]
(a) Mapuche expectation	0.321	0.243	0.328	0.326
(b) Non-Mapuche expectation	0.372	0.374	0.372	0.375
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.052	-0.124	-0.079	-0.050
C.I. (95%)	[-0.070, -0.034]	[-0.163, -0.079]	[-0.118, -0.033]	[-0.068, -0.032]
(a) Mapuche expectation	0.312	0.242	0.285	0.316
(b) Non-Mapuche expectation	0.364	0.366	0.364	0.366
No. of Mapuche ($\leq$ 10th pctl.)	3,916	552	641	4,026
No. of Non-Mapuche ($\leq$ 10th pctl.)	54,057	53,959	53,820	54,042

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 outcome is pretrial misconduct. The original set of covariates included 85 variables to be chosen: the predictors considered in Table 13, their squared terms, and their interactions. When Mapuche is defined as *at least one surname*, lasso selected 44 predictors, 47 when it is defined as *two surnames*, 46 when it is defined as *self-reported*, and 45 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 86% and 62% of the cases, respectively. The other characteristics of this table replicates Table 13. 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 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 200 repetitions.

Table 15: Prediction-Based Outcome Test, Using OLS to Estimate the Release Probability and Lasso to Select Predictors Using Judges Fixed Effects (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):	-0.050	-0.128	-0.081	-0.045
C.I. (95%)	[-0.071, -0.030]	[-0.167, -0.072]	[-0.122, -0.023]	[-0.066, -0.026]
(a) Mapuche expectation	0.287	0.209	0.255	0.291
(b) Non-Mapuche expectation	0.336	0.337	0.337	0.336
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.037	-0.126	-0.077	-0.033
C.I. (95%)	[-0.063, -0.013]	[-0.185, -0.056]	[-0.127, -0.012]	[-0.060, -0.010]
(a) Mapuche expectation	0.308	0.219	0.268	0.312
(b) Non-Mapuche expectation	0.345	0.345	0.345	0.345
No. of Mapuche ($\leq$ 5th pctl.)	1,954	292	321	2,010
No. of Non-Mapuche ($\leq$ 5th pctl.)	27,048	26,979	26,924	27,040
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	-0.114	-0.061	-0.040
C.I. (95%)	[-0.059, -0.031]	[-0.155, -0.081]	[-0.098, -0.018]	[-0.055, -0.029]
(a) Mapuche expectation	0.284	0.213	0.266	0.287
(b) Non-Mapuche expectation	0.327	0.327	0.327	0.326
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.045	-0.120	-0.069	-0.041
C.I. (95%)	[-0.066, -0.029]	[-0.164, -0.076]	[-0.118, -0.025]	[-0.063, -0.027]
(a) Mapuche expectation	0.287	0.213	0.264	0.291
(b) Non-Mapuche expectation	0.332	0.333	0.333	0.332
No. of Mapuche ($\leq$ 10th pctl.)	3,840	550	583	3,960
No. of Non-Mapuche ($\leq$ 10th pctl.)	54,164	53,991	53,907	54,140

Notes: This table presents the results from the P-BOT with the release probabilities predicted by using a linear model. The predictors were selected using Lasso. The outcome is pretrial misconduct, The original set of covariates included 1,187 variables to be chosen: the predictors considered in Table 13 (excluding the judge leniency and its square) and all the judges fixed effects. When Mapuche is defined as *at least one surname*, lasso selected 791 predictors, 810 when it is defined as *two surnames*, 811 when it is defined as *self-reported*, and 791 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 86% and 64% of the cases, respectively. The other characteristics of this table replicates Table 13. 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 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 200 repetitions.

Table 16: Prediction-Based Outcome Test, Using Probit to Estimate the Release Probability (Outcome: Non-Appearance in Court)

