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

Nicolás Grau[†]   Damián Vergara[‡]

First version: January 27, 2020
This version: July 14, 2021

Abstract

We propose an observational implementation of the outcome test that uses predicted selection status to identify marginal individuals. We provide conditions under which selected individuals with lower propensity scores are more likely to be marginal given their observables and propose empirical diagnostics to assess their plausibility. Our approach requires neither instruments nor the random assignment of decision-makers, allows for unrestricted correlation between observables and unobservables, and can accommodate non-monotone patterns of discrimination. We illustrate our method by analyzing prejudice in pretrial detentions against the Mapuche, the largest ethnic minority group in Chile, and find strong evidence of prejudice against them.

[*] This paper previously circulated under the title "A Simple Test for Prejudice in Selection Processes: The Prediction-Based Outcome Test." 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 Diego Amador, Hadar Avivi, Aureo De Paula, Marina Dias, Will Dobbie, Joseph Doyle, Peter Hull, Michael Jansson, Felipe Jordán, Patrick Kennedy, Patrick Kline, Attila Lindner, Charles Manski, Pablo Muñoz, Demian Pouzo, James Powell, Jesse Rothstein, Yotam Shem-Tov, Petra Todd, Chris Walters, Harry Wheeler, and John Wieselthier, as well as seminar participants at the 11th Annual Conference of the Spanish Association of Law and Economics, the Bolivian Conference on Development Economics, the Chicago/LSE Conference on the Economics of Crime and Justice (2021), CReAM/UCL, the Econometric Society World Congress (2020), the FROGEE Workshop on Gender Economics, King's College, the Online Economics of Crime Seminar (Summer 2020), PUC-Chile (Centro de Justicia y Sociedad), Toulouse School of Economics, UC Berkeley, Universidad de Chile, Universidad Mayor, University of Oxford, and Wilfrid Laurier University, for very helpful discussions and suggestions. We also thank Tomás Cortés for excellent research assistance. Nicolás Grau thanks the Centre for Social Conflict and Cohesion Studies (ANID/FONDAP/15130009) for financial support. 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

Many selection processes are based on predicted outcomes. For example, bail judges decide defendants' pretrial detention status based on expected pretrial misconduct if released. Normative and positive considerations suggest that prejudiced selection processes –that is, situations where decision-makers routinely set different effective selection thresholds for members of a particular group because of animus or systematic mispredictions of their expected outcomes– are problematic. However, testing for prejudice in selection processes is empirically challenging. A prominent approach is the outcome test (Becker, 1957, 1993) which is based on the idea that, if judges are not prejudiced, marginally released defendants from different groups should have equal pretrial misconduct rates. Then, testing for prejudice is reduced to comparing the average outcome of marginally selected individuals between groups, that is, to a simple difference in means.

While the outcome test has desirable properties, its implementation induces an empirical challenge: 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. The literature has taken different approaches to deal with this identification problem (see Hull, 2021 and Section 2 for a discussion). Quasi-experimental solutions usually rely on random assignment of decision-makers, which is unlikely to hold in many settings. On the other hand, observational proposals imply structural assumptions that may be seen as too restrictive in most applications. A more flexible observational approach for cases when instruments are unavailable is therefore missing in the literature.

This paper proposes a novel observational implementation of the outcome test, the Prediction-Based Outcome Test (P-BOT), that uses the predicted selection status (i.e., the propensity score) to identify marginal individuals. We motivate our approach with a model where judges decide over defendants' pretrial release status based on expected pretrial misconduct (i.e., non-appearance in court or pretrial recidivism). Our formal notion of prejudice is based on aggregate differences in effective selection thresholds across groups and accounts for judges' preferences and biased beliefs. This definition is closely aligned with Arnold, Dobbie, and Yang (2018) and Hull (2021) but differs from Canay, Mogstad, and Mountjoy (2020) who use a stricter characterization that compares thresholds after equalizing all observable characteristics across groups. We formally show that the outcome test is valid under our definition of prejudice and develop a critical discussion of its interpretation and normative relevance under different plausible scenarios.

Our main theoretical contribution is to provide sufficient conditions under which the released individuals that are more likely to be marginal given their observables also have lower propensity scores. That result reduces the challenge of identifying marginal individuals to a standard prediction problem, simplifying the implementation of the outcome test. The econometrician has to estimate the propensity score, rank released individuals according to their predicted probabilities to define samples of marginal individuals, compute group-specific pretrial misconduct rates within these samples of marginals, and perform a difference in means.

Relying on predicted values implies that the P-BOT is robust to omitted variables, since the structural interpretation of the prediction coefficients is not relevant. This intuition is corroborated by Monte Carlo simulations. Note that the argument for prediction-based identification of marginally selected individuals assumes the availability of good predictors, since the noise in the estimated ranking can induce bias in the outcome test. However, the predictive power of the observed covariates can be assessed by looking at the fit of the propensity score. We also propose a perturbation test to empirically assess the pervasiveness of this potential source of bias.

Our identification strategy is based on two assumptions. First, we assume that the selection equation has an additively separable representation between observables and unobservables. Through the lens of the model, this induces monotonicity on observables in the risk probabilities, meaning that the marginal effect of observables on latent pretrial misconduct does not depend on the unobserved component. Second, while we allow for unrestricted first moments in the joint distribution of observables and unobservables, which is an important improvement relative to the observational literature, we require the patterns of heteroskedasticity to hold a monotonicity property. We propose diagnostics to empirically assess the plausibility of both assumptions. We highlight that our test does not rely on random assignment of decision-makers and can accommodate non-monotone patterns of discrimination, appearing as an attractive alternative in situations where instrument-based approaches cannot be properly implemented.

As an application of the P-BOT, we test for prejudice against the largest ethnic minority group in Chile, the Mapuche, using nationwide administrative data. According to the last census, around 10% of the Chilean population reported themselves as being Mapuche. The Mapuche population is an interesting case of analysis for three reasons. First, a long-running conflict exists between the Mapuche and the Chilean state, dating back more than a century (Cayul et al., 2018). In this context, it is frequently claimed that Chilean institutions are biased against the Mapuche. Second, the Mapuche people are subject to numerous negative stereotypes, such as tendencies towards laziness, violence and alcoholism, from some quarters of Chilean society (Merino and Quilaqueo, 2003; Merino and Mellor, 2009). There is no evidence for any systematic difference in behavior between the Mapuche and the rest of the population. Third, Mapuche people are identifiable, mainly because of their surnames but also to some extent due to their physical appearance. Thus, discrimination against members of this group is feasible in this setting.

We use nationwide administrative data that covers more than 95% of criminal cases in Chile between 2008 and 2017. The data contains detailed information on cases and defendants and includes judges and attorneys identifiers. We merge the administrative records with a register of Mapuche surnames to create different measures of ethnicity that combine self-reporting and surname information. We provide evidence that suggests both our identification assumptions hold in this setting and implement the P-BOT by fitting different projection models for the release status using a wide set of predictors.

Results provide strong evidence of prejudice against Mapuche defendants in pretrial detention decisions. Our preferred specification shows that marginal Mapuche defendants are between 3 and 4 percentage points less likely to be engaged in pretrial misconduct relative to marginal non-Mapuche defendants. 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, Persico, and Todd, 2001) also suggests prejudice against Mapuche defendants, although the implied magnitude is smaller.

Since the Chilean setting is characterized by quasi-random assignment of judges for arraignment hearings at the court-by-time level, we also test for prejudice using the instrument-based approach proposed by Arnold, Dobbie, and Yang (2018). While the LATE for the non-Mapuche sample of defendants is precisely estimated, we show that the estimation is severely underpowered for the Mapuche sample. This prevents us from drawing precise conclusions from its application. We also perform the test proposed by Frandsen, Lefgren, and Leslie (2019) and systematically reject the null hypothesis of valid LATE assumptions. The fragility of the IV estimation in our setting illustrates that the P-BOT is an attractive alternative when the instrument-based approach cannot be properly implemented. Encouragingly, the LATE for the non-minority sample is similar to the P-BOT estimates of non-Mapuche pretrial misconduct rates at the margin, and the non-minority marginal defendants identified by the two methods have similar distributions of observables. This suggests that both approaches give similar results when both are expected to work properly.

We conclude the paper by estimating extensions that relate to the normative discussion on the definition of prejudice. First, we explore more complex patterns of prejudice by including additional regressors in the outcome equation. We present two examples that group defendants into two categories. In the first, we group defendants using *Mapuche* and *low income*, conjecturing that the discrimination patterns may interact with socioeconomic status. In the second, considering the geographical component of the Mapuche conflict, we group defendants using *Mapuche* and *Mapuche region*, conjecturing stronger discrimination patterns in those courts.[1] Our results show

[1] Colloquially, and for the purposes of this paper, the Mapuche region is the name given to the Araucanía Region, the Chilean administrative region that is the heartland of the indigenous Mapuche people and historically associated with the Mapuche conflict.

