Evaluating Counterfactual Policies Using Instruments*

Michal Kolesár José Luis Montiel Olea Jonathan Roth

December 31, 2025

Abstract

We study settings in which a researcher has an instrumental variable (IV) and seeks to evaluate the effects of a counterfactual policy that alters treatment assignment, such as a directive encouraging randomly assigned judges to release more defendants. We develop a general and computationally tractable framework for computing sharp bounds on the effects of such policies. Our approach does not require the often tenuous IV monotonicity assumption. Moreover, for an important class of policy exercises, we show that IV monotonicity—while crucial for a causal interpretation of two-stage least squares—does not tighten the bounds on the counterfactual policy impact. We analyze the identifying power of alternative restrictions, including the policy invariance assumption used in the marginal treatment effect literature, and develop a relaxation of this assumption. We illustrate our framework using applications to quasi-random assignment of bail judges in New York City and prosecutors in Massachusetts.

*All errors are our own. We thank numerous seminar participants, Isaiah Andrews, Josh Angrist, Yuehao Bai, Guido Imbens, Pat Kline, Jesse Shapiro, Jesper Sørensen, Alex Torgovitsky, and Ed Vytlacil for helpful comments. We thank Peter Hull for providing us with aggregate statistics on New York City bail judges, and Henrik Sigstad for providing data on panels of judges in São Paulo.

1 Introduction

The most common approach for leveraging an instrumental variable (IV) to tease out causal effects from observational data is to use two-stage least squares (TSLS) and interpret the resulting estimand as a local average treatment effect (LATE)—a (weighted) average of treatment effects for individuals whose treatment status changes in response to the IV (Imbens & Angrist, 1994). This interpretation relies on the well-known IV monotonicity condition.

Yet in many applications of IV designs, both statistical evidence and institutional details point to failure of IV monotonicity. For example, consider the popular leniency IV design, which harnesses as-good-as-random assignment of judges or other decision-makers who differ in their leniency—the propensity to grant treatment—to study the effects of treatments such as incarceration, pretrial detention, or bankruptcy on various outcomes.[1] IV monotonicity requires that every defendant released by a given judge would also be released by all judges who are more lenient on average. Effectively, the judges need to agree on the ranking of the defendants in terms of their risk; they only disagree on the release cutoff. This is a stringent requirement, as pointed out in the original Imbens and Angrist (1994) analysis (Example 2, p. 472). Institutional knowledge suggests that decision-makers may have different rankings owing to heterogeneity in skills (Chan et al., 2022) or preferences (Mueller-Smith, 2015). Moreover, IV monotonicity is frequently rejected by statistical tests (e.g. Frandsen et al., 2023), as well as by direct evidence from judicial panels (Sigstad, 2026). To address these concerns, a growing literature has proposed weaker or alternative conditions that still allow for a LATE-type interpretation of the TSLS estimand (e.g. de Chaisemartin, 2017; Frandsen et al., 2023; Mogstad et al., 2021; Small et al., 2017). Yet, the plausibility of these alternative assumptions is debated (Mogstad & Torgovitsky, 2024).

Often, however, researchers may not ultimately be interested in LATE *per se*, but rather in evaluating counterfactual policies.[2] Researchers conducting a leniency IV, for instance, may be interested in policies that nudge the decision-makers to be more (or less) lenient, by, say, imposing release quotas (Arnold et al., 2022), providing algorithmic recommendations (Angelova et al., 2025), or imposing a presumption of non-prosecution for low-level offenses (Agan et al., 2023). In the limit, universal release programs may treat everyone with certain observable characteristics (Albright, 2022).

In this paper, we develop a general framework for directly evaluating counterfactual

[1] The leniency IV design was pioneered by Kling (2006). Apart from leveraging the random assignment of judges (e.g. Aizer & Doyle, 2015; Dobbie & Song, 2015; Dobbie et al., 2018), researchers have also used leniency IV to study a variety of other treatments, from patent-granting (e.g. Sampat & Williams, 2019) to foster care placement (e.g. Doyle, 2007). For example, Table 1 in Frandsen et al. (2023) gives a selective list.

[2] A recent survey of the empirical literature by Locher et al. (2025) concludes that researchers are rarely interested in the LATE *per se*.

policies that alter treatment assignment. We make the goal of policy evaluation explicit by indexing potential treatments as $D(z, a)$, where z is the value of the instrument (say, the identity of the judge in the leniency IV example), and a is an indicator for whether the policy (say, a release quota) is implemented. We only observe data under the status quo ($a = 0$), with the observed treatment given by $D(Z, 0)$, and the observed outcome given by $Y(D(Z, 0))$, where $Y(d)$ is the potential outcome under treatment d. The parameter of interest is $\theta = E[Y(D(Z, 1))]$, the average outcome (say, recidivism) under the counterfactual treatment assignment, $D(Z, 1)$. Since θ is generally not point-identified, our goal is to derive bounds for it, i.e., an identified set of values consistent with the observable data.

Within this framework, we develop three sets of results. First, we derive tractable bounds for θ without imposing IV monotonicity. Second, we show that in a range of relevant cases, imposing IV monotonicity does not help tighten the identified set. Thus, while some form of monotonicity is essential for a LATE interpretation, our results show that IV monotonicity is neither necessary nor sufficient to learn about counterfactuals. Third, we consider other assumptions that do help tighten the bounds, and illustrate their power in two applications.

Specifically, our first main result provides a computationally tractable characterization of the identified set for θ without imposing IV monotonicity. In related IV settings (e.g. Bai, Huang, Moon, Shaikh, & Vytlacil, 2025; Kitagawa, 2021), the identified set is commonly computed by enumerating all possible "response types", defined by the values of $D(\cdot, \cdot)$. This proves computationally infeasible in our setting, because the number of response types grows exponentially in the number of judges K if one does not impose IV monotonicity. We show that one can bypass response type enumeration and compute the identified set as the solution to a linear program that scales linearly in K. The key observation is that it suffices to consider sets of marginals for $(Y(\cdot), D(z, \cdot))$, involving the decision of one judge z at a time. This is because the observable data does not depend on the joint distribution of $D(z, a)$ across judges z, and without imposing IV monotonicity, we are also not imposing any *ex ante* restrictions on this joint distribution. Our result builds on insights in Richardson and Robins (2014), who derived bounds on the average treatment effect in IV settings with binary outcomes involving $O(K^2)$ restrictions.

Our second set of results gives sufficient conditions under which the identified set for θ does not depend on whether one imposes IV monotonicity. Specifically, we show that this is the case when either (a) one of the potential outcomes is known, or (b) the policy counterfactual is a "sufficiently strong" encouragement (where the notion of "sufficiently strong" is formalized in Proposition 3 below). Condition (a) is satisfied in many criminal justice settings, where, for example, one cannot commit pretrial misconduct unless one is released. Condition (b) is trivially satisfied for universal release programs, like that studied

in Albright (2022), and may also plausibly be satisfied by other counterfactual policies that strongly encourage judges to release more defendants. An important practical takeaway is that when these sufficient conditions are satisfied, the debate over IV monotonicity is somewhat of a red herring: instead of worrying about IV monotonicity, researchers interested in policy counterfactuals should turn attention to evaluating the validity of other assumptions that may in fact tighten the identified set.

Our third set of results considers several such assumptions. We begin by evaluating the policy invariance assumption of Heckman and Vytlacil (2005). Intuitively, IV monotonicity effectively imposes that judges agree on their rankings of defendants *under the status quo*; policy invariance additionally imposes that this ranking is the same *under the counterfactual*. We show that policy invariance can indeed be helpful in tightening the identified set in some settings where IV monotonicity is not. Of course, since policy invariance is stronger than IV monotonicity, it may often be implausible in practice. We therefore develop a relaxation of policy invariance that may be more plausible yet still informative. In particular, while policy invariance imposes perfect agreement among judges in the ranking of defendants, our relaxation only imposes that judges do not disagree too often. We show how bounds on disagreement rates can be calibrated using Sigstad (2026)'s data on panels of judges. We further show that these disagreement bounds can be tractably incorporated into our linear program for calculating the identified set. We also discuss a variety of other economically-motivated restrictions that can be incorporated to further tighten the identified set. For example, in the pretrial release setting, judges are legally instructed to release defendants unless they are at high risk of committing pretrial misconduct. A natural assumption is then that the defendants released by judges under the status quo are lower-risk than the defendants they would marginally release in response to a policy encouragement.

We illustrate the usefulness of our results with two applications. The first studies bail judges in New York City, who determine whether to release defendants awaiting trial, using data from Arnold et al. (2022). We evaluate two counterfactual policies in this context. First, we consider a policy that releases all defendants ($D(Z, 1) = 1$), allowing us to evaluate what would happen if the universal release policy in Kentucky studied by Albright (2022) were applied in this context. This policy is also relevant for evaluating the "disparate impact" of pretrial release decisions by race, as in Arnold et al. (2022), whose disparate impact parameter depends only on the (race-specific) average outcomes under universal release and observable probabilities. We obtain informative bounds for the universal release counterfactual, imposing only IV validity (i.e., random IV assignment and exclusion). Our bounds suggest, for example, that the defendants marginally released by this policy will have higher misconduct rates than those released under the status quo, which is intuitive given

the judges' instructions to only detain high-risk defendants. Our bounds for the disparate impact parameter are only moderately wider than the range of estimates obtained by different parametric assumptions in Arnold et al. (2022), and our confidence interval excludes zero, indicating significant disparate impact by race.

We also use this data to evaluate a quota policy that requires the bottom 90% of judges in terms of leniency to increase their release rates to match the top 10%. For this policy, only imposing IV validity yields trivial bounds that allow marginally-released defendants to have misconduct rates anywhere between 0 and 1. Intuitively, this occurs because without any restrictions on agreement rates between judges, the judges in the bottom 90% of leniency may choose to marginally release only IV "never-takers" who were not released by any judge under the status quo, and the data thus contain no information about their misconduct potential. To tighten the bounds, we rule out such perfect discordance between judges above and below the 90th percentile of leniency by calibrating a bound on the average disagreement rate between judges above and below the 90th percentile using Sigstad (2026)'s data on panels of judges who rule on the same cases. Under the calibrated disagreement bounds, we once again obtain informative bounds on the counterfactual.

The second application uses data from Agan et al. (2023, ADH23), who study the effect of non-prosecution of misdemeanor cases by assistant district attorneys (ADAs) in Massachusetts on subsequent criminal justice involvement of the defendants. ADH23 report TSLS estimates, although the Frandsen et al. (2023) test formally rejects IV monotonicity. ADH23 also write that "if all arraigning ADAs acted like the most lenient ADAs in our sample... [we] would likely see a reduction in criminal justice involvement". Based on this discussion, we evaluate two counterfactual policies to make ADAs act more similarly to the most lenient ones (defined as the top 10%). First, we consider a simple re-allocation of cases to the most lenient ADAs. The impact of such a policy is point-identified, and our results suggest it would decrease crime, in line with the discussion in ADH23. As a more realistic policy, we consider a quota that requires the bottom 90% of ADAs to match the non-prosecution rate of the top 10%. For this policy, stringent bounds on agreement rates between ADAs are needed to conclude that the policy would decrease crime; the bounds appear unreasonably high given rejection of IV monotonicity under the status quo. More plausible restrictions on disagreement are attained, however, if we impose that treatment effects are not too heterogeneous among people on whom the ADAs disagree. Thus, the conclusion that increasing non-prosecution likely leads to a reduction in crime, as argued in ADH23, requires either a specific policy implementation (such as reallocation) or restrictions on the degree of treatment effect heterogeneity.

Our paper is motivated in part by the observation in previous work that the connection

between LATE and economic policies of interest is not entirely clear (e.g. Heckman & Vytlacil, 2005, 2007). Heckman and Vytlacil (1999, 2005) introduced the marginal treatment effects (MTE) framework for analyzing the impacts of policies that change treatment assignment. However, the canonical MTE framework requires policy invariance. Our framework allows for evaluation of a larger class of counterfactual policies by introducing the generalized potential treatment function $D(z, a)$ and allowing for a wide range of restrictions on it, including relaxations of policy invariance.[3] Our paper also relates to a large literature that has considered bounds on parameters other than LATE in IV settings (e.g. Bai, Huang, Moon, Santos, et al., 2025; Balke & Pearl, 1997; Chen & Flores, 2015; Manski, 1990; Swanson et al., 2018). These papers have primarily focused on parameters such as ATE, or average effects for principal strata, in contrast to our focus on specific counterfactual policies.

Finally, several papers have provided conditions under which IV monotonicity does not help to tighten the identified set for the average treatment effect or the marginal distributions of potential outcomes (Bai, Huang, Moon, Shaikh, & Vytlacil, 2025; Balke & Pearl, 1997; Kitagawa, 2021). By contrast, Kamat (2019) shows that IV monotonicity can matter for the copula of the potential outcomes. We complement this literature by focusing on partial identification of counterfactual policy effects, which need not correspond to the average treatment effect, and showing that IV monotonicity does not matter for a broader class of policies.

The next section sets up the model and defines the identified set for θ, the average outcome under the counterfactual. Section 3 derives a tractable characterization of the identified set without imposing IV monotonicity. Section 4 gives sufficient conditions under which imposing IV monotonicity does not help tighten the identified set, while Section 5 gives examples of restrictions that do. The identifying power of these restrictions is illustrated with two applications in Section 6. The appendices collect proofs and additional results.

2 Setup

2.1 Observable data and potential outcomes

We observe data on the triple (Y, D, Z), where Y is an outcome of interest, $D \in \{0, 1\}$ is a binary treatment indicator, and $Z \in \mathcal{Z} = \{1, \ldots, K\}$ is a multivalued instrument that

[3] In a complementary line of work, Ura and Zhang (2025) consider a multi-index generalization of the MTE framework that allows for relaxations of IV monotonicity. Although the framework allows for computation of bounds for a variety of treatment parameters (such as ATE or ATT), analyzing impacts of general counterfactual policies appears challenging without further assumptions. Likewise, Sigstad (2024) studies conditions under which standard MTE methods correctly estimate parameters such as ATE or LATE even when IV monotonicity is violated.

influences the treatment. We wish to use the data to learn about the effect of a *counterfactual policy*, which we denote by $a = 1$. We let $a = 0$ denote the status quo observed in the data. To properly define the effect of this policy, we use the potential outcome notation, with $Y(d)$ denoting the potential outcome under treatment status $d \in \{0, 1\}$, and $D(z, a)$ denoting the potential treatment under instrument value $z \in \mathcal{Z}$ and policy $a \in \{0, 1\}$. We assume that the support of the joint distribution of $(Y(0), Y(1))$ is contained in a compact set $\mathcal{Y}^2 \subseteq \mathbb{R}^2$.[4] The observed treatment and outcome are thus given by $D = D(Z, 0)$ and $Y = Y(D) = Y(D(Z, 0))$, while the potential treatments $D(z, 1)$ are not observed. The policy effect relative to the status quo is given by $\theta - E[Y]$, where $\theta := E[Y(D(Z, 1))]$ is the *average counterfactual outcome* under the new policy.[5] Since identification of $E[Y]$ is trivial, we focus on θ as the parameter of interest. We further note that in many settings, it may be reasonable to impose that the counterfactual policy has a monotone effect on the treatment in the sense that $D(z, 1) \geq D(z, 0)$ (which we formalize in Assumption 2 below), in which case:

$$E[Y(1) - Y(0) \mid D(Z, 1) > D(Z, 0)] = \frac{\theta - E[Y]}{E[D(Z, 1)] - E[D]}. \tag{1}$$

The left-hand side is the treatment effect for the *policy compliers* who are induced to adopt the treatment by the counterfactual policy (i.e., the marginally-released defendants). Thus, when the "first stage" impact of the counterfactual policy is identified, the average effect for policy compliers is also a simple point-identified transformation of θ. We follow the usual convention in identification analysis of suppressing any observable covariates, so that our setup and results may be understood as being conditional on covariates.