Data up 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):	-0.022	-0.059	-0.025	-0.021
C.I. (95%)	[-0.039, -0.006]	[-0.092, -0.020]	[-0.060, 0.020]	[-0.035, -0.004]
(a) Mapuche expectation	0.142	0.104	0.138	0.143
(b) Non-Mapuche expectation	0.164	0.164	0.164	0.164
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.013	-0.045	-0.024	-0.013
C.I. (95%)	[-0.033, 0.008]	[-0.083, 0.017]	[-0.061, 0.026]	[-0.033, 0.006]
(a) Mapuche expectation	0.152	0.121	0.142	0.152
(b) Non-Mapuche expectation	0.166	0.166	0.166	0.166
No. of Mapuche ($\leq$ 5th pctl.)	1,865	268	311	1,931
No. of Non-Mapuche ($\leq$ 5th pctl.)	27,121	26,988	26,920	27,103
Data up 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):	-0.032	-0.074	-0.024	-0.031
C.I. (95%)	[-0.042, -0.019]	[-0.109, -0.044]	[-0.054, 0.015]	[-0.041, -0.019]
(a) Mapuche expectation	0.153	0.111	0.161	0.154
(b) Non-Mapuche expectation	0.185	0.185	0.185	0.185
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.021	-0.055	-0.026	-0.020
C.I. (95%)	[-0.036, -0.004]	[-0.090, -0.016]	[-0.066, 0.009]	[-0.033, -0.001]
(a) Mapuche expectation	0.150	0.115	0.145	0.151
(b) Non-Mapuche expectation	0.171	0.171	0.171	0.171
No. of Mapuche ($\leq$ 10th pctl.)	3,699	497	585	3,817
No. of Non-Mapuche ($\leq$ 10th pctl.)	54,273	54,014	53,876	54,251

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 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 F. The confidence intervals are calculated using bootstrap with 200 repetitions.

Table 17: Prediction-Based Outcome Test, Using Probit to Estimate the Release Probability (Outcome: Pretrial Recidivism)

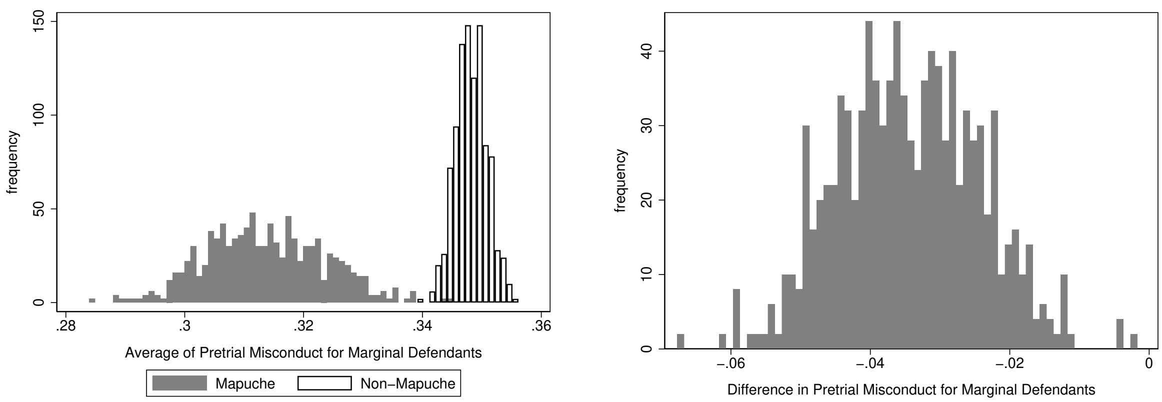
Data up 5th percentile	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Panel A: Simple Version				
Point estimate, (a)-(b):	-0.040	-0.089	-0.072	-0.036
C.I. (95%)	[-0.062, -0.014]	[-0.142, -0.033]	[-0.111, -0.008]	[-0.057, -0.011]
(a) Mapuche expectation	0.232	0.183	0.199	0.236
(b) Non-Mapuche expectation	0.272	0.272	0.272	0.272
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.033	-0.088	-0.059	-0.028
C.I. (95%)	[-0.057, -0.005]	[-0.151, -0.012]	[-0.115, 0.003]	[-0.053, -0.002]
(a) Mapuche expectation	0.249	0.194	0.222	0.254
(b) Non-Mapuche expectation	0.282	0.282	0.282	0.282
No. of Mapuche ($\leq$ 5th pctl.)	1,865	268	311	1,931
No. of Non-Mapuche ($\leq$ 5th pctl.)	27,121	26,988	26,920	27,103
Data up 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	-0.099	-0.052	-0.038
C.I. (95%)	[-0.055, -0.027]	[-0.135, -0.067]	[-0.088, -0.015]	[-0.054, -0.026]
(a) Mapuche expectation	0.224	0.165	0.212	0.226
(b) Non-Mapuche expectation	0.264	0.264	0.264	0.264
Panel B: Non-Parametric				
Point estimate, (a)-(b):	-0.037	-0.088	-0.057	-0.034
C.I. (95%)	[-0.059, -0.020]	[-0.135, -0.050]	[-0.101, -0.012]	[-0.055, -0.015]
(a) Mapuche expectation	0.235	0.185	0.216	0.239
(b) Non-Mapuche expectation	0.272	0.273	0.272	0.272
No. of Mapuche ($\leq$ 10th pctl.)	3,699	497	585	3,817
No. of Non-Mapuche ($\leq$ 10th pctl.)	54,273	54,014	53,876	54,251

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 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 F. The confidence intervals are calculated using bootstrap with 200 repetitions.