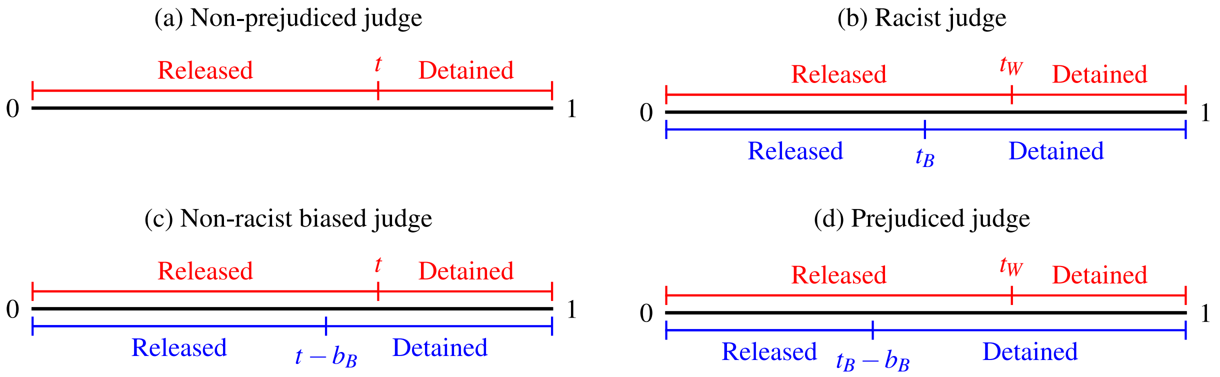
that prejudice patterns are stronger for Mapuche defendants that live in low-income municipalities and that, while there is prejudice against Mapuche defendants in all Chilean courts, it is slightly stronger in the Mapuche region. These results suggest that non-monotone patterns of discrimination are likely to occur in practice. Second, we estimate the outcome equations controlling for court-by-time fixed effects (the level at which judges are randomly assigned) and find that around a half of the overall effect is explained by the assignment rule of judges to defendants (that is, by Mapuche defendants being systematically assigned to courts with stricter judges).

This paper contributes to the literature on discrimination by proposing a simple methodology to test for prejudice (Guryan and Charles, 2013; Lang and Kahn-Lang, 2020; Small and Pager, 2020). More specifically, it adds to the literature that discusses the properties and the implementation of the outcome test; namely, Knowles, Persico, and Todd (2001), Anwar and Fang (2006), Arnold, Dobbie, and Yang (2018), Canay, Mogstad, and Mountjoy (2020), Feigenberg and Miller (2020), Marx (2020), Gelbach (2021), and Hull (2021). Throughout the paper, we argue that our approach is particularly appealing in settings when instrument-based approaches are weak or infeasible, thus constituting a complement to the existing literature.

Our empirical application also adds to a vast body of evidence on bias in different levels of the criminal justice system. See, for example, Knowles, Persico, and Todd (2001), Anwar and Fang (2006), Antonovics and Knight (2009), Abrams, Bertrand, and Mullainathan (2012), Anwar, Bayer, and Hjalmarsson (2012), Rehavi and Starr (2014), Simoiu, Corbett-Davies, and Goel (2017), Anwar, Bayer, and Hjalmarsson (2018), Cohen and Yang (2019), Fryer (2019), Durlauf and Heckman (2020), Feigenberg and Miller (2020), Arnold, Dobbie, and Hull (2020), Marx (2020), and Rose (2020). The paper more related to ours is Arnold, Dobbie, and Yang (2018), who find that bail judges are prejudiced against black defendants. Understanding racial disparities in incarceration is important beyond the normative concerns they raise because incarceration negatively affects employment, future crime, and education (Aizer and Doyle, 2015; Muller-Smith, 2015; Cortés, Grau, and Rivera, 2019). More specifically, pretrial detention affects conviction rates, employment, and the use of state benefits (Leslie and Pope, 2017; Dobbie, Goldin, and Yang, 2018; Grau, Marivil, and Rivera, 2019). The potential existence of prejudice in judicial decisions, therefore, is particularly costly from both a private and social perspective.

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

Figure 1: Selection Rule: Examples



Note: In each panel, the horizontal black line accounts for the domain of the true pretrial misconduct probability.

2 Preliminaries: Prejudice and the Outcome Test

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

2.1 Prejudice

In this paper, we analyze potential prejudice in selection rules that are based on expected outcomes. To fix ideas, consider a situation where judges decide whether or not to grant pretrial release for a defendant. Each judge has to predict the likelihood that the defendant will be engaged in pretrial misconduct (non-appearance in court or pretrial recidivism) if released during the investigation, compare that to a threshold, and make a decision. Given the legal principle of the presumption of innocence, judges should not detain defendants unless the expected risk of pretrial misconduct is significant. The question we address is whether there is prejudice against a specific group (e.g., black defendants) in the release decision.

Figure 1 illustrates a stylized selection rule for an individual judge. Panel (a) shows what the selection rule looks like for a non-prejudiced judge. The judge predicts the probability of pretrial misconduct using all the available information (including race) and releases defendants whenever that predicted probability is smaller than t. Panel (b) shows what the selection rule looks like for a racist judge. Because of animus, the judge sets a smaller threshold for black defendants with t_W and t_B being the thresholds set for white and black defendants, respectively. In this case, the selection rule is discriminatory against black defendants, given that only white defendants are released when the pretrial misconduct probability is between t_B and t_W. Now suppose the judge is non-racist, but systematically overestimates risk for black defendants: when the true probability of pretrial misconduct is p, the judge predicts $p + b_B$ if the defendant is black. This situation, illustrated in Panel (c), implies that the effective threshold is smaller for black defendants. This selection rule is also discriminatory against black defendants since defendants with pretrial misconduct probability between $t - b_B$ and t are released depending on their race. Finally, Panel (d) shows 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 animus (or taste-based discrimination) and biased beliefs (or inaccurate statistical discrimination). The framework we develop, as usual in this literature, is not able to separately identify between both sources of prejudice (Arnold, Dobbie, and Yang, 2018; Bohren et al., 2020; Hull, 2021).[2]

For an individual defendant, the stylized selection rule can be formalized by the following threshold-crossing model:

$$Release_i \quad = \quad 1\{p(G_i, Z_i) \leq h(G_i, Z_i, j(i))\}, \tag{1}$$

where i indexes defendants and j judges, $j(i)$ is a function that assigns judges to defendants, G_i is a group indicator variable (e.g., race), Z_i is a vector of characteristics of defendant i observed by the judge (e.g., type of crime and criminal record), $p(G_i, Z_i)$ is the true conditional probability of pretrial misconduct if released of defendant i, and $h(G_i, Z_i, j(i))$ is the effective threshold that can vary with G_i and Z_i because of animus or biased beliefs (or both), and is potentially heterogeneous across judges. In Appendix A, we present a very simple model that adds structure to the judge problem in the spirit of Figure 1 that works as a microfoundation of (1).[3]

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

DEFINITION 1 (PREJUDICE). *In the absence of prejudice*

$$\overline{h}(0) \quad = \quad \overline{h}(1). \tag{2}$$

[2] Note that the –statistically accurate– use of race for computing pretrial misconduct probabilities can be labeled as accurate statistical discrimination, which is a relevant (and, in many settings, illegal) source of discrimination (Kleinberg et al., 2019; Arnold, Dobbie, and Hull, 2020; Kline and Walters, 2020; Yang and Dobbie, 2020). While being robust to its presence, the analysis we develop is not informative of statistical discrimination.

[3] The exclusion of $j(i)$ from p is without loss of generality and made only for presentation purposes. Specifically, allowing $j(i)$ to enter p means that the assigned judge may have an impact on the probability of pretrial misconduct if released, which could be confused with p being a judge-specific prediction.

It follows that the decision process is prejudiced against defendants of group 1 whenever $\overline{h}(0) > \overline{h}(1)$. Note that this definition can be extended to a non-binary discrete G_i.

Our definition of prejudice is closely aligned with Arnold, Dobbie, and Yang (2018) and Hull (2021), but differs from Canay, Mogstad, and Mountjoy (2020). In particular, Canay, Mogstad, and Mountjoy (2020) say a judge j is racially unbiased if $h(0, j(i), Z_i) = h(1, j(i), Z_i)$, for all Z_i, so they equalize non-race characteristics at the moment of defining race-based prejudiced decision-making. This difference is not driven by the modeling choice, since equation (1) also allows discrimination patterns to depend on non-race characteristics. Instead, the difference follows a normative decision on the relevant notion of prejudice. We warn readers that are more sympathetic with Canay, Mogstad, and Mountjoy (2020) that some of the discussions developed below may not be valid under their definition of bias.