To fix ideas, as a running example, consider a judge IV design, where D is an indicator for release of a defendant, Z is the identification number of a judge who is randomly assigned to the case, and the policy $a = 1$ is an encouragement or a quota to release more defendants. Y may correspond to various outcomes of interest. In many applications of leniency IV designs, the outcome is binary, such as an indicator for pretrial misconduct (Arnold et al., 2022), recidivism or criminal complaints (Agan et al., 2023), conviction or employment (Dobbie et al., 2018), earning above the poverty rate (Kling, 2006), or an indicator or high-school graduation rate or entering adult prison when the sample consists of juvenile defendants

[4] For simplicity, we focus on the case where Y is bounded, which ensures that the bounds on the average policy effect are finite. Our results readily extend to the case with unbounded outcomes if one imposes additional restrictions on the outcome, such as those considered in Section 5.2, that guarantee that the average policy effect is finite. Even if the bounds are infinite, one could modify our framework to obtain bounds on quantiles, rather than the average policy effect.

[5] For ease of exposition, we assume that the marginal distribution of Z is invariant to the policy—e.g., imposing a quota on release rates does not affect the assignment of judges. It is straightforward to accommodate known changes in the marginal distribution of Z by weighting the average counterfactual outcome by the density of the counterfactual Z distribution relative to is status quo distribution.

(Aizer & Doyle, 2015). However, most of our results also allow for continuous outcomes, such as future academic performance of the defendant's child (Norris et al., 2021) or average earnings (Kling, 2006).

Since the potential outcomes are indexed by the treatment only, our notation implicitly embeds two exclusion restrictions: both the instrument (z) and the counterfactual policy (a) affect the outcome only by impacting the treatment. In the judge IV running example, the first exclusion restriction amounts to assuming that the judges affect the defendants' outcomes only through their release decisions. This would be violated if judges can make other decisions, such as set the terms of probation or fees owed, that may directly affect outcomes (Mueller-Smith, 2015). The exclusion restriction with respect to the counterfactual policy likewise imposes that, say, a quota on how many defendants are released affects defendants' outcomes only by changing which defendants are released. This could be violated if releasing more defendants reduces deterrence, thereby increasing crime directly; more generally, the exclusion restriction rules out general equilibrium effects of the policy.

Additionally, we assume throughout that the instrument is as-good-as-randomly assigned. In the judge IV running example, this holds if judges are as-good-as-randomly assigned to defendants, which is the case in many of the empirical papers cited above, at least once we condition on court-by-time fixed effects. Together with the exclusion restrictions described above, random assignment implies that the instrument is valid in that it is independent of the potential treatments and potential outcomes,

$$\{(Y(d), D(z,a))\}_{d,a\in\{0,1\},z\in\mathcal{Z}} \perp\!\!\!\perp Z. \tag{2}$$

The identification power of the instrument comes from its influence on the treatment.

2.2 The identified set for the average counterfactual outcome

To define the identified set for θ, let P denote the distribution of the observable data $(Y, D, Z) = (Y(D(Z,0)), D(Z,0), Z)$, and let P^* denote the distribution of the model primitives, that is, the joint distribution of potential outcomes, potential treatments, and the instrument, $(Y(\cdot), D(\cdot,\cdot), Z)$. We say that P^* *generates* P if the implied distribution of (Y, D, Z) under P^* matches the observed data—i.e., if under P^*, the distribution of the triple $(Y(D(Z,0)), D(Z,0), Z)$ is P.

We encode any *ex ante* restrictions on the model primitives by restricting the family $\mathcal{P}^*$ of possible distributions for P^*. At minimum, since we maintain IV validity throughout, we must have that $\mathcal{P}^* \subseteq \mathcal{P}^*_{\text{valid}}$, where $\mathcal{P}^*_{\text{valid}}$ is the set of distributions satisfying the IV validity condition (2). However, the family $\mathcal{P}^*$ may be smaller if the researcher imposes

other restrictions on the model that we discuss in more detail below, such as IV monotonicity. Given $\mathcal{P}^*$, the identified set for the model primitives is given by the set of distributions in $\mathcal{P}^*$ that could have generated the observed data:

$$\mathcal{P}_I^*(P; \mathcal{P}^*) := \{P^* \in \mathcal{P}^* \colon P^* \text{ generates } P\}.$$

The identified set for the average counterfactual outcome, $\theta = E[Y(D(Z,1))]$, corresponds to the set of counterfactual means consistent with these primitives:

$$\Theta_I(P; \mathcal{P}^*) := \{E_{P^*}[Y(D(Z,1))] \ : \ P^* \in \mathcal{P}_I^*(P; \mathcal{P}^*)\},$$

where we make explicit that the expectation is taken with respect to the distribution P^*. We suppress this dependence whenever it doesn't cause confusion.

For our analysis, it will be useful to distinguish between assumptions that restrict the joint distribution of decisions across decision-makers—i.e., restrict the dependence between $D(z,\cdot)$ and $D(z',\cdot)$ for $z \neq z'$—versus assumptions that only restrict the set of marginals $\{(Y(\cdot), D(z,0), D(z,1))\}_z$ involving one z at a time. We refer to the former as *cross-judge* restrictions, and the latter as *marginal* restrictions, as formalized by the next assumption.

Assumption 1 (Marginal restrictions)**.** The marginal distributions $\{(Y(0), Y(1), D(z,0), D(z,1))\}_{z \in \mathcal{Z}}$ are contained in some collection of distributions $\mathcal{R}^*$, specified by the researcher.

Next, we list examples of both marginal and cross-judge restrictions on the potential treatments that we will consider in our analysis. Sections 3, 4 and 5 then explore the identifying power of these restrictions (and variants thereof).

2.3 Examples of restrictions on potential treatments

Knowledge of policy implementation details will typically allow us to impose natural restrictions on the marginal distribution of $(D(z,1), D(z,0))$ for each z, which can be encoded by an appropriate choice of $\mathcal{R}^*$ in Assumption 1. In the context of our running example, for many counterfactual policies, the marginal release rates will either be known or consistently estimable, allowing us to impose that $E[D(z,1)] = \alpha_{1,z}$ for all z. For instance, under a quota policy that requires judges to release at least a fraction q of the defendants, we may set $\alpha_{1,z} = \max\{E[D \mid Z = z], q\}$ for each z, so that each judge who is currently below the quota increases their release rate to match it.

For quota policies or other policies that take the form of an encouragement, it is natural to also impose the following marginal restriction:

Assumption 2 (Policy monotonicity)**.** $D(z,1) \geq D(z,0)$ for all z.

If a policy is a directive to treat everyone, such as a universal release program, then policy monotonicity holds trivially since $D(z,1) = 1$ for all z. In the context of a quota policy, Assumption 2 simply imposes that any defendant released without the quota in place would also be released when the quota is in place, which seems quite reasonable. Likewise, an algorithm that flags low-risk defendants as candidates for release would be expected to have a monotonic effect on release rates.[6] We expect that for many counterfactual policies of interest policy monotonicity will be reasonable, and we will therefore maintain this assumption for many of our results.

A different type of monotonicity restriction that is commonly imposed is the Imbens and Angrist (1994) IV monotonicity assumption:

Assumption 3 (IV monotonicity)**.** For any pair z, z', either $D(z,0) \geq D(z',0)$ (a.s.) or $D(z',0) \geq D(z,0)$ (a.s.).

This assumption allows one to interpret the TSLS estimand as a local average treatment effect (LATE), a weighted average of treatment effects for individuals whose treatment depends on the instrument under the status quo, i.e., those for whom $D(z,0)$ varies with z. In contrast to policy monotonicity, IV monotonicity involves restrictions on the joint behavior of judges, so it is a cross-judge rather than a marginal restriction. Vytlacil (2002) shows that Assumption 3 is equivalent to the existence of a latent index U_0 (not depending on z) and thresholds α_z such that

$$D(z,0) = \mathrm{I}\{U_0 \leq \alpha_z\} \quad \text{for all } z.$$

Intuitively, this representation says that all judges have the same ranking of defendants under the status quo (U_0), but potentially disagree on the threshold to use to determine whether a defendant is released (α_z). As described in the introduction, Assumption 3 may often be questionable in empirical contexts—judges may differ in their rankings of defendants owing to idiosyncratic perceptions of risk, different preferences over crime types, or differences in skill—and in fact is often rejected using statistical tests (e.g. Frandsen et al., 2023), as well as direct measurement when we observe multiple judges making a decision about the same case (Sigstad, 2026).

[6] In some contexts, the algorithm may recommend release for some defendants, and recommend detention for others. In this case, we'd expect the algorithm to weakly increase release rates among those recommended to be released, and weakly decrease them among those recommended for detention. If the algorithmic recommendation is a function of observables, one could impose the appropriately signed version of monotonicity in each subpopulation.

In addition to IV monotonicity, which imposes a common ranking U_0 under the status quo, we will also consider the even stronger cross-judge assumption that there exists a common ranking U that is invariant to the policy a. This assumption underlies the use of marginal treatment effects (MTE) framework (Heckman & Vytlacil, 1999, 2005) for evaluating counterfactual policies.

Assumption 4 (Policy invariance)**.** $D(z,a) = \mathrm{I}\{\alpha_z + \beta_z \cdot a \geq U\}$, with $\beta_z \geq 0$.

Intuitively, Assumption 4 imposes that judges have a common ranking U under both the status quo and counterfactual but differ in their thresholds. Moreover, the restriction that $\beta_z \geq 0$ implies that implementing the counterfactual policy increases the threshold judges use for release. Assumption 4 strengthens the single index assumption underlying the MTE framework of Heckman and Vytlacil (1999, 2005) by assuming the index U is unchanged by the policy a—Heckman and Vytlacil (2005, 2007) call this condition policy invariance. A useful implication of Assumption 4 is that the change in the average outcome from implementing the counterfactual policy, $\theta - E[Y]$, is simply a functional of the MTE curve, $MTE(u) = E[Y(1) - Y(0) \mid U = u]$. Without Assumption 4, when U is not invariant to the policy, knowing the MTE curve may be insufficient for counterfactual policy analysis.

Note that Assumption 4 implies both Assumption 2 and Assumption 3, and thus will be questionable whenever Assumption 3 is questionable. Another way to see the difference between Assumptions 2, 3 and 4 is that if we had data under both the status quo and the counterfactual, we could consider two possible instruments, Z and A, where the latter is an indicator for whether the counterfactual policy is implemented. Assumption 2 corresponds to the IV monotonicity assumption where the binary policy variable A is the instrument. Assumption 3 corresponds to IV monotonicity for the multivalued instrument Z, while Assumption 4 corresponds to IV monotonicity for the two-dimensional instrument $\tilde{Z} = (Z, A)$. However, as argued in Heckman et al. (2006) and Mogstad et al. (2021), with multiple instruments, IV monotonicity imposes strong homogeneity conditions on treatment choice, making it hard to justify Assumption 4 in many contexts. Nevertheless, it is useful conceptually to consider what we can learn under Assumption 4, since as we will see below, for some policy exercises Assumption 4 will be useful while Assumption 3 will not. This suggests that relaxations of Assumption 4 may be useful in practice, a topic we turn to in Section 5 below.

3 The identified set without monotonicity or policy-invariance

In this section, we derive a tractable characterization of the bounds of the identified set Θ_I, without imposing IV monotonicity or policy invariance. Previous work that has considered partial identification in related IV settings has typically characterized the identified set by defining response types (a.k.a. principal strata) based on the values of $D(\cdot,\cdot)$ (e.g. Bai, Huang, Moon, Shaikh, & Vytlacil, 2025; Kitagawa, 2021). The observable data on (Y, D, Z) is then a mixture of the distributions of the potential outcomes across the response types, and one can characterize the identified set for the primitives $\mathcal{P}_I^*(P;\mathcal{P}^*)$ by searching over mixtures of types that match the observed data. The identified set for θ can then be calculated by computing the minimum and maximum value of $E_{P^*}[Y(D(Z,1))]$ among $P^* \in \mathcal{P}_I^*(P;\mathcal{P}^*)$.

The challenge with this approach is that if one does not impose instrument monotonicity, then the number of response types grows exponentially in the number of judges K. Specifically, if we impose no restrictions on $D(\cdot,\cdot)$, then there are four possible values of $(D(z,0), D(z,1))$ for each value of z, and hence 4^K possible response types. Imposing policy monotonicity (Assumption 2) rules out response types with $(D(z,0), D(z,1)) = (1,0)$ for some z, which still leaves 3^K possible response types. In either case, the number of response becomes extremely large with even a moderate number of judges: when $K = 30$, for example, we have $3^K \approx 2 \cdot 10^{14}$ and $4^K \approx 10^{18}$, which makes characterization of the identified set by type enumeration computationally infeasible.

Fortunately, we will show that it is not necessary to enumerate response types to derive the identified set for the counterfactual policy of interest. The key observation is that we need not search over possible distributions P^* for the model primitives; it is sufficient to restrict our attention to the collection of marginal distributions $\{Y(0), Y(1), D(z,\cdot)\}_z$ that involve the potential outcomes and the potential treatments *for one judge at a time*. To see why we need not consider the joint behavior of the judges, observe that the counterfactual outcome can be written

$$\theta = E_{P^*}[Y(D(Z,1))] = \sum_z P(Z=z) \cdot E_{P^*}[Y(D(z,1)) \mid Z=z],$$

which depends on P^* only through the collection of marginals for $\{Y(0), Y(1), D(z,1)\}_z$ and the marginal distribution of Z. Likewise, the probability distribution of the observable data takes the form

$$P(Y \in A, D=d, Z=z) = P(Z=z) \cdot P^*(Y(d) \in A, D(z,0)=d),$$

which also depends on P^* through the collection of marginals for $\{Y(0), Y(1), D(z,0)\}_z$ and the marginal distribution of Z. It follows that both the observable data distribution P and the average outcome under the counterfactual depend on P^* only through the marginal distributions $\{Y(0), Y(1), D(z,\cdot)\}_z$ and the marginal distribution of Z—they do not depend on the joint distribution of $D(z,\cdot)$ and $D(z',\cdot)$ for $z \neq z'$. Moreover, if we are not imposing any cross-judge restrictions, then our constraints on the model also only restrict the marginals of $\{Y(0), Y(1), D(z,\cdot)\}_z$. This suggests that to compute the identified set for θ, we can simply optimize over sets of marginals for $\{Y(0), Y(1), D(z,\cdot)\}_z$ that are consistent with the observable data.

Proposition 1 and Corollary 1 below formalize these observations. For ease of exposition, we state the results for the special case with discrete outcomes, and defer the general statement, which involves more notation, to Proposition 5 in Appendix A.1. Suppose that the support $\mathcal{Y}$ of the outcomes is discrete, so that the collection of marginal distributions $\{Y(0), Y(1), D(z,\cdot)\}_z$ can be characterized by the collection of marginal probability mass functions $\{\pi_z(\cdot)\}_z$, where $\pi_z(y_0, y_1, d_0, d_1) = P^*(Y(0) = y_0, Y(1) = y_1, D(z,0) = d_0, D(z,1) = d_1)$. Our first result gives a simple way of verifying whether a conjectured collection of marginals $\{\pi_z(\cdot)\}_z$ is consistent with the data in the sense that there exists a distribution P^* that generates both the data distribution P and the conjectured marginals.

Proposition 1. *Suppose that $\mathcal{Y}$ is finite. There exists a joint distribution $P^* \in \mathcal{P}_I^*(P; \mathcal{P}^*_{\text{valid}})$ with marginals $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$ if and only if $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$ satisfy the following three conditions:*

1. *They match the observable data: for every $y \in \mathcal{Y}$, $z \in \mathcal{Z}$*

$$P(Y = y, D = 1 \mid Z = z) = \sum_{y_0\in\mathcal{Y}, d_1\in\{0,1\}} \pi_z(y_0, y, 1, d_1),$$
$$P(Y = y, D = 0 \mid Z = z) = \sum_{y_1\in\mathcal{Y}, d_1\in\{0,1\}} \pi_z(y, y_1, 0, d_1).$$

2. *They imply the same distribution for $(Y(0), Y(1))$: for all $y_0, y_1 \in \mathcal{Y}$ and for any $z \in \mathcal{Z}$*

$$\sum_{(d_0,d_1)\in\{0,1\}^2} \pi_z(y_0, y_1, d_0, d_1) = \sum_{(d_0,d_1)\in\{0,1\}^2} \pi_1(y_0, y_1, d_0, d_1).$$

3. *They are valid probability mass functions: for all $(y_0, y_1, d_0, d_1) \in \mathcal{Y}^2 \times \{0,1\}^2$, and $z \in \mathcal{Z}$, $\pi_z(y_0, y_1, d_0, d_1) \geq 0$ with $\sum_{(y,y',d,d')\in\mathcal{Y}^2\times\{0,1\}^2} \pi_z(y, y', d, d') = 1$.*

Proposition 1 implies that to compute the identified set for θ, instead of searching over all distributions P^*, it suffices to search over the marginal probability mass functions that

satisfy the intuitive conditions 1–3. As the following corollary shows, this holds so long as the restrictions on $\mathcal{P}^*$ only take the form of marginal restrictions, as in Assumption 1.