Figure 8: Perturbation Test (Mapuche: At least one surname)

(a) Mapuche and Non-Mapuche averages

(b) Differences between Mapuche and Non-Mapuche

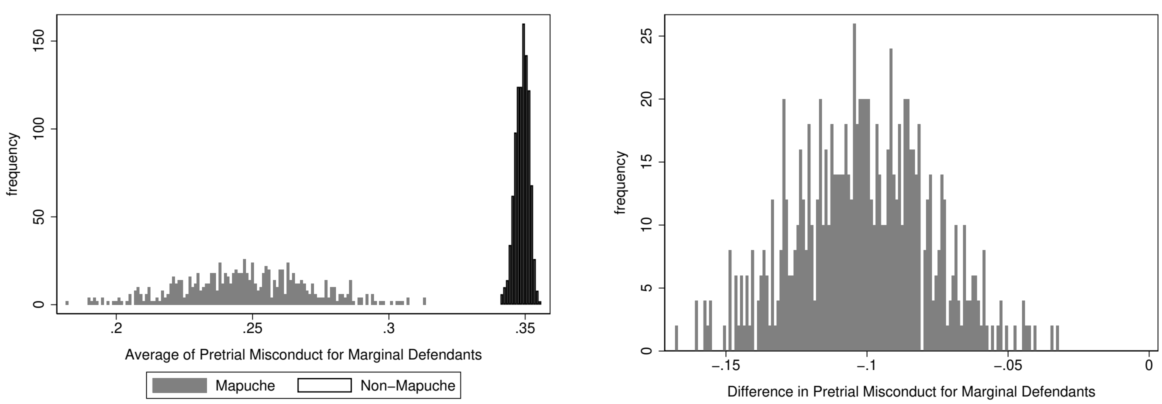


Note: These plots present the results of the perturbation test described in Section 4. They are produced in the following steps. First, we estimate the probit model (see Table 11). Then, for each released individual in the sample, we simulate 500 realizations from a standardized normal distribution to simulate $Release_i^*$ and redefine the samples of marginal individuals. Within each sample, we estimate $\mathbb{E}[PM_i|Mapuche_i = 1, Release_i^* = 0]$ and $\mathbb{E}[PM_i|Mapuche_i = 0, Release_i^* = 0]$. Panel (a) presents one histogram for each group. Panel (b) presents the histogram for the difference between these two estimated conditional expectations within each simulation.

Figure 9: Perturbation Test (Mapuche: Two surnames)

(a) Mapuche and Non-Mapuche averages

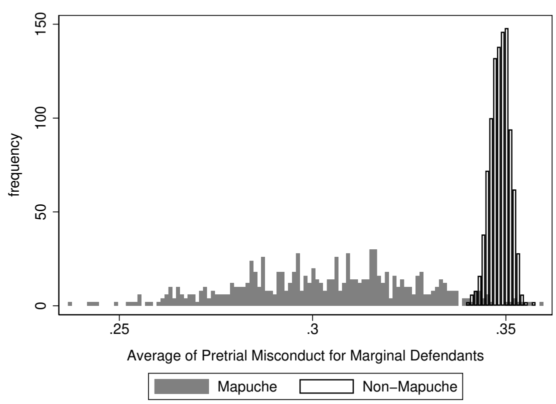
(b) Differences between Mapuche and Non-Mapuche



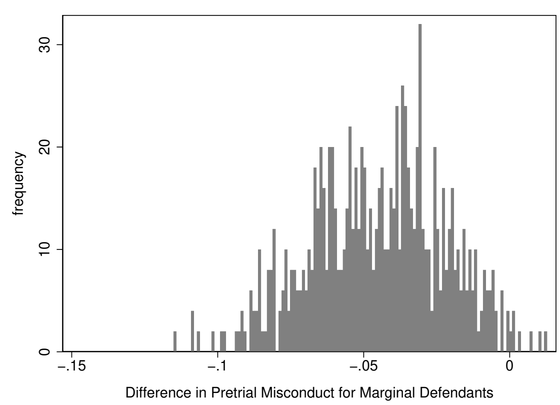
Note: These plots present the results of the perturbation test described in Section 4. They are produced in the following steps. First, we estimate the probit model (see Table 11). Then, for each released individual in the sample, we simulate 500 realizations from a standardized normal distribution to simulate $Release_i^*$ and redefine the samples of marginal individuals. Within each sample, we estimate $\mathbb{E}[PM_i|Mapuche_i = 1, Release_i^* = 0]$ and $\mathbb{E}[PM_i|Mapuche_i = 0, Release_i^* = 0]$. Panel (a) presents one histogram for each group. Panel (b) presents the histogram for the difference between these two estimated conditional expectations within each simulation.