There are several reasons why differences in *average* effective thresholds may not necessarily reflect the intuition depicted in Figure 1. In what follows, we discuss these reasons and argue that the alternative interpretations remain relevant from a normative point of view.

The role of Z_i in h Average thresholds integrate over non-race characteristics. If defendants of different groups have different distributions of Z_i, the differences in average thresholds could be recovering prejudice based on other characteristics. For example, suppose that judges do not care about race but discriminate based on place of living. If race is correlated with place of living, then Definition 1 can be violated even if effective thresholds do not depend on race.

We think this distinction is of second-order from a normative point of view since it is still the case that defendants of certain races are more frequently imprisoned for reasons unrelated with their probability of pretrial misconduct. This distinction, however, is of first-order when using Canay, Mogstad, and Mountjoy (2020) definition. Importantly, the approach introduced in the next section allows to test for patterns of prejudice that simultaneously depend on G_i and other observed variables and is also useful to test for the exclusion restrictions needed for the outcome test to identify Canay, Mogstad, and Mountjoy (2020) notion of prejudice.

Assignment of judges to defendants The related empirical literature usually focuses on cases where $j(i)$ is characterized by quasi-random assignment of judges to defendants. One of the advantages of the approach introduced in the next section is that it does not need random assignment of judges for identification. However, the nature of the assignment rule matters for interpreting differences in effective thresholds. To see why, consider two polar cases. In the first case, judges are completely unbiased (and hence the only variation in the effective thresholds comes from heterogeneity in the idiosyncratic leniency of judges), but stricter judges are systematically assigned to black defendants. In the second case, all judges are prejudiced against black defendants, but there is random assignment of judges. In both cases, Definition 1 is violated, but with different interpretations. The interpretation in the second case aligns with the intuition of Figure 1, while the first case reflects a situation where $j(i)$ can be said to be prejudiced.

Again, we believe both situations are relevant from a normative point of view, although Canay, Mogstad, and Mountjoy (2020) definition does not capture the first case. In our empirical application we show how our approach can be used to decompose between both sources of prejudice when judges are quasi-randomly assigned at some lower level (e.g., court-by-time).

Alternative objective functions The starting point of the analysis is that judges make (or, at least, should make) decisions based on predicted pretrial misconduct if released. It could be the case, however, that judges have different objective functions. This is related to the notion of "omitted payoff bias" defined in the literature of algorithmic decision-making (Kleinberg et al., 2018). The nature of the alternative objective functions determines the implications for our definition of prejudice. To see why, consider the following two cases. In the first case, judges are mandated by law to make decisions based on potential pretrial misconduct. However, in order to increase their chances of a promotion, they attempt to please their superiors. Thus, if their superiors demonstrate racist tendencies, these judges will routinely release white defendants and detain black defendants regardless of their predicted risk. As in the previous considerations, we see this subtlety as second-order since in this scenario is still the case that some defendants are discriminated against with respect to the normative standard provided by law. In the second case, consider an institutional setting that mandates by law the use of pretrial detention to all defendants that have prior convictions. Here, an unbiased selection process has different implications for effective thresholds as long as the distribution of prior convictions varies by group.[4]

The bottom line is that if the mandated selection rule is well defined, then individual deviations do not affect the normative relevance of our definition of prejudice. Moreover, we show how the approach presented in the next section can be used to indirectly assess if judges care about potential pretrial misconduct when making the release decisions.

2.2 Outcome test

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

[4] For example, Manski (2005, 2006) develops a model of police profiling where, if the deterrent effects of police searches vary by group, then the effective thresholds may be optimally different for reasons unrelated to discrimination.

its application is challenging since effective thresholds are rarely observable.

One approach used to overcome this challenge is the *outcome test* (Becker, 1957, 1993), which is based on the success rates at the margin of the selection process. To understand the intuition, consider the selection rule illustrated in Panel (a) of Figure 1. Define the marginally released defendants as defendants with true probability of pretrial misconduct equal to t (i.e., defendants that were released on a borderline decision). In expectation, $t\%$ of marginally released defendants should be engaged in 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 in some type of pretrial misconduct. Then, if there is prejudice in the selection 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 needs only to find a statistically significant correlation between pretrial misconduct and race for the defendants at the margin.

To formally define the outcome test, let the latent release status be given by $Release_i^* = h(G_i, Z_i, j(i)) - p(G_i, Z_i)$, hence $Release_i = 1\{Release_i^* \geq 0\}$. We say that a released defendant is marginal if $Release_i^* = 0$. The next proposition establishes that observed average behavior of marginal individuals of a given group coincides with the average effective threshold.

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

$$\mathbb{E}[PM_i | G_i = g, Release_i^* = 0] \quad = \quad \overline{h}(g). \tag{3}$$

Proof. See Appendix B.

In Proposition I, the expectation integrates across judges and non-race characteristics. Putting together Definition 1 and Proposition 1 formalizes the outcome test.

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

$$\mathbb{E}[PM_i | G_i = 0, Release_i^* = 0] \quad = \quad \mathbb{E}[PM_i | G_i = 1, Release_i^* = 0]. \tag{4}$$

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 selection process is prejudiced against group 1. Note that to properly perform this test the econometrician does not need to identify the causal effect of release status or group membership on pretrial misconduct. To reject the null hypothesis of no prejudice, only is required a statistically significant correlation between pretrial misconduct and group membership for the defendants at the margin.

Identification of marginal individuals While the outcome test implementation does not require observing effective thresholds, it induces an additional empirical challenge. The difference in means described above can be trivially implemented when knowing which released defendants are marginal. However, identifying who is marginal is challenging for the econometrician. This is important because the misspecification of marginal individuals may induce bias in the outcome test: when the risk distributions differ between groups, differences in pretrial misconduct rates computed away from the margin may not be informative about effective thresholds and, therefore, may result in misleading conclusions regarding prejudice. This is called the *inframarginality bias*.[5]

A solution that avoids imposing strong assumptions on judge behavior and the distribution of unobservables is proposed by Arnold, Dobbie, and Yang (2018). If the econometrician has an instrument for the release status, pretrial misconduct rates at the margin can be recovered by the expected treatment effects at the margin of release. Then, the outcome test can be implemented by comparing group-specific LATEs.[6] By exploiting quasi-random assignment of bail judges, the authors propose to use judge-specific leave-out mean release rates as an instrument. One problem with this approach is that it is equivalent to running a first-stage on judge fixed effects. This may induce power problems in settings where minority groups represent small shares of the population. Also, as emphasized by Muller-Smith (2015) and Frandsen, Lefgren, and Leslie (2019), the leave-out mean release rate may fail to meet the LATE monotonicity assumption.[7]

There are many situations, however, where decision-makers are not quasi-randomly assigned and alternative instruments are unavailable, or when power problems or non-monotone judge behaviors are likely to make the judges-design infeasible. These situations call for observational approaches to deal with the inframarginality bias. An example is Chandra and Staiger (2010), that derive an observational test for prejudice that relies on selection-on-observables assumptions. Another influential example is Knowles, Persico, and Todd (2001). In the context of motor vehicle searches for contraband, the authors model equilibrium conditions under which the marginally searched individuals demonstrate the same behavior as the average ones. Here, linear regressions of the outcome equation using the full sample of selected individuals are enough to test for prejudice. However, Anwar and Fang (2006) argue that Knowles, Persico, and Todd (2001) approach is affected by the inframarginality bias and, as noted by Arnold, Dobbie, and Yang (2018), the validity of OLS for this problem requires very strong distributional restrictions.

In the context of this discussion, we propose a novel observational approach to identify

[5] Section 2.2. of Simoiu, Corbett-Davies, and Goel (2017) and the Online Appendix C in Arnold, Dobbie, and Yang (2018) provide intuitive explanations of the inframarginality bias.

[6] Hull (2021) shows that the outcome test is equivalent to the difference between group-specific MTE frontiers.

[7] Arnold, Dobbie, and Hull (2020) develop a hierarchical MTE model that imposes additional structure to allow for deviations from strict monotonicity.

marginally released defendants. Our approach requires neither a valid instrument nor quasi-random assignment of judges for its implementation and allows for non-monotonicities in judge behavior, at the cost of assumptions that we argue are weaker than the implied restrictions of alternative observational approaches. Thus, we believe our approach is an attractive alternative in settings where the instrument-based approach cannot be properly implemented.

3 The Prediction-Based Outcome Test

In this section, we describe our observational proposal for identifying marginal individuals to implement the outcome test: the Prediction-Based Outcome Test (P-BOT). We discuss identification and estimation, as well as the virtues and weaknesses of our method.