Corollary 1. *Suppose that $\mathcal{Y}$ is finite and that P^* satisfies eq. (2) and Assumption 1 for some convex $\mathcal{R}^*$, but no further restrictions are placed on $\mathcal{P}^*$, that is $\mathcal{P}^* = \mathcal{P}^*_{\text{valid}} \cap \{P^* \colon \{\pi_z(\cdot)\} \in \mathcal{R}^*\}$. Then $\Theta_I(P; \mathcal{P}^*)$ is given by an interval, with the upper endpoint given by the optimization*

$$\sup_{\{\pi_z\}_{z\in\mathcal{Z}} \in \mathcal{R}^*} \sum_{z\in\mathcal{Z}} P(Z=z) \sum_{y_0,y_1,d_0,d_1 \in \mathcal{Y}^2\times\{0,1\}^2} (d_1 y_1 + (1-d_1)y_0) \cdot \pi_z(y_0, y_1, d_0, d_1) \tag{3}$$

subject to the constraints 1–3 in Proposition 1; the lower endpoint is given by an analogous minimization.

Computation via linear programming Corollary 1 implies that if $\mathcal{R}^*$ restricts the marginals $\{\pi_z\}_{z\in\mathcal{Z}}$ linearly, the identified set can be computed by solving a linear program that scales *linearly* with the number of instruments K.

To illustrate, suppose that the only restriction imposed by $\mathcal{R}^*$ is that it restricts the fraction of defendants released under the counterfactual, $E[D(z,1)] = \alpha_{1,z}$. For a given collection of marginal probability mass functions $\{\pi_z(\cdot)\}_z$, let π denote the vector of length $4|\mathcal{Y}|^2K$ that stacks all values of the marginals. Then the value of the objective function in eq. (3) can be written $\sum_{z,y_0,y_1,d_0,d_1} P(Z=z)(d_1y_1 + (1-d_1)y_0)\pi_z(y_0,y_1,d_0,d_1) = \omega'\pi$, where the weighting vector ω stacks the weights $P(Z=z)(d_1y_1 + (1-d_1)y_0)$. Furthermore, the constraints 1–3 in Proposition 1 can be written as $2|\mathcal{Y}|K + |\mathcal{Y}|^2(K-1) + 4|\mathcal{Y}|^2K + 1 = O(|\mathcal{Y}|^2K)$ linear constraints in π. Finally, the K constraints $E[D(z,1)] = \alpha_{1,z}$ can be written as $\sum_{y_0,y_1,d_0} \pi_z(y_0,y_1,d_0,1) = \alpha_{1,z}$, which are also linear in π. It follows that the optimization problem in eq. (3) can be written as a linear program in $O(K|\mathcal{Y}|^2)$ variables, subject to $O(K|\mathcal{Y}|^2)$ constraints. A related result appeared in Richardson and Robins (2014), who show that for binary outcomes, bounds on $E[Y(d)]$ when $\mathcal{R}^*$ is unrestricted can be obtained as a solution to $O(K^2)$ inequalities. Concurrent work by Song et al. (2025) shows that the joint distribution of $(Y(1), Y(0))$ when $\mathcal{R}^*$ is unrestricted can be characterized by $O(K)$ inequalities. Their result, however, does not directly cover policy counterfactuals like $\theta = E[Y(D(Z,1))]$, which is a functional of the joint distribution of $Y(\cdot)$ and $D(\cdot,\cdot)$.

The key feature of this program is that its dimension scales linearly with K, rather than exponentially, which makes it computationally fast even when K is large. The linear scaling is preserved if we restrict $\mathcal{R}^*$ in other ways, so long as the restrictions are linear. For example, imposing policy monotonicity (Assumption 2) amounts to adding the constraint that $\sum_{y_0,y_1} \pi_{y_0,y_1,1,0|z} = 0$ for all z.

Remark 1 (Discretizing Y)**.** If Y is continuous, one can obtain conservative bounds by considering a discretized version of Y. Given an initial grid of Q points, let Y^{ub} be Y rounded up to the nearest grid point. Since by construction $Y^{ub} \geq Y$, it follows that $E[Y^{ub}(D(Z,1))] \geq E[Y(D(Z,1))] = \theta$. Hence, computing the upper bound of the identified set for $E[Y^{ub}(D(Z,1))]$ yields a potentially non-sharp upper bound on θ. Likewise, one can obtain a conservative lower bound by computing the lower bound of the identified set after rounding down Y to the nearest grid point.

4 Does IV monotonicity help tighten the identified set?

The previous section characterized the identified set for the counterfactual outcome $\theta = E[Y(D(Z,1))]$ without imposing IV monotonicity. This section evaluates the extent to which imposing IV monotonicity helps tighten the identified set for θ. We present sufficient conditions under which imposing IV monotonicity alone does not help tighten the identified set. If these conditions hold, then the debate over the validity of IV monotonicity is somewhat of a red herring from the perspective of learning about counterfactuals, and the researcher should focus on alternative assumptions—such as policy invariance (and relaxations thereof)—that may help tighten the identified set.

Simple example for intuition. We begin with a simple example to illustrate why imposing IV monotonicity need not help tighten the identified set. Suppose that $a = 1$ corresponds to a universal release policy, so that $D(Z,1) = 1$ with probability 1. Suppose further that Y is binary and that $Y(0)$ is known to equal zero with probability one. This restriction arises frequently in the criminal justice setting, where, for example, one cannot fail to appear in court if they are detained awaiting trial.[7] The parameter of interest then reduces to $\theta = E[Y(1)]$.

In this setting, it is straightforward to show that without IV monotonicity, intersecting judge-specific intervals, as suggested in Manski (1990), in fact yields the sharp identified set that we characterized in Corollary 1 (Lemma 3.1 of Bai, Huang, Moon, Shaikh, and Vytlacil (2025) makes a similar observation). In particular, bounds for θ using *only* data on a given

[7]There are other leniency designs in which one of the potential outcomes is known to equal zero. For example, Baron et al. (2024) study child welfare investigations and define the outcome Y to be 1 if there is evidence of future misconduct in the child's original home, and thus by construction $Y = 0$ if a child is placed in foster care (i.e. $Y(1) = 0$). Likewise, Chan et al. (2022) study whether doctors diagnose patients with pneumonia, and define Y to be 1 if the patient subsequently shows signs of undiagnosed pneumonia. Thus, in their setting $Y(1) = 0$ by construction.

judge z are given by the interval

$$\mathcal{I}_z = [P(Y=1, D=1 \mid Z=z), 1 - P(Y=0, D=1 \mid Z=z)]. \tag{4}$$

The lower bound corresponds to the value of $E[Y(1)]$ if everybody not released by z has $Y(1) = 0$; the upper bound to the value if everybody not released by z has $Y(1) = 1$. The identified set for θ then corresponds to the intersection of the judge-specific intervals, $\Theta_I = \bigcap_z \mathcal{I}_z$.

On the other hand, under IV monotonicity, it is straightforward to show that the identified set corresponds to the judge-specific interval for the most lenient judge under the status quo, i.e. $\mathcal{I}_{z_{\max}}$, where $z_{\max} = \text{argmax}_z \, E[D(z,0)]$ is the identity of the most lenient judge under the status quo. Specifically, note that IV monotonicity implies that any defendant not released by judge $z_{\max}$ is also not released by any other judge, so that $D(z_{\max}, 0) = 0$ implies $D(z, 0) = 0$ for all z. It follows that the distribution of the observable data does not depend at all on the value of $Y(1)$ for defendants not released by judge $z_{\max}$, i.e., any binary distribution for $Y(1) \mid D(z_{\max}, 0) = 0$ is compatible with the observable data. Note, however, that by iterated expectations, $E[Y(1)]$ is a weighted average of $Y(1)$ for the defendants released and not released by judge $z_{\max}$, i.e. $E[Y(1)] = P^*(D(z_{\max}, 0) = 1)P^*(Y(1) = 1 \mid D(z_{\max}, 0) = 1) + P^*(D(z_{\max}, 0) = 0)P^*(Y(1) = 1 \mid D(z_{\max}, 0) = 0)$. The first term is point-identified from the defendants that judge $z_{\max}$ releases. The identified set therefore corresponds to the interval obtained by placing trivial bounds on the outcome for defendants not released by $z_{\max}$, $P^*(Y(1) = 1 \mid D(z_{\max}, 0) = 0) \in [0, 1]$, which yields the interval $\mathcal{I}_{z_{\max}}$.

We thus see that without imposing IV monotonicity, we obtain the identified set $\bigcap_z \mathcal{I}_z$, whereas if we do impose monotonicity, we obtain the identified set $\mathcal{I}_{z_{\max}}$. Since adding an assumption cannot widen the identified set, and since $\bigcap_z \mathcal{I}_z$ is contained in $\mathcal{I}_{z_{\max}}$, it follows that when IV monotonicity holds, the identified set is the same whether we impose IV monotonicity or not. On the other hand, if IV monotonicity does not hold and is rejected by the data, it may be the case that the set $\bigcap_z \mathcal{I}_z$ is a strict subset of $\mathcal{I}_{z_{\max}}$: the sharp identified set from Corollary 1 above without imposing monotonicity is then tighter than the *naïve* identified set $\mathcal{I}_{z_{\max}}$ that assumes IV monotonicity holds in the data (if IV monotonicity is rejected, then the identified set under IV monotonicity is formally empty).

The intuition for why IV monotonicity does not help in this setting is that under IV monotonicity, we learn nothing about the defendants not released by the most lenient judge, $z_{\max}$. By contrast, if IV monotonicity is violated, then some defendants not released by $z_{\max}$ *may be released* by another judge z', and thus the data provides some information about their value of $Y(1)$. Hence, IV monotonicity is in this sense the *least favorable* configuration

of judge release decisions, and thus imposing IV monotonicity does not help us tighten the identified set.

More general sufficient conditions. Our simple example above had two salient features: (i) one of the potential outcomes was known ($Y(0) = 0$), and (ii) the policy encouragement was very strong ($D(z, 1) = 1$). We next present a generalization showing that either of these features on its own is sufficient for IV monotonicity not to tighten the identified set. We first show that if $Y(0) = 0$, then the identified set for any counterfactual policy satisfying policy monotonicity ($D(z, 1) \geq D(z, 0)$) does not depend on whether we impose IV monotonicity, regardless of whether the policy encouragement is strong or weak. Second, we show that if both potential outcomes are non-trivial, then IV monotonicity does not help to tighten the identified set provided that one assumes the policy encouragement is "sufficiently strong", as formalized in condition (iii) below.

We first consider the setting where one of the potential outcomes is known. For simplicity of notation, we establish our result in the context of a counterfactual policy that imposes an *average policy quota* that restricts the average value of $D(Z, 1)$. Fix $\alpha \in [0, 1]$. Let $\mathcal{P}^*_\alpha$ be the set of distributions over the random vector $(Y(0), Y(1), D(\cdot, \cdot), Z)$ satisfying the IV validity condition (2), policy monotonicity (Assumption 3), and also

(i) *α-average policy quota:* $P^*(D(Z, 1) = 1) = \alpha$,

(ii) *Known outcome under $D = 0$:* $P^*(Y(0) = 0) = 1$.

Let $\mathcal{P}^*_{\alpha,Mon}$ be the subset of distributions in $\mathcal{P}^*_\alpha$ that further satisfy IV monotonicity (Assumption 3).

Proposition 2 (No identifying power of monotonicity with known $Y(0)$)**.** *If the identified set $\Theta_I(P; \mathcal{P}^*_{\alpha,Mon})$ is non-empty, then Assumption 3 has no identifying power in the sense that $\Theta_I(P; \mathcal{P}^*_{\alpha,Mon}) = \Theta_I(P; \mathcal{P}^*_\alpha)$.*

Although condition (i) focuses on a policy that imposes an average quota, it is possible to extend the proposition to the case in which there is a unit-specific quota α_z, at the cost of additional notation.

We next consider the setting with a sufficiently strong policy encouragement, in the sense that the policy satisfies the following condition:

(iii) *Sufficiently strong encouragement:* $P^*(D(z, 1) = 1 \mid D(z_{\max}, 0) = 1) = 1$ for all z.

Condition (iii) imposes that all defendants who would be released by the most lenient judge under the status quo would be released with probability 1 under the counterfactual. This is

trivially satisfied under a universal release program ($D(z,1) = 1$), but is somewhat weaker. For example, consider a program that releases all defendants except those predicted to be high-risk. Then (iii) will be satisfied, provided the most lenient judge is not currently releasing any such high-risk defendants under the status quo. Our next result shows that monotonicity also has no identifying if we replace condition (ii) defined above with condition (iii) in the statement of Proposition 2.

Proposition 3 (No identifying power of monotonicity with strong encouragement)**.** *The result in Proposition 2 holds if condition (ii) is replaced with condition (iii) in the definitions of* $\mathcal{P}^*_\alpha$ *and* $\mathcal{P}^*_{\alpha,Mon}$*.*

Remark 2 (Interaction of IV monotonicity with other constraints)**.** Propositions 2 and 3 provide conditions under which IV monotonicity *alone* does not help tighten the identified set for the average counterfactual outcome. As shown by Machado et al. (2019), if we impose additional restrictions —such as assume that all units share the same sign of the treatment effect $Y(1) - Y(0)$—then adding monotonicity can help to further tighten the identified set.

Remark 3 (Weaker monotonicity conditions)**.** Frandsen et al. (2023) propose a weakening of IV monotonicity called *average monotonicity*. Propositions 2 and 3 provide conditions under which imposing IV monotonicity does not help tighten the identified set. It follows immediately that under the same conditions, imposing the weaker notion of average monotonicity also does not help tighten the identified set.

Remark 4 (Relationship to literature)**.** Propositions 2 and 3 can be viewed as an extension of some important existing results in the literature which show that IV monotonicity has no identifying power for the average treatment effect parameter or the marginal distributions of potential outcomes (Bai, Huang, Moon, Shaikh, & Vytlacil, 2025; Balke & Pearl, 1997; Kitagawa, 2021). An important difference vis-à-vis our result is that we focus on more general policy counterfactuals.

Remark 5 (Can monotonicity help?)**.** In Appendix B.1, we give an example of a policy and a data distribution such that IV monotonicity sharpens the identified set for the average counterfactual policy outcome. This point relates to an observation in Kamat (2019), showing that the identified set for the joint distribution of $(Y(1), Y(0))$ shrinks under IV monotonicity. For policies that do not satisfy the strong encouragement condition (iii), IV monotonicity can help by restricting the possible couplings of the marginal distributions of $Y(0)$ and $Y(1)$, even though it has no identifying power for the marginal distributions.

5 What other assumptions help tighten the identified set?

The previous section outlined sufficient conditions under which IV monotonicity alone does not help to tighten the identified set. We now explore other assumptions that can potentially help to tighten it. We begin by showing that policy invariance can be helpful in some settings where IV monotonicity is not. Policy invariance may be too strong an assumption in practice, however, and we therefore introduce relaxations of policy invariance that may be more plausible, but nevertheless tighten the identified set. We then briefly discuss other economically motivated restrictions that may be reasonable and further help tighten the identified set.

5.1 Policy invariance and its relaxations

We begin by providing an intuitive example where the conditions of Proposition 2 are satisfied, so that IV monotonicity alone does not help tighten the identified set, but policy invariance is helpful.

Example: quota policy. Suppose that $Y(0) = 0$ (e.g., you cannot commit a crime while in jail). Consider a quota policy that requires all judges to increase their release rate to match that of the most lenient judge under the status quo, so that $E[D(z,1)] = E[D(z_{\max},0)]$ for all z, where again $z_{\max}$ denotes the identity of the most lenient judge. It seems reasonable to impose policy monotonicity in this setting ($D(z,1) \geq D(z,0)$), which states that defendants who would be released without the quota would also be released with the quota.