Figure 10: Perturbation Test (Mapuche: Self-reported)

(a) Mapuche and Non-Mapuche averages



(b) Differences between Mapuche and Non-Mapuche



Note: These plots present the results of the perturbation test described in Section 4. They are produced in the following steps. First, we estimate the probit model (see Table 11). Then, for each released individual in the sample, we simulate 500 realizations from a standardized normal distribution to simulate $Release_i^*$ and redefine the samples of marginal individuals. Within each sample, we estimate $\mathbb{E}[PM_i|Mapuche_i = 1, Release_i^* = 0]$ and $\mathbb{E}[PM_i|Mapuche_i = 0, Release_i^* = 0]$. Panel (a) presents one histogram for each group. Panel (b) presents the histogram for the difference between these two estimated conditional expectations within each simulation.

H Randomization Test

Table 18: Predicting Release Status

		Mapuche			
	Non-Mapuche	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Male	-0.012	-0.005	-0.020	0.008	-0.006
	(0.002)	(0.005)	(0.009)	(0.010)	(0.005)
Previous prosecution	-0.039	-0.028	-0.001	-0.042	-0.027
	(0.003)	(0.006)	(0.014)	(0.017)	(0.006)
Previous pretrial misconduct	-0.020	-0.014	-0.006	0.000	-0.015
	(0.002)	(0.004)	(0.009)	(0.009)	(0.004)
Previous conviction	0.013	0.005	-0.014	-0.001	0.006
	(0.002)	(0.006)	(0.012)	(0.015)	(0.006)
No. of previous prosecutions	-0.011	0.012	-0.016	-0.013	-0.012
	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)
Severity (previous prosecution)	-0.155	-0.140	-0.121	-0.123	-0.140
	(0.008)	(0.017)	(0.043)	(0.052)	(0.017)
Severity (current prosecution)	-1.003	-1.038	-1.036	-1.181	-1.039
	(0.021)	(0.026)	(0.050)	(0.048)	(0.027)
Drug crime	-0.026	-0.025	0.000	0.030	-0.026
	(0.012)	(0.017)	(0.028)	(0.031)	(0.017)
Homicide	-0.027	-0.037	-0.022	0.103	-0.029
	(0.014)	(0.024)	(0.059)	(0.054)	(0.025)
Property crime	0.002	0.011	0.011	0.014	0.010
	(0.003)	(0.005)	(0.011)	(0.010)	(0.005)
Court-by-time fixed effects	YES	YES	YES	YES	YES
Observations	643,204	50,917	9,665	9,327	52,093
Joint-F-test	585.09	326.23	84.78	157.04	294.39
p-value	0.000	0.000	0.000	0.000	0.000

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 court level.

Table 19: Predicting Judge Leniency

		Mapuche			
	Non-Mapuche	At least one surname	Two surnames	Self-reported	Self-reported or at least one surname
Male	0.0001	-0.0002	-0.0019	0.0013	-0.0003
	(0.0001)	(0.0006)	(0.0017)	(0.0029)	(0.0006)
Previous prosecution	-0.0000	0.0004	-0.0047	0.0008	0.0006
	(0.0001)	(0.0008)	(0.0031)	(0.0035)	(0.0008)
Previous pretrial misconduct	-0.0000	0.0006	-0.0017	0.0028	0.0007
	(0.0001)	(0.0007)	(0.0024)	(0.0017)	(0.0007)
Previous conviction	-0.0000	-0.0003	0.0050	-0.0016	-0.0005
	(0.0001)	(0.0009)	(0.0031)	(0.0033)	(0.0009)
No. of previous prosecutions	0.0000	0.0001	0.0003	-0.0004	0.0000
	(0.0000)	(0.0001)	(0.0003)	(0.0002)	(0.0001)
Severity (previous prosecution)	-0.0000	-0.0007	-0.0054	-0.0008	-0.0002
	(0.0003)	(0.0021)	(0.0072)	(0.0055)	(0.0018)
Severity (current prosecution)	-0.0005	0.0029	0.0132	0.0013	0.0025
	(0.0003)	(0.0016)	(0.0054)	(0.0059)	(0.0016)
Drug crime	0.0000	-0.0001	0.0016	0.0002	-0.0003
	(0.0001)	(0.0006)	(0.0039)	(0.0025)	(0.0006)
Homicide	0.0001	-0.0009	-0.0052	-0.0010	0.0001
	(0.0003)	(0.0017)	(0.0048)	(0.0036)	(0.0016)
Property crime	-0.0000	-0.0002	-0.0017	-0.0004	-0.0003
	(0.0001)	(0.0005)	(0.0013)	(0.0011)	(0.0005)
Court-by-time fixed effects	YES	YES	YES	YES	YES
Observations	642,778	49,544	7,960	7,733	50,770
Joint-F-test	0.76	1.44	1.71	1.06	1.35
p-value	0.669	0.167	0.091	0.403	0.208