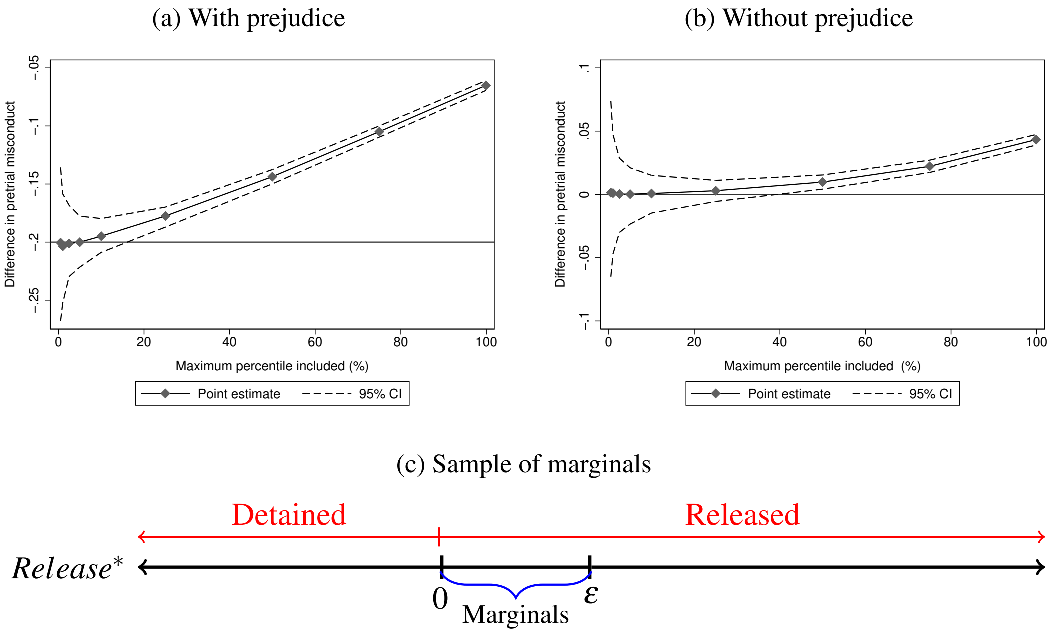
Notation In what follows, we classify all variables that affect the release decision into variables that are observed by the econometrician, X_i, and variables that are not, V_i. With this notation we can write $Release_i^* = f(X_i, V_i)$, where f is some function, so $Release_i = 1\{f(X_i, V_i) \geq 0\}$. Through the lens of equation (1), f takes a particular form. However, to make the analysis robust to alternative modeling choices, and for not taking an ex-ante stand on what is observed by the econometrician and what is not, we derive the analysis using this more general notation. The only variable we impose to belong to X_i is G_i since the identification of marginal individuals is, ultimately, an input for testing for prejudice against G_i, so not observing G_i makes the exercise meaningless.

3.1 Intuition

Suppose the econometrician wants to identify the defendants that were marginally released. Absent any guidance, the econometrician can compare the average behavior of released defendants. However, if risk distributions vary by group, differences in averages may be uninformative of the behavior at the margin. Our method helps the econometrician to restrict the sample of released defendants to the ones that are more likely to be marginal given their observables, so averages computed using these subsamples are less likely to be affected by the inframarginality bias.

Thought experiment Our approach tries to mimic the following thought experiment: Suppose the econometrician observes the latent release status, $Release_i^*$. If so, the latent variable can be used to rank released individuals and define arbitrary notions of the margin. Specifically, the lower $Release_i^*$ (conditional on $Release_i = 1$), the closer to the margin, so the inframarginality bias can be

Figure 2: Prediction-Based Outcome Test: Intuition


(a) With prejudice

(b) Without prejudice

(c) Sample of marginals

Note: Panels (a) and (b) use simulated data based on the model presented in Appendix A. See Appendix D for details on the simulation. The x-axis measures the maximum percentile of $Release^*$ considered for computing the difference in pretrial misconduct rates between races. That is, 100 means that the entire sample of released defendants is considered, 75 that only the 75% with lower $Release^*$ is considered, etc. The point estimates are the mean estimation across 200 Monte Carlo simulations. Confidence intervals correspond to the 2.5 and 97.5 percentiles of the simulations.

attenuated by excluding the observations with the larger values of $Release^*$. Based on this intuition, suppose that the econometrician labels released 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 could then be easily implemented by regressing PM_i on G_i within the sample of marginally released defendants.

Panels (a) and (b) of Figure 2 illustrate this intuition. Both figures use simulated data based on the model presented in Appendix A with group-specific distributions (see Appendix D for details). Panel (a) considers a case where there is prejudice with a corresponding difference in effective thresholds of 0.2 in favor of white defendants. Panel (b) considers a case with no prejudice, so the difference in effective thresholds is zero. The y-axis measures differences in pretrial misconduct between white and black defendants, while the x-axis considers different values of $\bar{q}$. Both figures show that, in this particular example, using the whole sample of released defendants gives wrong conclusions regarding prejudice, but that the differences in pretrial misconduct rates converge to the differences in effective thresholds as $\bar{q}$ decreases.

Following this intuition, marginally released defendants can be thought of as defendants with $Release^* \in [0, \varepsilon]$, with $\varepsilon > 0$ small. Panel (c) of Figure 2 illustrates this definition. Since $Release^*$ is truncated at 0 for released defendants, under a full support assumption, identifying the released defendants with $Release^* \in [0, \varepsilon]$ is equivalent to identifying the released defendants with the smaller latent indexes. Then, identifying a ranking of $Release^*$ among released defendants enables the creation of samples of marginal defendants.

Certainly, $Release^*$ is unlikely to be observed. If there are variables that judges use to make release decisions that the econometrician does not observe, $Release^*$ is also difficult to estimate. The econometrician, however, can try to identify the defendants more likely to have lower latent release indexes given their observables. This is what our approach does.

The P-BOT Assume the econometrician is interested in identifying released defendants that are more likely to be close to the margin given their observables, i.e., released defendants with large values of $\Pr\left(Release_i^* < \varepsilon | X_i, Release_i = 1\right)$. Then, following the logic of the thought experiment, the econometrician could rank released defendants based on this conditional probability and label as marginals the ones with the larger values of $\Pr\left(Release_i^* < \varepsilon | X_i, Release_i = 1\right)$.

Given that the distribution of V_i conditional on X_i is unknown, the econometrician cannot compute the aforementioned conditional probabilities without additional assumptions. In what follows, however, we provide sufficient conditions under which the ranking of released defendants based on $\Pr\left(Release_i^* < \varepsilon | X_i, Release_i = 1\right)$ is identified by the ranking of the predicted release probabilities, $\mathbb{E}\left[Release_i | X_i\right]$ (i.e., the propensity score). 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 $\mathbb{E}\left[Release_i | X_i\right]$, which can be achieved by fitting flexible projection models. In a sense, the identification of marginal individuals is reduced to a prediction problem. It is because of this feature that we call our method the Prediction-Based Outcome Test: prediction (rather than causal) models help solving the problem of identifying marginal individuals.

3.2 Identification of marginal individuals

Now we formalize the identification argument sketched above.

Assumptions Throughout the analysis, we make the following assumption:

ASSUMPTION 0 (A0). *The joint distribution of X_i and V_i is continuous and has full support.*

We need A0 for the rank-argument to work. Our identification argument identifies relative distance to the margin across defendants with different observables, so simply put, A0 implies that *the more marginals* are effectively marginals. Note, however, that this assumption is also needed for the outcome test to make sense. If there are no marginals, then it is not possible to estimate the conditional expectations at the margin. In this instance, we see A0 as a regularity condition for the more general idea of the outcome test, rather than a specific assumption for our approach.

To prove identification, we make two additional 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 additively separable representation of the selection equation. A1 can be empirically assessed by regressing $Release_i$ on X_i in samples of defendants with (presumably) different unobservables and comparing the estimated coefficients. Also, recall from equation (1) that $f(X_i,V_i) = h(X_i,V_i) - p(X_i,V_i)$. So through the lens of our model, sufficient conditions that do not require exclusion restrictions are given by (i) $h(X_i,V_i) = h_X(X_i) + h_V(V_i)$, and (ii) $p(X_i,V_i) = p_X(X_i) + p_V(V_i)$. While (i) is not testable, (ii) implies monotonicity on observables in the expected risk equation, and therefore can be empirically assessed by regressing PM_i on X_i in samples of released defendants with (presumably) different unobservables. Intuitively, changes in X_i should move the latent risk in the same direction for every defendant, regardless of the realization of V_i. We discuss both tests in Appendix F and illustrate them in our empirical application.