Strengthening policy monotonicity by imposing policy invariance leads to point identification. Intuitively, under policy invariance, all judges have the same ranking of defendants under both the status quo and counterfactual, but they disagree on the cutoff for when a defendant should be released. The quota policy forces them to use the same cutoff as the most lenient judge, so that the counterfactual outcomes for all judges will match that of the most lenient judge under the status quo: $E[Y(D(Z,1))] = E[Y(D(z_{\max},0))] = E[Y \mid Z = z_{\max}]$.

By contrast, Proposition 2 implies that imposing IV monotonicity in addition to policy monotonicity does not help tighten the identified set. In fact, the identified set may be trivial in the sense that it implies trivial bounds for the outcomes of policy compliers, $E[Y(1) \mid D(Z,1) > D(Z,0)]$. Consider, for example, a setting where there are two judges, who respectively release 10% and 20% of defendants. Under IV monotonicity alone, the first judge could choose to match the quota by marginally releasing only status quo *never-takers*—those

not released by either judge under the status quo. The data does not restrict $Y(1)$ for these individuals, so we only have trivial bounds $[0, 1]$ on their treated outcomes, implying trivial bounds for the policy complier treatment effect. Correspondingly, by eq. (1), the identified set for $\theta - E[Y]$ under IV monotonicity is given by multiplying the unit interval by the mass of policy compliers. Thus, policy invariance—which restricts agreement under judges under both the counterfactual and the status quo—has strong identifying power, leading to point identification. By contrast, IV monotonicity—which only restricts agreement under the status quo—does not help tighten the trivial bounds on the identified set.

Of course, policy invariance may be implausibly strong in many applied settings. Indeed, since policy invariance is *stronger* than IV monotonicity, if researchers doubt the validity of IV monotonicity (or it is rejected by the data) then they must necessarily doubt the validity of policy invariance as well. Nevertheless, the fact that there are relevant settings where policy invariance can help tighten the identified set, but IV monotonicity cannot, suggests that considering *relaxations* of policy invariance may be a more natural starting point than considering relaxations of IV monotonicity.

Relaxing policy invariance with disagreement bounds. To that end, we now introduce a relaxation of policy invariance that may be more plausible but nevertheless help to tighten the identified set. At a high level, policy invariance requires that judges perfectly agree on the ranking of defendants; we consider a relaxation that bounds the extent to which they can disagree. More concretely, consider two distinct judges z and z', and two policies $a, a' \in \{0, 1\}$. Without loss of generality, suppose that $E[D(z, a)] \geq E[D(z', a')]$ so that judge z releases more people under policy a than z' does under a'. Under policy invariance, judge z under policy a must release *all* defendants released by z' under policy a', so that $P^*(D(z, a) = 1 \mid D(z', a') = 1) = 1$. Such perfect agreement is likely to be too strong in many settings. Nevertheless, it also may be unreasonable to expect that the two judges perfectly *disagree*, so that judge z under policy a releases *none* of the defendants released by z' under a'. A natural middle-ground is to impose

$$P^*(D(z, a) = 0 \mid D(z', a') = 1) \leq \delta_{z,z',a,a'}, \tag{5}$$

so that judge z under policy a disagrees with no more than $\delta_{z,z',a,a'}$ fraction of defendants released by judge z' under policy a'. This nests policy invariance as the special case with $\delta_{z,z',a,a'} = 0$, but allows for non-trivial disagreement for $\delta_{z,z',a,a'} \in (0, 1)$.

Calculating the identified set under disagreement bounds. The disagreement bound in eq. (5) imposes restrictions on the joint distribution of $(D(z, a), D(z', a'))$. At first glance,

such restrictions appear to be difficult to incorporate into the linear programming approach for calculating the identified set described in Section 3, which only optimizes over the marginals $\{\pi_z(\cdot)\}$, and does not enumerate the joint distribution of $(D(z,a), D(z',a'))$. It turns out, however, that given a set of marginals $\{\pi_z(\cdot)\}$, there is a simple formula for the minimal value of $P^*(D(z,a)=0 \mid D(z',a')=1)$ consistent with a joint distribution of primitives that matches these marginals and satisfies policy monotonicity. This allows us to still tractably compute the identified set by optimizing only over the marginals $\{\pi_z(\cdot)\}$ even while imposing disagreement bounds such as eq. (5). This is formalized in the following results, which give analogs to Proposition 1 and Corollary 1 that allow for imposing disagreement bounds.

Proposition 4. *Suppose that $\mathcal{Y}$ is finite. Fix disagreement bounds $\delta_{z,z',a,a'} \in [0,1]$ for all z, z', a, a'. Let $\mathcal{P}^*_{\text{DB}}$ be the subset of distributions in $\mathcal{P}^*_{\text{valid}}$ satisfying Assumption 2 and the disagreement bounds in eq. (5) for all z, z', a, a'. Consider a collection $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$ of marginals that satisfy Assumption 2. There exists a joint distribution $P^* \in \mathcal{P}^*_I(P; \mathcal{P}^*_{\text{DB}})$ that generates $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$ if and only if the marginals $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$ are consistent with Assumption 2 and satisfy conditions 1–3 in Proposition 1 as well as*

4. *Disagreement bound: For all z, z', a, a',*

$$\sum_{y_1,y_0} \min\{\pi(y_0,y_1,D(z,a)=1), \pi(y_0,y_1,D(z',a')=1)\} \\ \geq (1-\delta_{z,z',a,a'}) \cdot \pi(D(z',a')=1), \quad (6)$$

where for any $(y_0,y_1) \in \mathcal{Y}^2$ and $(a,z) \in \{0,1\} \times \mathcal{Z}$,

$$\pi(y_0,y_1,D(z,a)=1) := \sum_{d_a=1, d_{1-a}\in\{0,1\}} \pi_z(y_0,y_1,d_0,d_1),$$
$$\pi(D(z,a)=1) := \sum_{d_a=1, d_{1-a}\in\{0,1\}} \sum_{(y_0,y_1)\in\mathcal{Y}^2} \pi_z(y_0,y_1,d_0,d_1).$$

Corollary 2. *Suppose that $\mathcal{Y}$ is finite and let $\mathcal{P}^* = \mathcal{P}^*_{DB} \cap \{P^* : \{\pi_z(\cdot)\} \in \mathcal{R}^*\}$ for some convex $\mathcal{R}^*$, where $\mathcal{P}^*_{DB}$ as defined in Proposition 4. Then $\Theta_I(P;\mathcal{P}^*)$ is given by an interval, with the upper endpoint given by the optimization*

$$\sup_{\{\pi_z\}_{z\in\mathcal{Z}}\in\mathcal{R}^*} \sum_{z\in\mathcal{Z}} P(Z=z) \sum_{y_0,y_1,d_0,d_1\in\mathcal{Y}^2\times\{0,1\}^2} (d_1y_1 + (1-d_1)y_0) \cdot \pi_z(y_0,y_1,d_0,d_1) \quad (7)$$

subject to the constraints 1–4 given in Propositions 1 and 4; the lower endpoint is given by

an analogous minimization.

The intuition for Proposition 4 is as follows. Note that eq. (5) can equivalently be written as a lower bound on the joint probability $P^*(D(z,a)=1, D(z',a')=1)$,

$$(1-\delta_{z,z',a,a'})P^*(D(z',a')=1) \leq P^*(D(z,a)=1, D(z',a')=1).$$

By the law of total probability, the right-hand side equals

$$\sum_{y_1,y_0} P^*(Y(0)=y_0, Y(1)=y_1, D(z,a)=1, D(z',a')=1) \leq$$
$$\sum_{y_1,y_0} \min\{P^*(Y(0)=y_0, Y(1)=y_1, D(z,a)=1), P^*(Y(0)=y_0, Y(1)=y_1, D(z',a')=1)\},$$

where the inequality uses the Fréchet–Hoeffding bound. Equation (6) follows from replacing $P^*(D(z,a)=1, D(z',a')=1)$ with the upper bound given in the previous display. This turns out not to come at the cost of sharpness, however, because given a set of marginals for $(Y(1), Y(0), D(z,\cdot))$, there always exists a coupling such that the inequality in the previous display holds with equality. In particular, the proof of Proposition 4 shows that the upper bound is achieved under a latent threshold crossing model wherein judges agree on the rankings of defendants *conditional* on their potential outcomes, so that $D(z,a) = \mathrm{I}\{\alpha_{z,y_0,y_1} + \beta_{z,y_0,y_1} a \geq V_{y_0,y_1}\}$ for a latent index V_{y_0,y_1} that depends only on the potential outcomes but not on z.

We note that it is straightforward to impose (6) in a linear program by introducing auxiliary parameters corresponding to the $\min\{\cdot,\cdot\}$ terms. Specifically, eq. (6) holds if and only if there exist constants $\eta_{z,z'}^{y_0,y_1} \geq 0$ that are less than the arguments of the $\min\{\cdot,\cdot\}$,

$$\eta_{z,z'}^{y_0,y_1} \leq \pi(Y(0)=y_0, Y(1)=y_1, D(z,a)=1), \tag{8}$$
$$\eta_{z,z'}^{y_0,y_1} \leq \pi(Y(0)=y_0, Y(1)=y_1, D(z',a')=1) \tag{9}$$

such that

$$\sum_{y_1,y_0} \eta_{z,z'}^{y_1,y_0} \geq (1-\delta_{z,z'})\pi(D(z',a')=1). \tag{10}$$

The above formulation is linear in both the marginals $\{\pi_z(\cdot)\}$ and the auxiliary parameter η. Therefore, we can implement the disagreement bound in eq. (5) by adding the constraints in eqs. (8) to (10) to the linear programming formulation in Section 3, and augmenting the parameter vector π by η.

Bounds on average disagreements. Instead of restricting pairwise disagreements between judges, we can alternatively bound disagreements between groups of judges. For concreteness, consider the quota policy from our empirical applications in Section 6, which asks the bottom 90% of judges to match the release rate of the most lenient decile. Let $\mathcal{Q}$ denote the set of judges in the bottom nine deciles of leniency, who are subject to the quota, and let $\mathcal{Q}^c$ denote the top decile of judges, who are not subject to the quota. To bound the average disagreement between the two groups of judges under the counterfactual, we can impose the average disagreement probability bound

$$P^*(D(Z_{\mathcal{Q}}, 1) = 0 \mid D(Z_{\mathcal{Q}^c}, 1) = 1) \leq \overline{\mathsf{DP}}, \tag{11}$$

where $Z_{\mathcal{Q}}$ is a randomly picked judge from the group $\mathcal{Q}$, and $Z_{\mathcal{Q}^c}$ is defined similarly. In other words, the disagreement probability bound $\overline{\mathsf{DP}}$ represents the average probability that a judge in the quota group disagrees with the release decision of a judge in the top decile. Similarly to the pairwise disagreement bounds, this average disagreement probability bound can be implemented as a linear program without losing sharpness of the resulting identified set. In particular, we can introduce auxiliary parameters $\eta_{z,z'}^{y_1,y_0}$, for each $z \in \mathcal{Q}$ and $z' \in \mathcal{Q}^c$, and impose the linear constraints eqs. (8) and (9), as well as the constraint $\sum_{y_1,y_0} \frac{1}{|\mathcal{Q}||\mathcal{Q}^c|} \sum_{z\in\mathcal{Q}, z'\in\mathcal{Q}^c} \eta_{z,z'}^{y_0,y_1} \geq (1-\delta)P^*(D(Z_{\mathcal{Q}^c}, 1) = 1)$.

The attractive feature of the average disagreement probability bound is that it only requires calibration of a single parameter, $\overline{\mathsf{DP}}$. In contrast, imposing pairwise disagreement bounds requires calibration of $\delta_{z,z'}$ for each pair of judges, which may be difficult in practice.[8] In Appendix B.2, we discuss how researchers can calibrate the parameter $\overline{\mathsf{DP}}$ using a Gaussian signal model parametrized to match the setting in Sigstad (2026), in which we observe decisions of multiple judges on the same case. We use this calibration to inform our empirical application in Section 6.[9]

[8] Our average disagreement bound focuses on the case where we only bound disagreement under the counterfactual. This is sufficient for the quota counterfactual, since we assume judges in the top decile do not change their behavior. For other policies, it may be informative to impose bounds on $P^*(D(Z_{\mathcal{Q}}, a) = 0 \mid D(Z_{\mathcal{Q}^c}, 1) = a')$ for pairs of (a, a') other than $(1, 1)$.

[9] In principle, one could use the Sigstad (2026) data to directly calculate the sample analog of the disagreement probability $P^*(D(Z_{\mathcal{Q}}, 1) = 0 \mid D(Z_{\mathcal{Q}^c}, 1) = 1)$. However, the disagreement probability is not invariant to the marginal release rates (if two judges have release rates equal to 99%, they cannot disagree on more than 2% of cases; by contrast, two judges with release rates of 50% could disagree 100% of the time). We therefore use a Gaussian signal model to compute the implied disagreement probability when the marginal release rates are matched to those in our application. The Gaussian signal model implicitly assumes that judges serving on panels rule the same way as when making decisions on their own. To weaken this assumption, one could estimate a richer model that allows for spillovers (e.g. Iaryczower & Shum, 2012).

Limitations of policy invariance. The quota example given above showed that policy invariance can help tighten the identified set (and even restore point identification) in a setting where IV monotonicity alone is not informative. It turns out, however, that there are some situations in which neither IV monotonicity nor policy invariance helps to tighten the identified set. To gain intuition, consider our earlier example of a universal release policy such that $D(Z,1) = 1$ with probability 1. In that case, by assumption, all judges perfectly agree on their decisions under the counterfactual policy. It follows that imposing policy invariance—which imposes IV monotonicity and restricts agreement under the counterfactual—is *equivalent* to imposing IV monotonicity in this setting, which we know is not helpful by Proposition 3. In settings like this, researchers will therefore have to turn to alternative restrictions to try to tighten the identified set.

5.2 Outcome restrictions

We now outline several other restrictions that researchers may consider imposing to help tighten the identified set.

Policy complier (PC) bounds. In the pretrial release setting, bail judges are legally instructed to release defendants based on pretrial misconduct potential if released ($Y(1)$). Consider a policy that encourages judges to release more defendants (e.g., a quota). Unless the judges' estimates of pretrial misconduct risk are terribly miscalibrated, policy compliers—individuals only released under the quota—should be on average higher risk than those currently released, but lower risk than those who are never released:

$$E[Y(1) \mid D(z,0) = 1] \leq E[Y(1) \mid D(z,0) < D(z,1)] \leq E[Y(1) \mid D(z,1) = 0]. \tag{12}$$

When the marginal release rates under the counterfactual ($E[D(z,1)]$) are known, eq. (12) can be written as a linear restriction on the marginals $\pi_z(\cdot)$, and is thus is straightforward to incorporate in the linear program for the identified set bounds.

Treatment effect bounds. It may sometimes be reasonable to impose bounds on the sign or magnitude of the treatment effects. For example, in some settings, it may be natural to impose the monotone treatment response assumption that $Y(1) \geq Y(0)$, as in Manski (1997) and Machado et al. (2019). This is straightforward to impose by setting $\pi_z(y_0, y_1, d_0, d_1) = 0$ whenever $y_1 < y_0$. Similarly, in some settings it may be reasonable to impose that the treatment effect for the policy compliers is not too large, e.g. $E[Y(1) - Y(0) \mid D(z,0) <$

$D(z,1)] \leq c$, which can be implemented as a linear constraint, similar to eq. (12).[10]

Outcome disparity bounds. Some counterfactual policies directly encourage a subset $\mathcal{Q}$ of decision-makers to act more like a benchmark group $\mathcal{Q}^c$. For example, in implementing a quota policy directive that asks the set $\mathcal{Q}$ to match the treatment rate of the group $\mathcal{Q}^c$, a policy-maker may encourage the decision-makers to also emulate their outcomes. We consider such a setting in our empirical application in Section 6.2 below. In such cases, an alternative to bounding the disagreements between their treatment decisions is to bound the disparity in the outcomes between the two sets of decision-makers by imposing

$$|E[Y(D(Z_{\mathcal{Q}},1)) - Y(D(Z_{\mathcal{Q}^c},1))]| \leq \overline{\mathsf{OD}}, \tag{13}$$

where, again, $Z_{\mathcal{Q}}$ is a randomly picked judge from the group $\mathcal{Q}$, and $Z_{\mathcal{Q}^c}$ is defined similarly.