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 race-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 court level.

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). Since our IV model is only well-behaved in the sample of non-Mapuche defendants, we limit the comparison to this group.

Since the P-BOT explicitly identifies marginally released defendants, it is straightforward to characterize their distribution of observables. 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 $\Pr(Release_i = 1|Z_i = \overline{z}) - \Pr(Release_i = 1|Z_i = \underline{z}) = \Pr(Release_i(\overline{z}) > Release_i(\underline{z}))$. 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.[44] 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., $\Pr(Release_i(\overline{z}) > Release_i(\underline{z})|X_i = x)$. Then, by Bayes rule

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

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

Tables 20 presents these conditional probabilities for the marginal defendants identified by the 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 other words, under both methodologies, marginally released defendants are more likely to be male, to have previous prosecutions, to have been engaged in pre-

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

trial misconduct in the past, to have been convicted in the past, and to be accused of more severe crimes. We interpret this as strong 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 20: Characteristics of Marginal Defendants

	$\Pr[X=x]$	$\Pr[X=x \mid \text{Marginal}]$ IV	$\Pr[X=x \mid \text{Marginal}]$ P-BOT (5%)	$\Pr[X=x \mid \text{Marginal}]$ P-BOT (10%)
Male	0.885	0.890	0.918	0.920
	(0.0004)	(0.0096)	(0.0017)	(0.0012)
Female	0.115	0.113	0.082	0.080
	(0.0004)	(0.0098)	(0.0017)	(0.0012)
At least one previous prosecution	0.567	0.738	0.861	0.765
	(0.0006)	(0.0149)	(0.0029)	(0.0024)
No previous prosecution	0.433	0.258	0.139	0.235
	(0.0006)	(0.0147)	(0.0029)	(0.0024)
At least one previous pretrial misconduct	0.757	0.875	0.943	0.895
	(0.0005)	(0.0113)	(0.0016)	(0.0017)
No previous pretrial misconduct	0.243	0.128	0.057	0.105
	(0.0005)	(0.0107)	(0.0016)	(0.0017)
At least one previous conviction	0.629	0.809	0.892	0.824
	(0.0006)	(0.0132)	(0.0022)	(0.0021)
No previous conviction	0.371	0.192	0.108	0.176
	(0.0006)	(0.0130)	(0.0022)	(0.0021)
High severity (previous case)	0.510	0.648	0.743	0.664
	(0.0006)	(0.0153)	(0.0032)	(0.0025)
Low severity (previous case)	0.490	0.357	0.257	0.336
	(0.0006)	(0.0153)	(0.0032)	(0.0025)
High severity (current case)	0.515	0.805	0.999	0.993
	(0.0006)	(0.0116)	(0.0001)	(0.0004)
Low Severity (current case)	0.485	0.156	0.001	0.007
	(0.0006)	(0.0115)	(0.0001)	(0.0004)
Drug crime	0.124	0.167	0.014	0.064
	(0.0004)	(0.0124)	(0.0008)	(0.0013)
Non-drug crime	0.876	0.824	0.986	0.936
	(0.0004)	(0.0138)	(0.0008)	(0.0013)
Property crime	0.183	0.082	0.001	0.005
	(0.0004)	(0.0095)	(0.0001)	(0.0004)
Non-property crime	0.817	0.913	0.999	0.995
	(0.0004)	(0.0094)	(0.0001)	(0.0004)

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 (200 repetitions).