A1 imposes restrictions on the joint effect of X_i and V_i on the decision rule. It does not, however, impose restrictions on their joint distribution. The required distributional restrictions are summarized in A2.

ASSUMPTION 2 (A2). *The structure of W_i is given by $W_i = r_1(X_i) + r_2(X_i)\zeta_i$, with ζ_i scalar, independent from X_i, and with log-concave cdf, $r_2(X_i) > 0$ for all X_i, and r_2 non-increasing in the expected distance from the margin.*

Log-concavity is a standard regularity condition, and assuming that the unobserved component is of the form $W_i = \lambda(X_i,\zeta_i)$ with ζ_i scalar and λ strictly increasing in ζ_i has been assumed for identification in other contexts (e.g., Imbens and Newey, 2009). Assuming that λ is linear in ζ_i is a stronger restriction. However, note that linearity still can accommodate fairly general dependence structures, since the conditional mean and variance of W_i given X_i are unrestricted. In our view, the restrictive element of A2 is the monotone behavior of $r_2(X_i)$. This restriction states that the volatility of the unobservables cannot be larger for released defendants that are less likely to be marginal given their observables. This restricts the patterns of heteroskedasticity.

While we acknowledge the restrictiveness of A2, two things are worth discussing. First, A2 is weaker than selection-on-observables or stronger independence assumptions since it allows for unrestricted conditional first moments (i.e., for any correlation level between W_i and X_i) and can accommodate some forms of heteroskedasticity. Then, we argue this assumption constitutes an improvement relative to the literature in the absence of plausible exogenous variation. Second, the monotone behavior of $r_2(X_i)$ is a sufficient but not necessary condition. In Appendix C we present examples that suggest that deviations from this restriction should be large to invalidate identification. That is, conditional second moments should be strongly increasing in the expected distance from the margin in order to compromise identification. We think this alleviates potential concerns regarding A2. While this restriction is not directly testable, in Appendix F we propose a test to empirically assess our identification argument, and illustrate it in our empirical application.

Discussion To assess the restrictiveness of both assumptions, it is illustrative to compare them to the assumptions required by other methods. Alternative observational approaches rely on stronger restrictions. Chandra and Staiger (2010) approach is identified under selection on observables, which is stronger than A2. Knowles, Persico, and Todd (2001) assumptions rely on a behavioral model of police search for contraband and, therefore, a comparison to our assumptions is less direct. However, their recommendation of using the average behavior of selected individuals imply strong restrictions on the conditional distributional of unobservables to avoid inframarginality bias (equal risk distributions across groups or constant treatment effects across the risk distribution, see Arnold, Dobbie, and Yang, 2018). In this regard, we see our sufficient conditions as an improvement relative to the observational literature, making the P-BOT an attractive alternative in the absence of quasi-experimental variation.[8]

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

[8] It is important to note that A1 and A2 do not imply the absence of inframarginality bias. To the extent that the distribution of (X_i, V_i) varies with G_i, it is still the case that both groups may have very different risk distributions.

Identification Proposition II summarizes the identification argument.

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

$$\begin{aligned} \Pr\left(Release_i^* \leq \varepsilon | X_i = x_1, Release_i = 1\right) &> \Pr\left(Release_i^* \leq \varepsilon | X_i = x_2, Release_i = 1\right) \\ \iff \quad \mathbb{E}\left[Release_i | X_i = x_1\right] &< \mathbb{E}\left[Release_i | X_i = x_2\right]. \end{aligned} \tag{5}$$

Proof. See Appendix B.

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 identifies the relative distance to the threshold in probability. The result produces two aspects that warrant further discussion.

Prediction The identification argument relies on the predicted release status but not on the specifics of the prediction model. This makes our approach robust to omitted variable bias. That is, as V_i is not observed, it biases the estimated coefficients of the prediction model, but the same bias improves the prediction of the conditional expectation. In fact, Monte Carlo exercises presented in Appendix D show that the P-BOT behaves better when the correlation between observables and unobservables is large, in particular, by increasing precision. This implies that omitted variables do not bias the estimation of the expected proximity to the margin. The reason is that the econometrician only needs to know *who* are close to the margin, not *why* they are close.

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 W_i (i.e., the variance of ζ_i). Variance in the estimated ranking implies that inframarginal defendants are potentially included in the sample of marginals. As a consequence, the noise in the estimated ranking may generate inframarginality bias. This suggests that an implicit assumption in the application of our method is the availability of good predictors. In Appendix D we present Monte Carlo simulations that show that as this measurement error increases, our test converges to Knowles, Persico, and Todd (2001)'s test. Intuitively, when the predictive power of X_i is very weak, the ranking of predicted probabilities flattens and the sample of marginals converges to a random sample of released individuals.

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. Specifically, the selection rule can be written as $Release_i =$

$1\{Release_i^* \geq 0\} = 1\left\{\frac{d(X_i)-r_1(X_i)}{r_2(X_i)} \geq \zeta_i\right\}$. Since the econometrician observes $Release_i$ and X_i, it is possible to estimate the left-hand-side and the variance of ζ_i. The estimated variance of ζ_i can be then used to simulate perturbations that alter the estimated ranking and, therefore, the defendants that are considered to be marginals. By recomputing the outcome test on each of these simulations, the econometrician can check how the test varies with the perturbations. In the next subsection we describe in more detail how to implement this test.[9]

3.3 Estimation and implementation

Proceeding as per the thought experiment, the econometrician can estimate the propensity score, use the predicted release probabilities to rank released defendants, and estimate the outcome equation on a sample of defendants at a given margin definition. We propose two approaches for implementing the P-BOT. To simplify notation, let $\hat{R}_i$ denote the estimated propensity score.

Simple approach This approach involves defining the sample of marginally released individuals based on the quantiles 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 and significant estimates of the coefficient on G_i constitute evidence of prejudice against group $G_i = 1$. Note that there is a 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 also implies that the outcome equation is estimated using a larger share of inframarginal individuals. This leads to a natural inframarginality test: the econometrician can assess the pervasiveness of the inframarginality problem by analyzing the sensitivity of the estimation to the choice of $\overline{q}$.

Note that testing for more complex patterns of prejudice can be easily done by adding discrete regressors to the outcome equation. Moreover, if there is quasi-random assignment of judges to defendants after the appropriate controls (e.g., court-by-time fixed effects), including them may help the econometrician to assess the extent of overall estimated prejudice that is driven by the assignment rule. We illustrate these extensions in our empirical application.

Non-parametric approach As a refinement, we suggest performing non-parametric local regressions to estimate $\mathbb{E}\left[PM_i|G_i = 0, q(\hat{R}_i) = 1\right]$ and $\mathbb{E}\left[PM_i|G_i = 1, q(\hat{R}_i) = 1\right]$, and to assess the

[9] Note that this source of bias does not depend on the relative sample sizes of the different groups since the prediction model is estimated using all defendants. Small sample sizes may induce noise in the estimated conditional expectations in the outcome equation, which is implicitly captured by the confidence intervals of the outcome test.

extent of prejudice by computing $\mathbb{E}\left[PM_i|G_i=1,q(\hat{R}_i)=1\right]-\mathbb{E}\left[PM_i|G_i=0,q(\hat{R}_i)=1\right]$.[10] An advantage of this approach is that it weights observations according to their relative distance to the margin definition.

Weights As Arnold, Dobbie, and Yang (2018) and Hull (2021) note, the conditional expectations at the margin can be recovered after estimating MTEs of $Release_i$ on PM_i. Specifically, following Zhou and Xie (2019) notation, the Marginal Policy Relevant Treatment Effect (MPRTE) (i.e., the MTE evaluated at the margin) recovers the conditional expectation at the margin. Our approach does not estimate MTEs since we purposely abstract from imposing exclusion restrictions. However, the comparison with the MTE framework is useful for rationalizing the weighting schemes used by our approaches.

To see why, suppose the variance of ζ_i is almost zero. In this case, under A0, A1, and A2, the defendants that are more likely to be marginal given their observables are also the (unconditional) marginals. Then, computing averages using the mass of released defendants with the lowest propensity score would be sufficient for recovering the MPRTE. That would be problematic, however, because of (at least) two reasons. First, in practice, the variance of ζ_i is likely to be non-zero and, therefore, there is measurement error in the estimated ranking. Second, if the propensity score is continuous, there would not be a large mass of defendants with the lowest propensity score. Then, because of sampling error, it would be desirable to include additional observations to compute more precise estimations.

Then, in our setting, it makes sense to add additional observations for estimating the conditional expectations. Since increasing the sample size with inframarginal defendants may add bias to the estimation, we truncate the outcome equation sample to only consider the lower part of the (estimated) propensity score distribution. For simplicity and transparency, the simple approach equally weights each observation of this sub-sample. The non-parametric regression weights according to the estimated propensity score to give more importance to the observations that are closer to the margin in expectation. Then, these weighting schemes allow our approach to approximate the notion of MPRTE, which is the relevant structural estimand for the outcome test.