To interpret this bound, it can be helpful to relate it to the disagreement probability bound in eq. (11) above. To this end, let $\mathcal{C}_{z,z'} = \mathrm{I}\{D(z,1) = 0, D(z',1) = 1\}$ denote the event that z' would release an individual but not z under the counterfactual. If the treatment rates of the two groups are equal, then $E[\mathcal{C}_{Z_{\mathcal{Q}},Z_{\mathcal{Q}^c}}] = E[\mathcal{C}_{Z_{\mathcal{Q}^c},Z_{\mathcal{Q}}}] = P^*(D(Z_{\mathcal{Q}},1) = 0, D(Z_{\mathcal{Q}^c} = 1))$. Using this, we may write the left-hand side of eq. (13) as

$$\begin{aligned} E[Y(D(Z_{\mathcal{Q}},1)) - Y(D(Z_{\mathcal{Q}^c},1))] = P^*(D(Z_{\mathcal{Q}},1) = 0, D(Z_{\mathcal{Q}^c} = 1)) \\ \times (E[Y(1) - Y(0) \mid \mathcal{C}_{Z_{\mathcal{Q}},Z_{\mathcal{Q}^c}}] - E[Y(1) - Y(0) \mid \mathcal{C}_{Z_{\mathcal{Q}^c},Z_{\mathcal{Q}}}]). \end{aligned}$$

This is the *product* of the aggregate disagreement probability, $P^*(D(Z_{\mathcal{Q}},1) = 0, D(Z_{\mathcal{Q}^c} = 1)) = P^*(D(Z_{\mathcal{Q}},1) = 0 \mid D(Z_{\mathcal{Q}^c} = 1))P^*(D(Z_{\mathcal{Q}^c} = 1))$ times the *difference* between treatment effects for individuals about whom there is disagreement. Thus, if we bound the treatment effect difference by Δ_{TE}, imposing eq. (11) implies that eq. (13) holds with

$$\overline{\mathsf{OD}} = \Delta_{TE}\overline{\mathsf{DP}} \cdot q, \tag{14}$$

where $q = P^*(D(Z_{\mathcal{Q}^c} = 1))$ is the marginal release rate of the group $\mathcal{Q}^c$. This relationship can be used to gauge the strength of the outcome disparity restriction, as we illustrate in Section 6.2 below.

[10] Frandsen et al. (2023) argue, for example, that it may be reasonable to restrict the magnitude of the treatment effect of incarceration among compliers between different sets of judges; here we impose an analogous constraint on the compliers with respect to a policy of interest.

6 Empirical Applications

6.1 NYC bail judges

Background. We study pretrial release in New York City (NYC) using data from Arnold et al. (2022, henceforth ADH22). A judge (Z) decides whether to release a defendant or hold them in pretrial detention. The defendant is categorized as released ($D = 1$) if they are released without preconditions, or after paying money bail; otherwise they are categorized as detained ($D = 0$). We then see whether a defendant commits pretrial misconduct ($Y \in \{0, 1\}$) by either failing to appear for a court hearing or by being arrested for a new crime. By construction, a defendant cannot commit pretrial misconduct if they are detained, so $Y(0) = 0$. We also observe the defendant's race, denoted by $R \in \{b, w\}$, where b and w denote black and white. We consider two counterfactual policy exercises in this context. First, we consider a policy that releases all defendants. Albright (2022) studies a universal release program for defendants (meeting certain conditions) in Kentucky, so this policy counterfactual directly addresses what would happen if such a program were implemented for ADH22's sample of defendants in NYC. Moreover, as we describe below, identification of the parameter of interest in ADH22, the disparate impact parameter, is equivalent to identification of race-specific average outcomes under this counterfactual. Second, we consider a quota policy that induces the bottom 90% of judges to match the leniency of the top 10%.

Disparate impact in ADH22. ADH22's primary goal is to estimate the "disparate impact" of pretrial release decisions by race. We now show that their measure of disparate impact is solely a function of observable probabilities and $E[Y(1) \mid R]$. Hence, learning about the disparate impact parameter is isomorphic to learning about the mean outcome (for each race) under the counterfactual in which everyone is released. To be more precise about the measures ADH22 consider, let $FPR_r := P(D = 1 \mid Y(1) = 1, R = r)$ be the false positive rate for race r, i.e., the fraction of defendants who are released despite having $Y(1) = 1$. Observe that we can write $FPR_r = \frac{P(D=1,Y=1|R=r)}{P^*(Y(1)=1|R=r)}$. The numerator is an observable probability, and thus the only challenge in learning about FPR_r is learning about the misconduct rate under the counterfactual in which everyone is released, $P^*(Y(1) = 1 \mid R = r)$. ADH22 similarly define $TPR_r := P(D = 1 \mid Y(1) = 0, R = r)$ to be the true positive rate, which we can write as $TPR_r = \frac{P(D=1,Y=0|R=r)}{1-P^*(Y(1)=1|R=r)}$. Again, the numerator is an observable probability. ADH22's main parameter of interest is a weighted average of the differences in FPR_r and TPR_r across races,

$$\Delta := (1 - E[Y(1)]) \cdot (TPR_b - TPR_w) + E[Y(1)] \cdot (FPR_b - FPR_w).$$

We thus see that identification of the disparate impact parameter Δ is isomorphic to identification of the race-specific counterfactual outcomes if everyone is released, $E[Y(1) \mid R = r]$.

Identification in ADH22. ADH22 consider an identification-at-infinity type argument for $E[Y(1) \mid R]$. Let $P_{z,r} = E[D(z,0) \mid R = r]$ be the judge-specific release rate for race r. ADH22 consider various functional forms for $E[Y(1) \mid D = 1, P_{z,r} = p, R = r]$ and then extrapolate to $p = 1$. Specifically, they report linear, local linear, and quadratic extrapolations. For comparison with our nonparametric approach, we replicate the point estimates and confidence intervals based on these three parametric extrapolation approaches.

Data. We analyze aggregated statistics from the main estimation sample in ADH22, which consists of the universe of arraignments made in NYC between 2008–2013 involving white or black defendants charged with a felony or misdemeanor, where the defendant is not already serving jail time for an unrelated charge. The sample comprises 284,598 cases involving white defendants and 310,588 cases involving black defendants. We do not have access to the individual microdata, and only observe estimates of the judge-specific release rates ($E[D(z,0) \mid R = r]$) and misconduct rates $E[Y(1) \mid D(z,0) = 1, R = r]$ that are obtained from linear regressions that adjust for court-by-time fixed effects, along with the number of observations for each judge and race (these estimates are plotted in Figure 2 in ADH22). Because we do not have access to the microdata, our inference ignores the statistical uncertainty stemming from the covariate adjustment: we treat the covariate-adjusted rates as if they were sample means. We show below that our replication of the results in ADH22 yields very similar standard errors as in the original, suggesting that the impact of the covariate adjustment on inference is minimal. To avoid small-sample issues stemming from observing only a few cases per judge, we pool all judges with fewer than 300 race-specific cases into a single judge, which leaves us with $K = 167$ and $K = 160$ distinct values for the instrument for the samples of black and white defendants, respectively.[11]

Summary statistics. Figure 1 plots the judge-specific misconduct rates against their release rates. We see that the misconduct rates increase sharply with the release rate, which is intuitive since by construction $Y = 0$ for non-released defendants. Correspondingly, the TSLS estimate of the effect of release on misconduct, which is equivalent to fitting a weighted least squares line through this scatterplot, is quite high, and equals 49.7% for blacks and 48.1% for whites (with very tight standard errors: 1.2 and 2.0). The average release rate

[11] Since the quota policy we consider below applies a quota to judges below the 90th percentile of leniency, we separately pool judges with fewer than 300 cases who lie below the 90th percentile of leniency, and those lying above it. Tables B.2 and B.3 show that our inference is virtually unchanged when we do not pool.

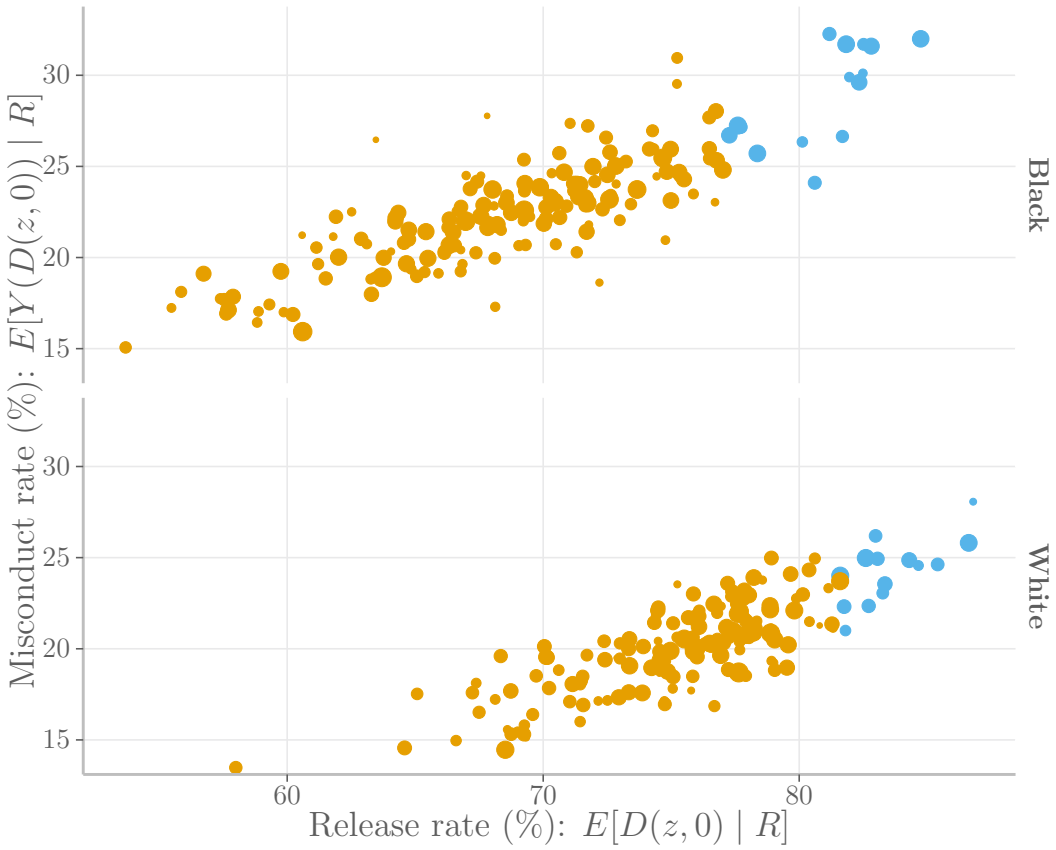

Figure 1: Judge- and race-specific sample release and misconduct rates, based on NYC bail judge data from Arnold et al. (2022).

Notes: Both rates adjusted for court-by-time fixed effects. Each dot corresponds to a single judge; dot size scales with the number of cases handled by the judge. Judges with fewer than 300 cases are pooled into a single dot. The top decile of most lenient judges is plotted in blue, the bottom 90% are plotted in orange.

equals 69.7% for blacks and 76.5% for whites. Unconditionally, the average misconduct rate equals 22.9% for blacks and 20.6% for whites; conditional on release, these rates equal 32.9%, and 26.9%.

Results for universal release policy. We compute estimates of the identified set for $E[Y(1) \mid R]$, the race-specific average outcomes under a universal release program, along with 95% confidence intervals computed by projection (see Appendix B.2 for inference details). Because $Y(0) = 0$, the sharp bounds on $E[Y(1) \mid R = r]$ imposing only IV validity (i.e., randomization and exclusion) are given by an intersection of judge-specific intervals given in eq. (4) above. For illustration, Figure 2 shows the form of the bounds (along with CIs) for black defendants: we compute sample analogs of the judge-specific bounds in eq. (4), and intersect them to estimate the identified set. While under monotonicity all the information

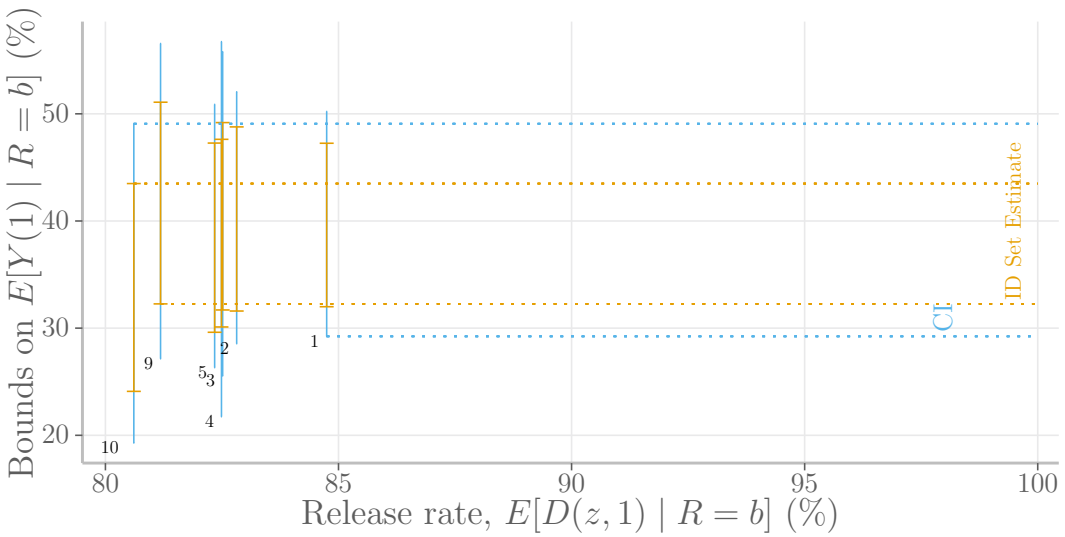

Figure 2: Illustration of identified set for $E[Y(1) \mid R = b]$ using NYC bail judge data.

Notes: The x-axis plots estimates of judge-specific release rates for each judge, adjusted for court-by-time fixed effects, while the y-axis plots estimates (in orange) and sup-t confidence intervals (in blue) of bounds on $E[Y(1) \mid R = b]$ using data on each judge separately. The plug-in estimate of the identified set and the associated confidence interval are formed by intersecting the judge-specific bounds and CIs, and are depicted on the right-hand side (the orange dotted lines give the plug-in estimates, and the blue dotted lines the CI). For clarity, we plot the judge-specific bounds only for a subset of the judges. Judges are numbered by their leniency, with 1 corresponding to the most lenient judge.

is obtained from the most lenient judge, we see that in practice the estimate of the identified set exploits information from other judges and thus is tighter than the interval obtained by looking at the most lenient judge alone.

Columns (1) and (3) of Table 1 give the estimates and confidence intervals for the race-specific average misconduct rates computed using our linear program as well as the parametric extrapolation methods in ADH22. For direct comparability to the confidence intervals using our approach, we compute standard errors without accounting for covariate adjustment.[12] Table B.1 in the appendix shows that the results are very similar to those reported in ADH22, suggesting that accounting for covariate adjustment doesn't substantively affect inference.

To help interpret the estimates, we also report estimates of the treatment effects for policy compliers, $E[Y(1) - Y(0) \mid D(Z,1) > D(Z,0)] = E[Y(1) - Y(0) \mid D(Z,0) = 0]$ in columns (2) and (4). The lower bound for policy compliers is above the misconduct rate for white defendants released under the status (31.9% vs. 26.9%) and very similar to the status quo rate for black defendants (30.8% vs. 32.9%), suggesting that the marginally-released

[12]ADH22 also weight their extrapolations by the inverse of the estimated sampling variance for the judge-level covariate-adjusted outcome means. Since we do not have these weights, we instead weight by the number of defendants the judge released.