Inference The distributions of the two proposed estimators of prejudice must consider that the sample definition criterion is estimated. In addition, the can be noise in the estimated conditional expectations if group-specific sample sizes are small. We therefore suggest using bootstrap to

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

calculate confidence intervals.[11]

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 what follows we propose an implementation.

We focus on instances where the propensity score is estimated using a probit model. The test can be implemented as follows. 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 corresponds to the (standarized) ζ_i from the previous subsection.[12] Finally, for each of the K realizations, and given the estimated parameters of the probit model, simulate $Release_i^*$ for all released defendants, define samples of marginally released defendants, and estimate the group-specific pretrial misconduct rates for marginal defendants. With the estimated pretrial misconduct rates, the econometrician can assess the bias induced by the measurement error by examining the distribution of the P-BOT estimate across all simulations. We illustrate this test in our empirical application.

3.4 Discussion

We think our approach has three main good properties. First, since the strategy is based on predictions, the identification of marginal individuals is robust to standard omitted variable bias. Second, the P-BOT requires neither instruments nor the random assignment of judges, and allows for non-monotone discrimination patterns. Finally, its implementation is simple: testing for prejudice is reduced to projection models and linear regressions. Notwithstanding these good properties, we see two main limitations. First, our identification strategy relies on assumptions that may be restrictive in some settings. Second, the P-BOT's ability to deal with the inframarginality problem

[11] The bootstrap is not always valid in two-step estimations (Cattaneo and Jansson, 2018; Cattaneo, Jansson, and Ma, 2019). This could be problematic for our approach, especially considering the similarities between the P-BOT, RDD, and propensity score-based procedures (Abadie and Imbens, 2008; Calonico, Cattaneo, and Titiunik, 2014). Our approach, however, is a simple difference in means using a generated regressor that determines the sample of the second step. To the extent that the process that generates the regressor is continuous, the bootstrap is consistent for the P-BOT. This implies that this inference strategy is valid whenever the propensity score is continuous in X_i. When that is not the case, inference via bootstrap may be problematic. Yet, in that case, the implementation of the P-BOT is also compromised since the lack of continuity flattens the ranking of released defendants.

[12] 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. Formally, if $\zeta_i \sim \mathcal{N}(\mu_\zeta, \sigma_\zeta^2)$, we can write $Release_i = 1\left\{\frac{1}{\sigma_\zeta}\left(\frac{(d(X_i)-r_1(X_i))}{r_2(X_i)} - \mu_\zeta\right) \geq \tilde{\zeta}_i\right\}$, where $\tilde{\zeta}_i \sim \mathcal{N}(0,1)$. Then, the probit model estimates the left-hand-side and simulations of $\tilde{\zeta}_i$ can be used to perturb the estimated ranking.

depends on the availability of good predictors. As discussed throughout the section, we propose empirical diagnostics to assess the plausibility of our identification assumptions and the relevance of the potential bias due to measurement error. We illustrate these tests in our empirical application.

4 Empirical Application: Institutional Setting and Data

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

4.1 Setting

The current criminal justice system in Chile was implemented in 2005 and works uniformly throughout the territory. We focus on pretrial detentions. The procedure to define pretrial detention for arrested people 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. Following the legal principle of presumption of innocence, judges should not incarcerate defendants unless there is clear danger of escape (i.e., a high probability of failing to appear in court), the defendant represents a danger to society (i.e., a high probability of committing a different crime during the investigation), or imprisonment aids the investigation of the criminal case. In general, the arraignment hearing is very brief (lasting about 15 minutes) and is carried out by quasi-randomly assigned judges.

We test for prejudice against the largest ethnic minority group in Chile, the Mapuche. According to the last census, around 10% of the Chilean population reported themselves as being Mapuche. The Mapuche population is an interesting case of analysis for three reasons. First, a long-running conflict exists between the Mapuche and the Chilean state dating back more than a century (Cayul et al., 2018). In this context, it is frequently claimed that the Chilean institutions are biased against the Mapuche. Second, the Mapuche people are subject to numerous negative stereotypes, such as tendencies towards laziness, violence and alcoholism, from some quarters of Chilean society (Merino and Quilaqueo, 2003; Merino and Mellor, 2009). There is no evidence for any systematic difference in behavior between the Mapuche people and the rest of the population. Third, Mapuche people are identifiable, mainly because of their surnames but also to some extent due to their physical appearance. Thus, discrimination against members of this group is feasible.

4.2 Data

We use administrative records from the Public Defender's Office (PDO). The PDO is a centralized public service under the oversight of the Ministry of Justice. It offers criminal defense services to all individuals accused of or charged with a crime; as such, it ensures the right to a defense by a lawyer and due process in criminal trials. Our estimation sample covers more than 95% of the criminal cases for the period between 2008 and 2017, and contains detailed case and defendant characteristics. In addition, we can identify the judges and attorneys assigned to each case at the beginning of the criminal process (i.e., when the determination of pretrial detention occurs).

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

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

Descriptive statistics Table 1 presents the descriptive statistics of our estimation sample. Mapuche defendants represent 7.4% of the total sample when we consider our most comprehensive definition of Mapuche ($52,002/699,732$). Release occurs in about 84% of the cases, with a minor difference in favor of Mapuche defendants. In terms of the outcomes that pretrial detention seeks to avoid, conditional on being released, between 23% and 30% of the defendants (depending on the group) engage in at least one type of pretrial misconduct, either non-appearance in court or pretrial recidivism. Across all measures of pretrial misconduct, released Mapuche defendants demonstrate

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

Table 1: Descriptive Statistics

		Mapuche			
	Non-Mapuche	At least one surname	Two surnames	Self-Reported	Self-Reported or at least one surname
Released	0.84	0.85	0.87	0.85	0.85
Outcomes (only for released)					
Non-appearance in court	0.17	0.16	0.14	0.16	0.16
Pretrial recidivism	0.19	0.17	0.13	0.16	0.17
Pretrial misconduct	0.30	0.27	0.23	0.27	0.27
Individual Characteristics					
Male	0.88	0.89	0.91	0.92	0.89
At least one previous case	0.68	0.66	0.60	0.65	0.66
At least one previous pretrial misconduct	0.40	0.37	0.29	0.36	0.37
At least one previous conviction	0.65	0.63	0.57	0.62	0.63
No. of previous cases	4.59	4.25	3.47	4.13	4.28
Severity previous case	0.08	0.07	0.06	0.07	0.07
Severity current case	0.18	0.17	0.15	0.16	0.17
Court Characteristics					
Average severity (year/Court)	0.09	0.09	0.08	0.08	0.09
No. of cases (year/Court)	3,053	2,729	2,311	1,802	2,717
No. of judges (year/Court)	46	40	32	20	40
Observations (released)	541,743	42,987	8,455	7,992	43,952
Observations (non-released)	105,988	7,830	1,255	1,431	8,049

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

better conduct during prosecution than released non-Mapuche defendants. On average, the criminal records of Mapuche defendants are less severe, measured as both the number of previous cases and their severity. The current cases of Mapuche defendants are also slightly less severe.

5 Empirical Application: Results

This section presents the results of our empirical application. First, we assess the validity of the identification strategy. We then discuss the prediction model for the release status and 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. Then, we perform alternative tests for prejudice and compare the results. Finally, we develop extensions to the basic model to discuss the interpretation of the outcome test.

5.1 Identification strategy

First, we present evidence that suggests that our identification assumptions are plausible in this setting. Details on the tests' implementation are discussed in Appendix F.

A0 One way of assessing the plausibility of A0 (continuity and full support) is to look at the empirical distribution of the propensity score. Appendix F shows the propensity score distributions for Mapuche and non-Mapuche released defendants. The figures suggest that A0 is met in our setting, especially for the more comprehensive Mapuche definitions.

A1 Recall that A1 implies monotonicity in observables in the selection equation which, through the lens of the model, implies monotonicity in observables in the risk equation. Appendix F shows that the coefficients of the 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. For example, the marginal effect of having previous cases on the probability of being released is -0.028 for Mapuche defendants and -0.029 for non-Mapuche defendants. We include several observables in each regression and consider eight different criteria for splitting the sample. In 96% of the cases considered, the sign of the coefficient is consistent between subsamples. We interpret this as strong evidence in favor of A1.

A2 and ranking validity A2 is more difficult to test since a formal diagnostic requires stronger structural assumptions. Moreover, A2 is sufficient but not necessary. Accordingly, we propose a second diagnostic that assesses, in more general terms, the validity of the propensity score-based ranking. Noting that the relevant unobservables are variables observed by the judges, we can 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 these predictions, we can compute rank correlations between the (restricted) propensity scores among released defendants by groups of observables, and the conditional probabilities of being marginal that can be recovered from the unrestricted estimation. Appendix F shows, using different rankings, statistics, and excluded variables, that the rank correlations are very large in all cases. We interpret this as broad support for our identification argument.

5.2 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, the number of previous prosecutions, the severity of previous prosecutions, whether the individual engaged in pretrial misconduct during a previous prosecution, whether the individual has been convicted in the past, the severity of the current prosecution, the number of cases seen in the court during the year of the prosecution, the number of judges working at the court during the year of the prosecution, the assigned public attorney's quality and its square, the assigned judge's leniency and its square, and year of prosecution fixed effects. 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, we also compute the release probabilities using a linear probability model adding court fixed effects. We also use Lasso to select regressors considering all interactions and squared terms, and judge fixed effects. Finally, we also fit a heteroskedastic probit model. Since results are consistent between models, we restrict our discussion to the probit case. Results using alternative prediction models can be found in Appendices G and I.

Appendix G shows the results of the probit model. Considering 0.5 as the probability threshold, 85% or more of the cases are correctly classified by the prediction model (86% for Mapuche and 85% for non-Mapuche defendants). We also perform an out-of-sample cross-validation exercise that gives similar conclusions.[14] Finally, we apply the methods of inference for rankings set out in Mogstad et al. (2020) and conclude that more than 80% of the released defendants labeled as marginals have true propensity scores in the bottom 5% of the distribution, with 95% confidence.[15]

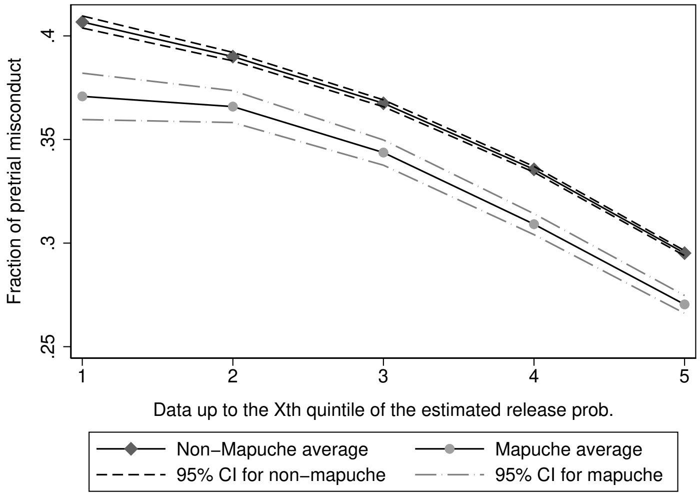
5.3 Outcome equation

To formally test for prejudice against Mapuche defendants, we use the predicted release probabilities to rank released defendants and build samples of marginal individuals. As a first exploratory analysis, we analyze how the outcome test varies as we increase the estimation sample. We achieve this by sequentially adding defendants with a higher predicted probability of being released. We

[14] 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.

[15] We would like to thank Daniel Wilhelm for answering questions about the code. In concrete terms, we calculate standard errors for the predictions based on the probit model and compute the joint (simultaneous) confidence sets for the ranks. Then, we count how many individuals labeled as marginals have ranking upper bounds within the bottom 5% (as in their τ-worst suggested procedure). For computational feasibility, we consider the 40,000 observations of released individuals with lower estimated propensity scores, use three decimals for the predicted probabilities and their standard errors, and derive critical values using 100 bootstrap repetitions. The specific shares for each definition of Mapuche are 82.1%, 81.9%, 81.6%, and 82.2%, respectively.

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



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

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 (i.e., 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 result of this exercise for the most comprehensive definition of Mapuche (self-reported or at least one surname). The plots for the other definitions are presented in Appendix H. The outcome is defined as any pretrial misconduct (i.e., non-appearance in court or pretrial recidivism). Three aspects 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. This result can be thought of as a test of model specification: defendants that are more likely to be released are also less likely to be engaged in pretrial misconduct. This suggests that judges care about expected outcomes when making pretrial detention decisions. Finally, the two lines are mostly parallel with a slightly wider gap in the first quintile. This suggests that in our setting the potential inframarginality bias exists but is modest.

Going beyond the graphical evidence, Table 2 presents the results of the implementation of the P-BOT. We focus on the most comprehensive definition of Mapuche. Results for the other defini-

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

	A: Simple Version		**B: Non-Parametric**	
Max. percentile considered:	5	10	5	10
Point estimate, (a)-(b):	-0.040	-0.040	-0.031	-0.036
C.I. (95%)	[-0.064, -0.020]	[-0.054, -0.024]	[-0.058, -0.007]	[-0.056, -0.017]
(a) Mapuche expectation	0.368	0.363	0.393	0.375
(b) Non-Mapuche expectation	0.408	0.403	0.425	0.411
No. of Mapuche	1,986	3,901	1,986	3,901
No. of Non-Mapuche	27,299	54,669	27,299	54,669

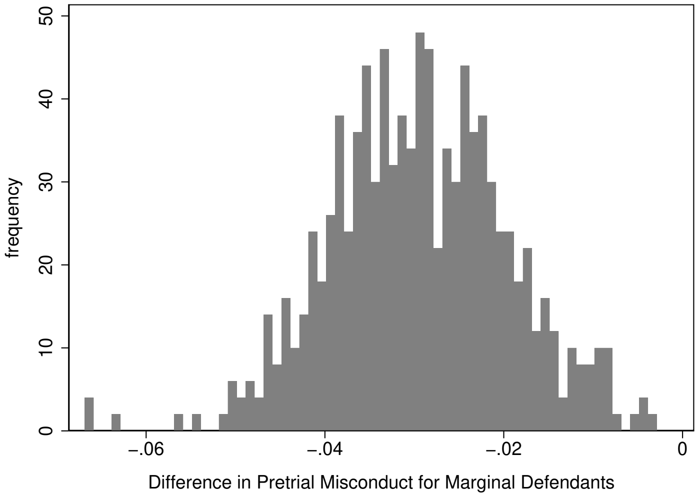
Note: This table presents the results from the P-BOT using the data described in Table 1, considering two approaches to estimate the outcome equation and two criteria to determine who is the margin. Mapuche is defined as self-reported or at least one surname. Release probabilities are predicted using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

tions are presented in Appendix H. 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 marginal defendants. In Panel B, we implement the non-parametric version, where the point estimate is obtained by subtracting the Mapuche and non-Mapuche conditional expectations for pretrial misconduct, which are non-parametrically calculated at the first percentile of the estimated release probability distribution. We consider two criteria to define the margin: the bottom 5% and bottom 10% of the predicted release probability distribution of released defendants. A negative point estimate constitutes evidence of prejudice against Mapuche defendants.

Table 2 shows that point estimates are negative and statistically significant. Marginally released Mapuche defendants are between 3 and 4 percentage points less likely to be engaged in pretrial misconduct relative to marginal non-Mapuche defendants. This provides evidence of prejudice against Mapuche defendants. Results are robust to considering non-appearance in court and pretrial recidivism as separate outcomes (see Appendix I). Prejudice is more than three times larger when we identify Mapuche defendants using both surnames (see Appendix H), which we conjecture is explained by the salience of the ethnicity measure. Finally, consistent with the modest potential for inframarginality bias, results are similar between the different criteria for defining the margin.

Perturbation test Depending on the fit of the propensity score, the noise in the ranking estimation may induce bias in the outcome test. To assess the extent of this concern, we perform the

Figure 4: Perturbation Test



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

perturbation test proposed in Section 3. We implement the test using the coefficients of the probit model. For each individual in our sample of released defendants, we simulate 500 realizations from a standardized normal distribution to simulate $Release_i^*$, recompute the ranking, and redefine the sample of marginals. Then, in each of the 500 simulations, we estimate the outcome test using the simple approach. Finally, we plot the distribution of the outcome test across simulations.

Figure 4 shows the results for the most comprehensive Mapuche definition. Reassuringly, the perturbation test suggests that our results are robust to this potential bias. With the exception of the self-reported measure (our least preferred Mapuche indicator, see Appendix H), the distributions of the outcome test do not include the zero. That is, even in the worst-case scenario induced by this test, the conclusion of prejudice is not reversed. This is consistent with the good fit of the propensity score estimation.