	Blacks		Whites		
	$E[Y(1) \mid R]$	PC TE	$E[Y(1) \mid R]$	PC TE	Δ
Specification	(1)	(2)	(3)	(4)	(5)
Valid IV only	$[32.3, 43.5]$	$[30.8, 67.9]$	$[28.1, 39.2]$	$[31.9, 79.2]$	$[4.0, 8.2]$
	$(29.2, 49.1)$	$(20.9, 86.3)$	$(23.4, 41.9)$	$(12.0, 90.7)$	$(1.0, 9.6)$
Linear extrap.	40.0	56.3	33.5	54.8	5.4
	$(38.9, 41.0)$	$(52.9, 59.6)$	$(32.3, 34.6)$	$(49.8, 59.8)$	$(5.0, 5.8)$
Quadratic extrap.	42.8	65.6	36.2	66.4	4.8
	$(39.2, 46.4)$	$(53.8, 77.4)$	$(32.7, 39.7)$	$(51.5, 81.4)$	$(3.4, 6.3)$
Local linear extr.	43.6	68.3	35.0	61.3	4.3
	$(39.1, 48.1)$	$(53.5, 83.0)$	$(31.0, 38.9)$	$(44.6, 78.0)$	$(2.9, 5.7)$

Notes: Cols. (1) and (3) report estimates of the average counterfactual outcome under the universal release policy, $E[Y(1) \mid R]$, separately for blacks and whites; cols. (2) and (4) report the estimates of the treatment effect for policy compliers, $E[Y(1) - Y(0) \mid D(Z, 0) = 0]$ implied by these estimates. Col. (5) reports estimates of the disparate impact parameter Δ. For specifications leading to set identification, identified set estimates are reported in brackets. 95% confidence intervals in parentheses. "Valid IV only" assumes only IV validity (eq. (2)). Linear, quadratic and local linear extrapolation correspond to the parametric extrapolation methods considered in ADH22.

Table 1: Estimates for universal release policy and the disparate impact parameter using NYC bail judge data from Arnold et al. (2022).

defendants will have similar or higher crime rates to those released under the status quo. This is intuitive, since we expect judges to be releasing defendants who they deem to be the lowest risk. The confidence intervals are also reasonably tight, allowing us to conclude that a universal release program will lead to a substantial increase in the misconduct rates relative to the status quo: for blacks, they imply that the misconduct rate will increase by at least 6.3 percentage points, which is a substantial increase relative to the status quo rate, 22.9%.

We also compute an estimate of the identified set and confidence intervals for the disparate impact parameter Δ. Comparing the resulting estimates to those from the parametric extrapolation methods, we see in Table 1 that our estimate of the identified set, 4.0–8.2%, is somewhat wider than the range of estimates we obtain based on the parametric extrapolations, 4.3–5.4%. Our 95% CIs are also informative: they equal $(1.0, 9.6)$%, so we can reject the null hypothesis of no disparate impact. Thus, the conclusion in ADH22 that $\Delta > 0$ is robust to dropping their functional form assumptions.[13]

Results for the quota policy. Although universal release programs have been implemented in some settings (Albright, 2022), it may often be more realistic to implement a policy that encourages judges to release more defendants but does not fully remove judicial discretion. For example, beginning in 2016, New York City began encouraging judges to release some defendants through a supervised release program rather than requiring them to post bail (Skemer et al., 2020). Likewise, Kentucky passed a law in 2011 that, among other things, changed the presumptive default from monetary bail to non-monetary release, yet provided judges with the discretion to override the default (Stevenson, 2018). Both policies were found to increase the fraction of defendants who were released, but did not increase this rate to one. As a stylized evaluation of such a counterfactual encouragement policy, we evaluate a counterfactual quota policy that asks the bottom 90% of judges in terms of leniency to increase their release rates to match the average release rate among the top 10% of judge under the status quo. This policy may also be viewed as a reasonable approximation to other types of policies that nudge or encourage the bottom 90% of judges to behave more similarly to the most lenient decile.

To benchmark the effect of this policy, we first compute the policy effect of a simpler policy that achieves the same release rates: reallocating all cases to the most lenient decile. Here the policy effect is point identified under random assignment of the instrument, and it is given by the difference between the current average misconduct rate versus the average misconduct

[13] In a robustness check, ADH22 construct non-sharp bounds on $E[Y(1) \mid R]$ and Δ using the aggregate mean outcome among released defendants, rather than constructing the sharp bounds by intersecting the judge-specific intervals. This yields wider bounds than what we obtain. For example, the plug-in bounds on Δ from this approach are -1% to 10%.

	no PC bound		PC bounds	
	Policy effect	PC TE	Policy effect	PC TE
Specification	(1)	(2)	(3)	(4)
A: Blacks				
Reallocation	5.6	53.2		
	(5.2, 6.1)	(49.2, 57.3)		
Constant treatment	5.2	49.7		
	(5.0, 5.5)	(47.4, 52.0)		
Valid IV only	[0.0, 10.5]	[0.0, 100.0]	[3.4, 7.0]	[32.3, 66.5]
	(0.0, 10.7)	(0.0, 100.0)	(1.8, 10.7)	(16.4, 100.0)
$\overline{DP} = 0.025$	[4.7, 6.1]	[44.9, 58.0]	[4.9, 5.9]	[46.7, 56.3]
	(2.8, 7.9)	(26.3, 76.7)	(2.8, 7.9)	(26.3, 76.7)
B: Whites				
Reallocation	3.8	56.6		
	(3.3, 4.2)	(49.7, 63.5)		
Constant treatment	3.2	48.1		
	(3.0, 3.5)	(44.5, 51.6)		
Valid IV only	[0.0, 6.7]	[0.0, 100.0]	[1.8, 5.2]	[26.2, 77.2]
	(0.0, 6.9)	(0.0, 100.0)	(0.7, 6.9)	(9.9, 100.0)
$\overline{DP} = 0.021$	[2.4, 5.1]	[35.7, 76.9]	[2.5, 5.0]	[37.7, 75.2]
	(1.2, 5.9)	(17.0, 91.1)	(1.2, 5.9)	(17.0, 91.1)

Notes: Cols. (1) and (3) report estimates of the policy effect $E[Y(D(Z,1)) - Y]$ for a quota policy that asks the bottom 90% of judges to match the release rate of the most lenient decile (which equals 80.2% for blacks and 83.2% for whites). Cols. (2) and (4) report the treatment effect for policy compliers implied by the policy effect estimates in cols. (1) and (3). Cols. (1) and (2) report estimates without imposing the PC bound in eq. (12), while cols. (3) and (4) impose it. 95% confidence intervals in parentheses; these do not account for covariate adjustment. For specifications leading to set identification, identified set estimates are reported in brackets.

Table 2: Estimates for quota policy using NYC bail judge data from Arnold et al. (2022).

rate for individuals assigned to judges in the most lenient decile.[14] We also consider a simple parametric benchmark that assumes homogeneous treatment effects, which also leads to point identification: the policy effect is identified by the TSLS estimate, multiplied by the change in the release rate relative to the status quo. As shown in Table 2, both benchmarks imply a similar policy effect, equal to about 5% for blacks and 3–4% for whites.

We then estimate the policy effect under a conservative specification that assumes only IV validity and policy monotonicity. Unlike for the universal release policy, the bounds for this policy are trivial in the sense that they imply the trivial bounds $[0, 100]$ for the policy complier treatment effect. The reason for this is that we cannot rule out that there is a sufficient mass of instrument never-takers under the status quo (those with $D(z, 0) = 0$ for all z) for whom we only have trivial bounds on $Y(1)$, and without further restrictions, policy compliers may be drawn from this pool in an adversarial way.

However, we can tighten these bounds by imposing two additional sets of restrictions that exploit the institutional details. First, we can impose the aggregate disagreement probability bound in eq. (11). To calibrate $\overline{\mathsf{DP}}$, we assume that the judges make release decisions according to a Gaussian signal model similar to that considered in Chan et al. (2022). The judges observe correlated signals U_z, releasing the individual if the signal is high enough (see Appendix B.2 for details). To calibrate the correlation matrix, we use the data from Sigstad (2026) on panels of judges ruling on criminal cases in the São Paulo Appeal Court. We focus on the first three judges to rule in each case.[15] In this data, the implied signal correlation between a randomly selected judge in the bottom 90% and one in the most lenient decile is 0.989.[16] We set the correlation between any pair of signals U_z and $U_{z'}$ where z is a judge in the bottom 90% and z' is in the top decile to this value in our Gaussian signal model, and set the judges' cutoffs for release to match those under our quota counterfactual. We calculate the value of $\overline{\mathsf{DP}}$ as the average disagreement probability in this model, which yields $\overline{\mathsf{DP}} = 0.025$ blacks and 0.021 for whites.

Second, the sole legal objective of bail judges is to allow most defendants to be released before trial while minimizing the risk of pretrial misconduct. In line with this objective, we can impose the policy complier bounds given by eq. (12) for each judge z in the bottom 90%.

[14] We note that for analyzing the impact of such a reallocation policy, we do not actually require IV exclusion.

[15] At least three judges rule on each case. Additional judges may provide an opinion in some cases, but only if the initial three judges disagree. To estimate disagreement rates, we therefore focus on the first three judges to avoid selection bias related to the initial three judges' decisions.

[16] The raw disagreement rates are as follows: for a randomly-selected pair of judges A and B with judge A in the top decile of leniency and judge B in the bottom nine deciles in Sigstad's data, we have $P(D_A = 1, D_B = 0) = 3.8\%$ and $P(D_A = 1, D_B = 0) = 0.6\%$, where D_A, D_B are the decisions of judges A, B, respectively.

Table 2 gives the results for the quota policy when we impose the disagreement probability bound in eq. (11) with this calibration. Confidence intervals are computed by projection, as detailed in Appendix B.2. Column (1) gives the results without other restrictions, while column (3) additionally imposes the bound on policy compliers in eq. (12).

With the $\overline{\mathsf{DP}}$ bound imposed, the identified set estimate is fairly tight, implying that the treatment effects for policy compliers are not too different from the TSLS estimates. The estimates of the bounds exceed the status quo misconduct rates among those currently released, $E[Y(1) \mid D(Z,1) = 1]$: these rates equal 32.9% for blacks and 26.9% for whites, while the lower bounds on the average value of $Y(1)$ for policy compliers, reported in column (2), equal 44.9 and 35.7, respectively. While the confidence intervals are wider, the lower endpoints are close to the status quo misconduct rates, particularly for blacks (26.3% vs. 32.9%), implying the policy compliers have similar or higher misconduct rates than the inframarginal individuals (those currently released). Additionally, imposing policy complier bounds helps tighten the point estimates, but not the confidence intervals once the $\overline{\mathsf{DP}}$ bound is imposed.

6.2 Suffolk County prosecutors

Background. We reanalyze the data from Agan et al. (2023, henceforth AHD23), who are interested in the effect of non-prosecution of nonviolent misdemeanor offenses on the subsequent criminal justice involvement of the defendants. Here the decision-makers (Z) are assistant district attorneys (ADAs), who decide whether to prosecute a case after an initial arraignment ($D = 1$) or drop it ($D = 0$). We then see whether a criminal complaint against the defendant has been filed within two years of arraignment ($Y \in \{0, 1\}$). ADH23 use the jackknife IV estimator to estimate the treatment effect of non-prosecution, and find robustly negative IV estimates. Based on this analysis, they conclude (ADH23, p. 1455):

> The results of our analysis imply that if all arraigning ADAs acted more like the most lenient ADAs in our sample when deciding which cases to prosecute, Suffolk County would likely see a reduction in criminal justice involvement for these nonviolent misdemeanor defendants.

If we assume homogeneous treatment effects, then the negative IV estimates do indeed imply that asking ADAs to be more lenient would result in lower criminal complaint probability for the defendants. We are interested in evaluating the robustness of this conclusion to dropping the homogeneous treatment effects assumption. Importantly, when ADH23 test IV monotonicity (Assumption 3) for each of the 9 courts in their dataset using the test developed by Frandsen et al. (2023), the test rejects IV monotonicity in three of the courts. Correspondingly, we conduct the analysis without imposing Assumption 3 or Assumption 4.

To formalize what it means to ask ADAs to act "more like the most lenient ADAs", we consider the same quota policy as in Section 6.1, whereby we impose a quota on the bottom 90% of ADAs to match the non-prosecution rate of the most lenient decile.

Data. The data comes from the Suffolk County District Attorney's Office in Massachusetts. We focus on the main analysis main sample in ADH23, which restricts the data to be between 2004 and 2008, drops felony charges, violent crimes, and prosecutors with fewer than 30 cases. This yields a sample with $67,060$ cases. ADH23 argue that conditional on court by time fixed effects, the ADAs are as-good-as-randomly assigned to cases, and they control for these fixed effects linearly in all their specifications. In their main specification, they also consider an extended set of controls that include case and defendant characteristics. Correspondingly, to estimate ADA-specific probabilities $P^*(Y(d) = y, D(z, 0) = d)$ we also linearly adjust for the court-by-time fixed effects and case and defendant characteristics. Since we have access to the microdata, our inference accounts for this covariate adjustment. As in the NYC bail judge application, we pool ADAs with fewer than 300 cases in the bottom 90% and those in the top decile of leniency into a single ADA to mitigate finite-sample issues. Table B.4 shows that our inference results remain the same when we do not pool. This leaves us with $K = 69$ distinct values for the instrument.

Summary statistics. Figure 3 plots the ADA-specific non-prosecution rates against the criminal complaint rates. Relative to Figure 1, there is much more variation in the outcomes for ADAs with similar non-prosecution rates. An implication of IV monotonicity is that ADAs with the same prosecution rate must have the same outcomes, since they must be prosecuting the same set of individuals; if the prosecution rates differ by x, the outcomes cannot differ by more than x times the range of the support for the outcome. The IV monotonicity test of Frandsen et al. (2023) tests precisely this implication. Thus, the large outcome variability in Figure 1 can be seen as giving visual evidence against IV monotonicity, in line with the rejection of the Frandsen et al. (2023) test in several subsets of the data.

Figure 3 also shows that the relationship between non-prosecution and criminal complaints is negative—the IV estimate correspondingly equals -28.8%, albeit the estimate is an order of magnitude noisier than that in ADH22, with standard error 10.0 (this replicates Table III, col. (4) in ADH23). The overall non-prosecution rate in the data is 20.4%, and the overall criminal complaint rate is 34.1%.

Results for the quota policy. To benchmark our estimates, we again start out by computing the policy effect of a reallocation policy that assigns all cases to the most lenient

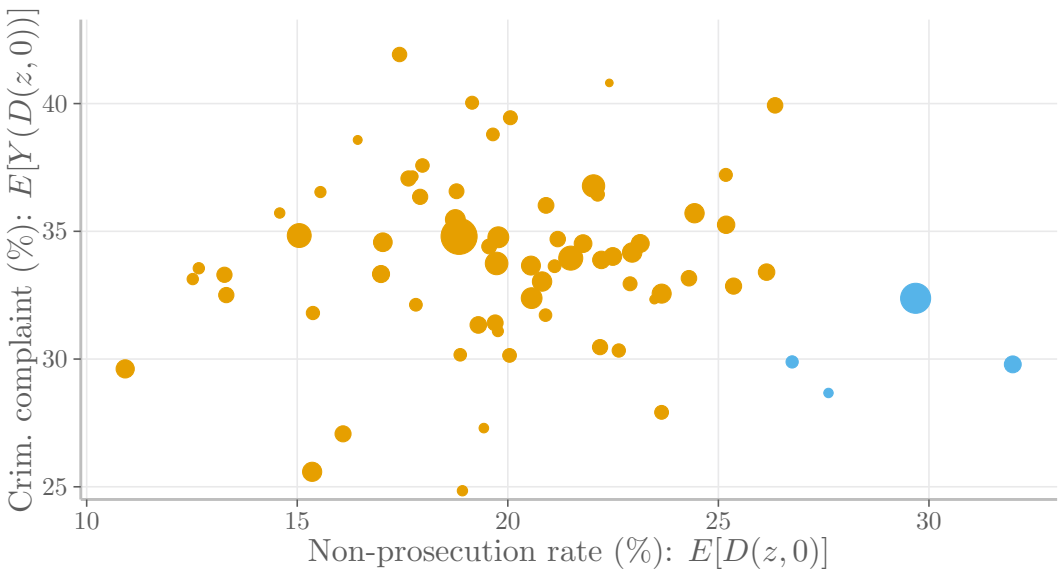

Figure 3: ADA-specific sample non-prosecution rates and criminal complaint rates, based on Suffolk County prosecutor data from Agan et al. (2023).

Notes: Both rates are adjusted for court-by-time fixed effects and case and defendant characteristics. Each dot corresponds to a single ADA; dot size scales with the number of cases handled by the ADA. Judges with fewer than 300 cases are pooled into a single dot. The top decile of most lenient ADAs is plotted in blue, the bottom 90% are plotted in orange.