5.4 Alternative tests

To assess the relative performance between the P-BOT and other approaches, we also test for prejudice using alternative methods. We consider the outcome test using the full sample (Knowles, Persico, and Todd, 2001) and the instrument-based approach (Arnold, Dobbie, and Yang, 2018). For the latter, we exploit the quasi-random assignment of judges to pretrial detention hearings that

Table 3: Alternative Tests for Prejudice

	Outcome test (full sample)	**IV-Outcome test (Mapuche)**	**IV-Outcome test (non-Mapuche)**
Coeff.	-0.023	0.240	0.363
Robust SE	(0.003)	(0.478)	(0.059)
Observations	699,732	50,802	647,701

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

characterizes the Chilean setting.[16]

Table 3 presents the results for the alternative methods using our preferred Mapuche definition. The outcome test using the full sample, as expected, provides evidence of prejudice. Consistent with Figure 3, however, we note that the inframarginality bias is biasing the estimation downwards. The most interesting analysis relates to the application of the instrument-based approach. While the estimated LATE for the non-Mapuche defendants is precisely estimated, the Mapuche estimation is severely underpowered. Point estimates (the difference between both estimated LATEs) support the existence of prejudice, 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 (see Appendix H). In Appendix J we report the first-stage F-tests, which corroborates the lack of power of the instrument in the minority sample. Therefore, our setting is one in which the instrument-based approach is not well-behaved because of power problems.[17]

Moreover, recall from Table 2 and Appendix I that the P-BOT's estimate of the pretrial misconduct rate of marginally released non-Mapuche defendants is between 37.6% and 42.5%. The estimated LATE using the instrument-based test in the non-Mapuche sample is 36.3%. Therefore, the estimation of the pretrial misconduct behavior of non-Mapuche marginal defendants is

[16] Appendix J presents the results of the randomization test suggested by Arnold, Dobbie, and Yang (2018).

[17] We also perform the test proposed by Frandsen, Lefgren, and Leslie (2019) and reject the null hypothesis of monotonicity. While their procedure jointly tests for exclusion and monotonicity, the institutional setting of our application suggests exclusion holds and, therefore, we interpret rejections of the null as deviations from strict monotonicity. We would like to thank Emily Leslie for answering questions about the code. We parametrize the test following the recommendations of the authors. For computational feasibility, we compute the test for random subsamples. In concrete terms, we generate random samples considering (i) 25% of court-by-year cells, and (ii) 25% of bail judges. For each criterion, we build 10 random subsamples. In all subsamples, the composite p-value is 0.000. We only consider the subsample of non-Mapuche defendants. Since courts, years, and judges vary in their caseloads, random samples have different sizes. Among the 20 samples used, the average sample contains 159,069 observations. The smaller (larger) sample contains 142,675 (176,704) observations.

similar between both methods. In addition, in Appendix K we perform a complier analysis and show that the non-Mapuche defendants identified as marginals by both methods have comparable distributions of observables. This suggest that both methods yield similar results in cases where they are expected to work properly. Then, although the instrument-based method does not report reliable estimations of discrimination in our setting, its application is reassuring for the efficacy of the P-BOT and reinforces the complementarity argument developed throughout the paper.[18]

5.5 Extensions

Recall the discussion in Section 2 that argues that the interpretation of differences in average effective thresholds may depend on some structural features of the selection process. While we believe that under alternative interpretations our notion of prejudice remains relevant from a normative perspective, it may be of interest to disentangle between sources. In the reminder of the section we revisit this discussion and illustrate how the P-BOT can be used to explore these distinctions.

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

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

The table shows that prejudice patterns are more complex than the binary model case. This becomes clear when looking at the differences in the four conditional means. In Panel A, results show that prejudice against Mapuche defendants is mainly relevant for those Mapuche who live in low-

[18] Deviations from strict monotonicity suggest that the estimated conditional expectation at the margin using instrumental variables is potentially biased. However, since treatment effects can still be identified under weaker notions of monotonicity (e.g., Frandsen, Lefgren, and Leslie, 2019), to the extent that those weaker assumptions hold in our data, the bias in the estimated behavior at the margin should be limited. Then, small differences between both methods are still reassuring for the P-BOT's assessment.

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	-0.015	Mapuche	-0.031
C.I. (95%)	[-0.038, 0.014]	C.I. (95%)	[-0.048, -0.015]
Low income	0.017	Mapuche region	-0.069
C.I. (95%)	[0.009, 0.026]	C.I. (95%)	[-0.092, -0.045]
Mapuche and low income	-0.037	Mapuche and mapuche region	-0.019
C.I. (95%)	[-0.076, -0.006]	C.I. (95%)	[-0.068, 0.034]
Pretrial misconduct expectation for:		**Pretrial misconduct expectation for:**	
Mapuche and low income	0.327	Mapuche and mapuche region	0.285
Non-Mapuche and low income	0.378	Non-Mapuche and mapuche region	0.336
Mapuche and high income	0.347	Mapuche and non-mapuche region	0.374
Non-Mapuche and high income	0.361	Non-Mapuche and non-mapuche region	0.405
Observations:		**Observations:**	
Mapuche and low income	1,765	Mapuche and mapuche region	466
Non-Mapuche and low income	22,515	Non-Mapuche and mapuche region	1,495
Mapuche and high income	1,382	Mapuche and non-mapuche region	3,435
Non-Mapuche and high income	21,036	Non-Mapuche and non-mapuche region	53,174

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

income municipalities. This suggests that the relevant prejudice is against low-income Mapuche defendants. In Panel B, results suggest that Mapuche defendants are slightly more prejudiced against in the conflict region, however, the interaction is non-significant. These results suggest that non-monotone patterns of discrimination are likely to occur in practice.

Assignment rule for judges The assignment rule matters for interpreting whether the aggregate estimated prejudice is driven by judges being, on average, prejudiced, or by Mapuche defendants visiting courts that are, on average, less lenient. When information on judges is available, the relevance of these two sources of prejudice can be tested. When judges are randomly assigned at the court-by-time level, implementing our simple P-BOT regression while controlling for court-by-year fixed effects will yield an estimate for prejudice net of the role of the assignment rule. This is what we present in Table 5. Point estimates are about a half of the baseline results. This suggests that prejudice driven by the assignment rule is an important force behind our results.

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

Max. percentile considered:	5	10
Point estimate:	-0.020	-0.020
C.I. (95%)	[-0.046, 0.003]	[-0.035, -0.003]
No. of Mapuche	1,986	3,901
No. of Non-Mapuche	27,299	54,669

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

6 Conclusion

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

In this paper, we propose a novel observational 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 and extensively discuss our notion of prejudice. 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 propose a series of empirical diagnostics for assessing their validity. We argue that the P-BOT is an attractive methodology in the absence of well-behaved instruments.

Our identification strategy significantly 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, therefore, 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 is not a particularly strong requirement.

We use the P-BOT to test for prejudice in pretrial detentions against the Mapuche, the largest ethnic minority in Chile, 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. We also 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. We also show that discrimination patterns are likely to be more complex than commonly assumed, and that the assignment rule of judges to defendants partly explains the overall estimated effect.

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, Gary Becker's original ideas 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 (and ex-post measurable) outcome. The fact that the P-BOT does not require instruments for its implementation may foster the application of the outcome test in a broader range of settings where testing for prejudice is important.

References

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

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

Antonovics, K. and B. G. Knight (2009). A new look at racial profiling: Evidence from the Boston Police Department. *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. *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.

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.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2020). Inaccurate statistical discrimination. *Working Paper*.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Canay, I. A., M. Mogstad, and J. Mountjoy (2020). On the use of outcome tests for detecting bias in decision making. *Working Paper*.

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

Cattaneo, M. D., M. Jansson, and X. Ma (2019). Two-step estimation and inference with possibly many included covariates. *Review of Economic Studies 86*(3), 1095–1122.

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). *Working Paper*.

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

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.

Durlauf, S. N. and J. Heckman (2020). An empirical analysis of racial differences in police use of force: A comment. *Journal of Political Economy 108*(10), 3998–4002.

Feigenberg, B. and C. Miller (2020). Racial disparities in motor vehicle searches cannot be justified by efficiency. *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.

Gelbach, J. (2021). Testing economic models of discrimination in criminal justice. *Working Paper*.

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.

Hull, P. (2021). What marginal outcome tests can tell us about racially biased decision-making. *Working Paper*.

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

Kleinberg, J., J. Ludwig, S. Mullainathan, and C. R. Sunstein (2019). Discrimination in the age of algorithms. *Journal of Legal Analysis 10*.

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

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

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

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. (2020). An absolute test of racial prejudice. *Journal of Law, Economics, and Organization*.

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.

Mogstad, M., J. P. Romano, A. Shaikh, and D. Wilhelm (2020). Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries. *Working Paper*.

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

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

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

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

Zhou, X. and Y. Xie (2019). Marginal treatment effects from a propensity score perspective. *Journal of Political Economy 127*(6), 3070–3084.