ADAs, and a parametric benchmark that assumes homogeneous treatment effects, which allows for extrapolation of the IV estimates. As shown in Table 3, these imply that the quota policy would lead to a reduction in the probability of criminal complaints by about 1–3 percentage points. Notably, however, under the reallocation policy the confidence interval only marginally excludes zero.

We then estimate the policy effect under a specification that only assumes IV validity and policy monotonicity. Like the NYC bail judge data, the data is consistent with there being many instrument never-takers. The policy only moves treatment by 9.0 percentage points, so without further assumptions, we cannot rule out that the individuals marginally released—the policy compliers—are picked from the set of instrument never-takers in an adversarial way. As a result, the estimate of the identified set (row 3 of Table 3) implies trivial bounds for the policy complier treatment effect.

We next consider an additional assumption that can help us tighten these bounds. In contrast to our previous application, imposing bounds on the outcomes for policy compliers as in eq. (12) is hard to motivate in this setting, since prosecution decisions are based in large part on the strength of evidence against the defendant, which is not directly related to their recidivism risk. Likewise, we do not have external data from panels of prosecutors to

Specification	Policy effect (1)	PC TE (2)
Reallocation	-1.7	-19.0
	$(-3.3, -0.1)$	$(-36.7, -1.2)$
Constant treatment	-2.6	-28.8
	$(-4.4, -0.8)$	$(-48.5, -9.1)$
Valid IV only	$[-9.0, 9.0]$	$[-100.0, 100.0]$
	$(-9.3, 9.3)$	$(-100.0, 100.0)$
$\overline{OD} = 0.019$	$[-3.4, 0.0]$	$[-37.9, 0.0]$
	$(-5.3, 1.9)$	$(-61.2, 21.9)$

Notes: Col. (1) reports estimates of the policy effect $E[Y(D(Z,1)) - Y]$ for a quota policy that asks bottom 90% of ADAs to match the release rate of the most lenient decile (which equals 29.4%). Col. (2) reports the treatment effect for policy compliers implied by the policy effect estimates. 95% confidence intervals in parentheses. For specifications leading to set identification, identified set estimates reported in brackets.

Table 3: Estimates for quota policy using Suffolk County prosecutor data from Agan et al. (2023).

help calibrate disagreement probabilities across ADAs.

We therefore pursue a simple approach that directly restricts the counterfactual outcome disparity between ADAs in the bottom 90% and those in the top decile as in eq. (13) for some bound $\overline{\mathsf{OD}}$. Imposing $\overline{\mathsf{OD}} = 0$ amounts to assuming that the policy effect matches that under the reallocation policy. As a simple benchmark, we calibrate $\overline{\mathsf{OD}}$ to the status quo outcome disparity between the two groups of ADAs. Since the criminal complaint rate for those in the top decile is 1.9 percentage points lower than for those in the bottom 90%, we set $\overline{\mathsf{OD}} = 0.019$. This is motivated by the informal discussion in ADH23 who consider a counterfactual where the stricter ADAs are asked to act "more like the most lenient ADAs." If the policy directive asks them to act this way, one can reasonably expect the outcome disparity to be lower under the counterfactual, so that this value of $\overline{\mathsf{OD}}$ can be interpreted as a conservative bound. As Table 3 shows, imposing this bound substantially tightens the estimate of the identified set. As column (2) indicates, the implied treatment effect for policy compliers lies between 0 and about 130% of the IV estimate. The confidence intervals include small positive values.

There is an alternative view of this bound that is useful. In particular, setting $\overline{\mathsf{OD}}$ to the status quo disparity is the maximal value of $\overline{\mathsf{OD}}$ that one can allow for while still guaranteeing

identification of the sign of the policy effect.[17] Using eq. (14), we can thus ask if the implied disagreement probability is reasonable: if so, this suggests that identification of the sign of the treatment effect is possible under reasonable restrictions on the disagreement probability. If we make no restrictions on treatment effect heterogeneity, the implied value of $\overline{DP}$ is 0.032, which corresponds to a signal correlation equal to $\rho = 0.998$. This is substantially higher than the corresponding value calibrated from Sigstad (2026) (0.989), implying that prosecutors in Massachusetts would have to be much more strongly in agreement than the judges in Sigstad's data. While it is possible that agreement rates differ across contexts, this strikes us as a stringent requirement, especially given the rejection of IV monotonicity, which implies disagreement under the status quo. We can, however, learn about the sign of the policy impact under weaker restrictions on disagreement if we also impose some restrictions on treatment effect heterogeneity. If we impose that Δ_{TE} is at most twice the IV estimate, then we only require $\overline{DP} = 0.112$, which corresponds to $\rho = 0.972$ in the Gaussian signal model. This is lower than the value in Sigstad, and thus seems more reasonable.

[17] Since ADAs above the quota are assumed not to change their behavior, the policy effect may be written as the difference between status quo, and counterfactual outcome disparity, $\theta - E[Y] = n_{\mathcal{Q}}/n \cdot (E[Y(D(Z_{\mathcal{Q}}, 1)) - Y(D(Z_{\mathcal{Q}^c}, 1))] - (E[Y \mid Z \in \mathcal{Q}] - E[Y \mid Z \in \mathcal{Q}^c]))$, with $n_{\mathcal{Q}}$ denoting the number of cases assigned to the ADAs who are subject to the quota. Since the status quo outcome disparity is negative, assuming that the counterfactual outcome disparity as at most as big as that under the status quo implies that the policy effect cannot be positive.

Appendix A Proofs

A.1 Proof of Proposition 1

Proposition 5 below gives a more general version of Proposition 1 that does not restrict the distribution of the potential outcomes. Stating this result requires setting up additional notation. The model primitives $(Y(0), Y(1), D(\cdot,\cdot), Z)$ take values in the space $\mathbb{R}^{2+2K+1}$. We endow this space with its standard Borel σ-algebra. We consider general Borel probability measures for the model primitives with support $\mathcal{Y}^2 \times \{0,1\}^{2K} \times \mathcal{Z}$, where $\mathcal{Y} \subseteq \mathbb{R}$. Proposition 10.2.8 in Dudley (2002) implies that any joint probability distribution over the model's primitives has well-defined conditional distributions, for any subset of conditioning variables. We use this fact repeatedly in what follows.

For each $z \in \mathcal{Z}$, let Π_z^* denote the marginal distribution of the random vector $(Y(0), Y(1), D(z,0), D(z,1))$ under P^*. It is without loss of generality to assume these marginal distributions are all dominated by some probability measure, μ (for example, set $\mu := (1/K)\sum_{k=1}^{K} \Pi_k^*$). We denote the densities of Π_z^* with respect to μ by π_z^*. We refer to the collection of densities $\{\pi_z^*\}_{z\in\mathcal{Z}}$ as the *marginal z-densities* of P^*.

The dominating measure μ can be thought of as the law of some random vector (Y_0, Y_1, D_0, D_1). Let $\mu_{D_0,D_1|Y_0=y_0,Y_1=y_1}$ denote the conditional distribution of μ given $(Y_0, Y_1) = (y_0, y_1)$, which we abbreviate to $\mu_{D_0,D_1|y_0,y_1}$ whenever it doesn't cause confusion. We denote the marginal distribution of (Y_0, Y_1) under the probability measure μ by μ_{Y_0,Y_1}. We follow the same convention to denote other conditional and marginal distributions.

The observable data given $Z = z$, $(Y(D(z,0)), D(z,0))$, take values in the sample space $\mathcal{Y} \times \{0,1\}$. Thinking of μ as a probability, the implied measure on this space is given by $\tilde{\mu}(F \times G) = \mathrm{I}\{1 \in G\}\mu_{Y_1|D_0=1}(F)\mu_{D_0}(\{1\}) + \mathrm{I}\{0 \in G\}\mu_{Y_0|D_0=0}(F)\mu_{D_0}(\{0\})$. Let $P_{Y,D|z}$ denote the conditional distribution of P given $Z = z$.

Proposition 5. *There exists $P^* \in \mathcal{P}_I^*(P; \mathcal{P}^*_{\text{valid}})$ with marginal z-densities $\{\pi_z\}_{z\in\mathcal{Z}}$ (with respect to some probability measure μ) if and only if $\{\pi_z\}_{z\in\mathcal{Z}}$ satisfy the following conditions:*

1. *They match the observable data: for every $z \in \mathcal{Z}$, $P_{Y,D|z}$ is absolutely continuous with respect to $\tilde{\mu}$ with density $p_{Y,D|z}$ given $\tilde{\mu}$-a.e. by*

$$p_{Y,D|z}(y,1) = \int_{\mathcal{Y}\times\{0,1\}} \pi_z(y_0, y, 1, d_1)\, d\mu_{Y_0,D_1|Y_1=y,D_0=1},$$
$$p_{Y,D|z}(y,0) = \int_{\mathcal{Y}\times\{0,1\}} \pi_z(y, y_1, 0, d_1)\, d\mu_{Y_1,D_1|Y_0=y,D_0=0}.$$

2. *They imply the same distribution for* $(Y(0), Y(1))$*: for all* $z, z' \in \mathcal{Z}$*, the equality*

$$\int_{\{0,1\}^2} \pi_z(y_0, y_1, d_0, d_1)\, d\mu_{D_0,D_1|y_0,y_1} = \int_{\{0,1\}^2} \pi_{z'}(y_0, y_1, d_0, d_1)\, d\mu_{D_0,D_1|y_0,y_1}$$

holds μ_{Y_0,Y_1}*-a.e.*

3. *They are valid density functions: for all* $z \in \mathcal{Z}$*,* $\pi_z(y_0, y_1, d_0, d_1) \geq 0$ *(*μ*-a.e.) and* $\int_{\mathcal{Y}^2\times\{0,1\}^2} \pi_z(y_0, y_1, d_0, d_1)\, d\mu = 1$*.*

Proof. We first show the $\implies$ part of the statement. Namely, if there exists $P^* \in \mathcal{P}_I^*(P; \mathcal{P}^*)$ with marginal z-densities $\{\pi_z\}_{z\in\mathcal{Z}}$, then conditions 1–3 hold. This is conceptually straightforward, and most of the work just involves translating statements about probabilities to statements about densities.

To establish condition 1, observe that by definition of the density $p_{Y,D|z}$, for any measurable $F \subseteq \mathcal{Y}$,

$$P_{Y,D|z}(F \times \{1\}) = \int_{F\times\{1\}} p_{Y,D|z}(y, 1)\, d\tilde{\mu}.$$

Further, observe that

$$\begin{aligned}
P_{Y,D|z}(F \times \{1\}) &= P^*(Y(1) \in F, D(z, 0) = 1) \\
&= \int_{\mathcal{Y}\times F\times\{0,1\}\times\{1\}} \pi_z(y_0, y_1, d_0, d_1)\, d\mu \\
&= \int_{\{1\}} \int_F \int_{\mathcal{Y}\times\{0,1\}} \pi_z(y_0, y_1, d_0, d_1)\, d\mu_{Y_0,D_1|y_1,d_0} d\mu_{Y_1|D_0=d_0} d\mu_{D_0} \\
&= \int_F \left[\int_{\mathcal{Y}\times\{0,1\}} \pi_z(y_0, y_1, 1, d_1)\, d\mu_{Y_0,D_1|Y_1=y_1,D_0=1} \right] d\mu_{Y_1|D_0=1} \cdot \mu_{D_0}(\{1\}) \\
&= \int_{F\times\{1\}} \left[\int_{\mathcal{Y}\times\{0,1\}} \pi_z(y_0, y_1, 1, d_1)\, d\mu_{Y_0,D_1|Y_1=y_1,D_0=1} \right] d\tilde{\mu}_{Y,D}
\end{aligned}$$

where the first line uses the fact that since $P^* \in \mathcal{P}_I^*(P; \mathcal{P}^*)$, it generates P and satisfies eq. (2), the second line uses the definition of the density π_z, the third line uses law of iterated expectations (e.g. Part II of Theorem 10.2.1 in Dudley (2002)), the fourth line follows by algebraic manipulation, and last line uses the definition of $\tilde{\mu}$. Since the last two displays hold for all measurable F, it follows that $p_{Y,D|z}(y, 1) = \int_{\mathcal{Y}\times\{0,1\}} \pi_z(y_0, y_1, 1, d_1)\, d\mu_{Y_0,D_1|Y_1=y_1,D_0=1}$ ($\tilde{\mu}$-a.e.), which gives the first expression in condition 1. The second expression in condition 1 can be verified by using an analogous argument to show that

$$P_{Y,D|z}(F \times \{0\}) = \int_{F\times\{0\}} p_{Y,D|z}(y, 0)\, d\tilde{\mu} = \int_{\mathcal{Y}\times\{0,1\}} \pi_z(y_0, y_1, 0, d_1)\, d\mu_{Y_1,D_1|Y_0=y_0,D_0=0}.$$

To establish condition 2, fix any $z \in \mathcal{Z}$. Then for any measurable sets $F_0, F_1 \subseteq \mathcal{Y}$:

$$\begin{aligned} P^*(Y(0) \in F_0, Y(1) \in F_1) &= \int_{F_0 \times F_1 \times \{0,1\}^2} \pi_z(y_0, y_1, d_0, d_1)\, d\mu \\ &= \int_{F_0 \times F_1} \int_{\{0,1\}^2} \pi_z(y_0, y_1, d_0, d_1)\, d\mu_{D_0,D_1|y_0,y_1}\, d\mu_{Y_0,Y_1}, \end{aligned}$$

where the first equality follows by definition of the density π_z, and the second by the law of iterated expectations. Thus, for any measurable $F_0, F_1 \subseteq \mathcal{Y}$ and any pair $z, z' \in \mathcal{Z}$,

$$\begin{aligned} &\int_{F_0 \times F_1} \int_{\{0,1\}^2} \pi_z(y_0, y_1, d_0, d_1)\, d\mu_{D_0,D_1|y_0,y_1}\, d\mu_{Y_0,Y_1} \\ &\qquad = \int_{F_0 \times F_1} \int_{\{0,1\}^2} \pi_{z'}(y_0, y_1, d_0, d_1)\, d\mu_{D_0,D_1|y_0,y_1}\, d\mu_{Y_0,Y_1}, \end{aligned}$$

which implies condition 2. Finally, condition 3 is immediate from the definition of a marginal z-density.

We now show the $\Longleftarrow$ part of the statement. That is, given $\{\pi_z\}_{z \in \mathcal{Z}}$ that satisfy conditions 1–3, we need to show that we can build a distribution P^* with marginal z-densities $\{\pi_z\}_{z \in \mathcal{Z}}$ that generates P. The key step is the observation that since the marginals imply the same distribution of $(Y(0), Y(1))$, to couple them together to form a joint distribution P^*, we just need to construct a joint distribution for $D(\cdot, \cdot)$ given $Y(0), Y(1)$. To this end, we show this can be done by assuming that conditional on $(Y(0), Y(1))$, the treatment decisions $D(z, 0), D(z, 1)$ can be assumed to be independent across judges z, with the distribution of $D(z, 0), D(z, 1)$ set to match that implied by the marginal π_z.

Formally, to construct P^*, for any measurable $F_0, F_1 \subseteq \mathcal{Y}$, $G = (G_0^1 \times G_1^1 \ldots G_0^K \times G_1^K) \subseteq \{0, 1\}^{2K}$, and $H \subseteq \mathcal{Z}$ we set

$$\begin{aligned} &P^*\left(Y(0) \in F_0, Y(1) \in F_1, D(\cdot, \cdot) \in G, Z \in H\right) \\ &\qquad := P^*\left(Y(0) \in F_0, Y(1) \in F_1, D(\cdot, \cdot) \in G\right) \cdot P(Z \in H), \end{aligned}$$

set the conditional distribution of P^* given $(Y(0), Y(1)) = (y_0, y_1)$ to

$$P^*_{D(\cdot,\cdot)|y_0,y_1} := \Pi_{D(1,\cdot)|y_0,y_1}\left(G_0^1 \times G_1^1\right) \cdots \Pi_{D(K,\cdot)|y_0,y_1}\left(G_0^K \times G_1^K\right), \tag{15}$$

where $\Pi_{D(z,\cdot)|y_0,y_1}$ is the conditional distribution of $(D(z, 0), D(z, 1))$ given (y_0, y_1) implied by

the marginal z-density π_z, and set the marginal distribution of P^* over $(Y(0), Y(1))$ to

$$P^*(Y(0) \in F_0, Y(1) \in F_1) := \int_{F_0 \times F_1} \int_{\{0,1\}^2} \pi_z(y_0, y_1, d_0, d_1)\, d\mu_{D_0,D_1|y_0,y_1}\, d\mu_{Y_0,Y_1}. \tag{16}$$

By conditions 2 and 3, this marginal distribution is well-defined and doesn't depend on z.

Observe that by construction, P^* satisfies eq. (2). It thus remains to show that P^* generates P. Given that the marginals over Z match by construction, it suffices to show that for any $z \in \mathcal{Z}$, and $F \subseteq \mathcal{Y}$,

$$P_{Y,D|z}(Y \in F, D = 1) = P^*\left(Y(1) \in F, D(z,0) = 1\right) \tag{17}$$

and

$$P_{Y,D|z}(Y \in F, D = 0) = P^*\left(Y(0) \in F, D(z,0) = 0\right). \tag{18}$$

Let Π_z denote the distribution implied by the densities π_z. Equation (17) follows from the following set of equalities:

$$\begin{aligned}
P_{Y,D|z}(Y \in F, D = 1) &= \int_F \int_{\mathcal{Y}\times\{0,1\}} \pi_z(y_0, y, 1, d_1)\, d\mu_{Y_0,D_1|Y_1=y,D_0=1}\, d\mu_{Y_1|D_0=1}\mu_{D_0}(\{1\}) \\
&= \int_{\mathcal{Y}\times F\times\{1\}\times\{0,1\}} \pi_z(y_0, y_1, d_0, d_1)\, d\mu = \Pi_z(F \times \mathcal{Y} \times \{1\} \times \{0,1\}) \\
&= \int_{\mathcal{Y}\times F} \int_{\{1\}\times\{0,1\}} d\Pi_{D(z,\cdot)|y_0,y_1}\, dP^*_{Y(0),Y(1)} \\
&= \int_{\mathcal{Y}\times F} \int_{\{1\}\times\{0,1\}} dP^*_{D(z,0),D(z,1)|y_0,y_1}\, dP^*_{Y(0),Y(1)} \\
&= P^*(Y(1) \in F, D(z,0) = 1).
\end{aligned}$$

where the first equality uses the fact that $P_{Y,D|z}$ has density $p_{Y,D|z}$ and condition 1, the second equality uses iterated expectations, the third equality uses iterated expectations and the fact that by definition of P^* in eq. (16), its marginal distribution over $(Y(0), Y(1))$ matches Π_z, the fourth line uses the fact that by eq. (15), the conditional distribution $P^*_{D(z,0),D(z,1)|y_0,y_1}$ matches $\Pi_{D(z,\cdot)|y_0,y_1}$, and the last line follows by iterated expectations. Analogous arguments establish eq. (18). □

A.2 Proof of Corollary 1

The fact that the identified set is an interval follows by convexity of the optimization problem. To show that eq. (3) gives the upper endpoint of the interval, note that, by definition of the supremum, there exists a sequence of distributions $P^*_n \in \mathcal{P}^*$ such that

$\sup_{P^* \in \mathcal{P}^*} E_{P^*}[Y(D(Z,1))] = \lim_{n\to\infty} E_{P_n^*}[Y(D(Z,1))]$. Let v_0 be the value function of the optimization problem in (3). Since $P_n^* \in \mathcal{P}^*$, the marginals of P_n^*, call them $\{\pi_z^n(\cdot)\}_z$, are feasible in the optimization problem (3) by Proposition 1. It follows that $\sup_{P^* \in \mathcal{P}^*} E_{P^*}[Y(D(Z,1))] \leq v_0$. To show that $\sup_{P^* \in \mathcal{P}^*} E_{P^*}[Y(D(Z,1))] \geq v_0$, suppose that $\{\widetilde{\pi}_z^n(\cdot)\}_z$ is a sequence in $\mathcal{R}^*$ such that the value of (3) is given by the limit

$$\lim_{n\to\infty} \sum_{z\in\mathcal{Z}} P(Z=z) \sum_{y_0,y_1,d_0,d_1 \in \mathcal{Y}^2\times\{0,1\}^2} (d_1 y_1 + (1-d_1)y_0) \cdot \widetilde{\pi}_z^n(y_0,y_1,d_0,d_1).$$

Then, since $\{\widetilde{\pi}_z^n(\cdot)\}_z$ satisfy the constraints 1–3 in Proposition 1, there exists some sequence $\tilde{P}_n^* \in \mathcal{P}^*_{\text{valid}}$ with marginals $\{\widetilde{\pi}_z^n(\cdot)\}_z$. Furthermore, since $\{\widetilde{\pi}_z^n(\cdot)\}_z \in \mathcal{R}^*$, it follows that $\tilde{P}_n^* \in \mathcal{P}^*$. Thus, $\sup_{P^* \in \mathcal{P}^*} E_{P^*}[Y(D(Z,1))] \geq v_0$.

A.3 Proof of Proposition 2

Let $\alpha_0 = P(D=1)$ denote the status quo release rate. If $\alpha = \alpha_0$, then θ is point-identified by $E_P[Y]$; assume therefore that $\alpha > \alpha_0$. Since $Y(0) = 0$ and $D(Z,1) \geq D(Z,0)$ by policy monotonicity, it follows by iterated expectations that

$$\begin{aligned} E_{P^*}[Y(D(Z,1))] &= E_{P^*}[Y(1) \mid D(Z,1) > D(Z,0)](\alpha - \alpha_0) + E_{P^*}[Y(1) \mid D(Z,0) = 1]\alpha_0 \\ &= E_{P^*}[Y(1) \mid D(Z,1) > D(Z,0)](\alpha - \alpha_0) + E_P[Y \mid D = 1]\alpha_0, \end{aligned}$$

where we use $P^*(D(Z,1) > D(Z,0)) = \alpha - \alpha_0$, and $P^*(D(Z,0)) = \alpha_0$. Apart from the average treated outcome for policy compliers, $E_{P^*}[Y(1) \mid D(Z,1) > D(Z,0)]$, all other quantities are known. It therefore suffices to show that the identified set for $E_{P^*}[Y(1) \mid D(Z,1) > D(Z,0)]$ does not depend on IV monotonicity.

Lemma 1 in Appendix A.7 derives lower and upper bounds on the cdf F_{P^*} of $Y(1) \mid D(Z,1) > D(Z,0)$ over $P^* \in \mathcal{P}_I^*(P, \mathcal{P}_\alpha^*)$ of the form $F^{lb} \leq F_{P^*} \leq F^{ub}$. Lemma 2 shows that if the set $\mathcal{P}_I^*(P, \mathcal{P}_{Mon,\alpha}^*)$ is not empty, there exist probabilities $P^{ub}, P^{lb} \in \mathcal{P}_I^*(P, \mathcal{P}_{Mon,\alpha}^*)$ that generate the cdfs F^{lb} and F^{ub}. Since first-order stochastic dominance, $P(Y > t) \leq Q(Y > t)$, implies $E_P[Y] \leq E_Q[Y]$ (e.g. Marshall et al., 2011, Proposition A.2, p. 694), it follows by convexity of $\mathcal{P}_I^*(P, \mathcal{P}_\alpha^*)$ that whenever this set is not empty, the identified set for θ is given by an interval, with endpoints given by the expected value of the distributions F^{ub} and F^{lb}, respectively, whether or not IV monotonicity is imposed.

A.4 Proof of Proposition 3

Let $\alpha_{\max} = E[D(z_{\max}, 0)]$, where, by definition, $z_{\max} \in \text{argmax}_{z \in \mathcal{Z}} E[D(z, 0)]$. By iterated expectations, for any $P \in \mathcal{P}_I^*(P; \mathcal{P}_\alpha^*)$,

$$
\begin{aligned}
&E_{P^*}[Y(D(Z,1))] \\
&\quad = \alpha_{\max} E_{P^*}[Y(D(Z,1)) \mid D(z_{\max},0) = 1] + (1 - \alpha_{\max}) E_{P^*}[Y(D(Z,1)) \mid D(z_{\max},0) = 0] \\
&\qquad = \alpha_{\max} E_{P^*}[Y(1) \mid D(z_{\max},0) = 1] + (1 - \alpha_{\max}) E_{P^*}[Y(D(Z,1)) \mid D(z_{\max},0) = 0],
\end{aligned}
$$

where the second inequality follows by the definition of the sufficiently strong encouragement condition (iii). Since the first term is point identified, $E_{P^*}[Y(1) \mid D(z_{\max}, 0) = 1] = E_P[Y \mid D = 1, Z = z_{\max}]$, if $\alpha_{\max} = 1$, then the proposition holds trivially. Thus, we focus on the case $\alpha_{\max} < 1$. We will show that the identified set for the conditional distribution $Y(1), Y(0), D(Z, 1) \mid D(z_{\max}, 0) = 0$ doesn't depend on whether IV monotonicity (Assumption 3) is imposed. Since $E_{P^*}[Y(D(Z, 1)) \mid D(z_{\max}, 0) = 0]$ is a functional of this distribution, the result will then follow.

The distribution of $D(Z, 1) \mid D(z_{\max}, 0) = 0$ is Bernoulli with parameter $(\alpha - \alpha_{\max})/(1 - \alpha_{\max}) \geq 0$, regardless of whether we impose IV monotonicity. This follows from the fact that

$$
\begin{aligned}
\alpha &= P^*(D(Z,1) = 1) \\
&= P^*(D(Z,1) = 1, D(z_{\max,0}) = 1) + P^*(D(Z,1) = 1, D(z_{\max,0}) = 0) \\
&= \alpha_{\max} + P^*(D(Z,1) = 1, D(z_{\max,0}) = 0),
\end{aligned}
$$

where the last equality holds because P^* satisfies the strong encouragement condition.

Moreover, since the distribution of the observable data does not depend on $D(Z, 1)$, any joint distribution for $Y(1), Y(0), D(Z, 1) \mid D(z_{\max}, 0) = 0$ is in the identified set if the implied marginal for $Y(1), Y(0) \mid D(z_{\max}, 0) = 0$ is in the identified set and $P(D(Z, 1) = 1 \mid D(z_{\max}) = 0)$ is Bernoulli with parameter $(\alpha - \alpha_{\max})/(1 - \alpha_{\max})$.

Thus, it is sufficient to show that the identified set for the conditional distribution $Y(1), Y(0) \mid D(z_{\max}, 0) = 0$ does not depend on IV monotonicity. To this end, note that $Y(0) \mid D(z_{\max}, 0) = 0$ is identified by the distribution of $Y \mid D = 0, Z = z_{\max}$. Thus, without imposing IV monotonicity, the identified set for $(Y(1), Y(0)) \mid D(z_{\max}, 0) = 0$ is a weak subset of the possible joint distributions that match the identified marginal for $Y(0) \mid D(z_{\max}, 0) = 0$ and the constraint that $Y(1) \in \mathcal{Y}$. However, any joint distribution in this set is achievable under IV monotonicity, since IV monotonicity implies that the observed data do not depend on $Y(1) \mid D(z_{\max}, 0) = 0$, and thus any choice for $P(Y(1) \mid Y(0), D(z_{\max}, 0) = 0)$ is consistent with the observable data.

A.5 Proof of Proposition 4

We first show the $\implies$ direction that if a distribution $P^* \in \mathcal{P}_I^*(P; \mathcal{P}_{\text{DB}}^*)$ generates the collection of marginal probability mass functions $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$, then these marginals satisfy conditions 1–4. Since $\mathcal{P}_{\text{DB}}^* \subseteq \mathcal{P}_{\text{valid}}^*$, Proposition 1 implies that the collection $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$ satisfies conditions 1–3. Hence, it only remains to verify condition 4. To this end, note that (5) is equivalent to

$$\begin{aligned} P^*(D(z,a)=1, D(z',a')=1) &\geq (1-\delta_{z,z',a,a'})P^*(D(z',a')=1) \\ &= (1-\delta_{z,z',a,a'})\pi(D(z',a')=1). \end{aligned}$$

By the law of total probability, the left-hand side equals

$$\sum_{y_1,y_0} P^*(Y(0)=y_0, Y(1)=y_1, D(z,a)=1, D(z',a')=1) \leq \sum_{y_1,y_0} \min\{\pi(y_0,y_1,D(z,a)=1), \pi(y_0,y_1,D(z',a')=1)\},$$

where the inequality uses the Fréchet–Hoeffding bound, and the identity $\pi(y_0, y_1, D(z,a) = 1) = P^*(Y(0) = y_0, Y(1) = y_1, D(z,a) = 1)$. Combining the preceding two displays then yields the condition 4.

Now we show the $\impliedby$ direction that if a collection of marginals $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$ satisfies conditions 1–4 and is consistent with Assumption 2, then there exists a distribution $P^* \in \mathcal{P}_I^*(P; \mathcal{P}_{\text{DB}}^*)$ that generates $\{\pi_z(\cdot)\}_{z\in\mathcal{Z}}$. At a high-level, the specified $\{\pi_z\}_z$ pin down the marginal distributions of $\{Y(\cdot), D(z,\cdot)\}_z$, so what remains is to specify a coupling of these marginals under P^* to match the disagreement bound. The key idea is to construct $D(\cdot,\cdot)$ via a latent threshold crossing model conditional on the potential outcomes, so that $D(z,a) = \mathrm{I}\{V_{y_1,y_0} \leq \alpha_z(a,y_1,y_0)\}$, where V_{y_1,y_0} is a common ranking conditional on $Y(1) = y_1, Y(0) = y_0$. In other words, the judges under this DGP agree on the ranking of defendants within groups of people with the same potential outcomes, but possibly disagree on the potential outcome-specific release cutoffs.

Formally, to construct P^*, we set

$$\begin{aligned} &P^*\left(Y(0)=y_0, Y(1)=y_1, \{D(z,0)=d_{z0}, D(z,1)=d_{z1}\}_{z\in\mathcal{Z}}, Z=z\right) \\ &\qquad := P^*\left(Y(0)=y_0, Y(1)=y_1, \{D(z,0)=d_{z0}, D(z,1)=d_{z1}\}_{z\in\mathcal{Z}}\right) \cdot P\left(Z=z\right), \end{aligned}$$

set the marginal distribution of P^* over $(Y(0), Y(1))$ to equal the distribution of $(Y(0), Y(1))$

implied by the marginals π_z,

$$P^*(Y(0)=y_0, Y(1)=y_1) := \pi(y_0,y_1), \quad \pi(y_0,y_1) := \sum_{d_0,d_1\in\{0,1\}^2} \pi_z(y_0,y_1,d_0,d_1),$$

(by condition 2, $\pi(y_0,y_1)$ doesn't depend on z), and finally, to define the conditional distribution of $\{D(z,0),D(z,1)\}_{z\in\mathcal{Z}}$ given $(Y(0),Y(1))=(y_0,y_1)$ we set

$$D(z,a) = \mathrm{I}\{V_{y_0,y_1} \le \alpha_z(a,y_0,y_1)\}, \quad \alpha_z(a,y_0,y_1) = \frac{\pi_z(y_0,y_1,1,1)}{\pi(y_0,y_1)} + a\frac{\pi_z(y_0,y_1,0,1)}{\pi(y_0,y_1)}, \quad (19)$$

where V_{y_0,y_1} is uniform on $[0,1]$ conditional on $Y(1)=y_1, Y(0)=y_0$.

This construction implies that eq. (2) holds. Further, since $\alpha_z(1,y_0,y_1) \ge \alpha_z(0,y_0,y_1)$ by construction, this implies that $P^*(D(z,0)=1, D(z,1)=0 \mid Y(0)=y_0, Y(1)=y_1)=0$, so that Assumption 2 holds. The construction also implies that

$$P^*(D(z,0)=0, D(z,1)=1 \mid Y(0)=y_0, Y(1)=y_1) = \frac{\pi_z(y_0,y_1,0,1)}{\pi(y_0,y_1)},$$
$$P^*(D(z,0)=1, D(z,1)=1 \mid Y(0)=y_0, Y(1)=y_1) = \frac{\pi_z(y_0,y_1,1,1)}{\pi(y_0,y_1)},$$

so that P^* generates the marginals $\{\pi_z\}$. Moreover, condition 1 implies that P^* generates P. It remains to show that P^* satisfies the disagreement bounds in (5). It follows from the definition of $D(z,a)$ in eq. (19) that

$$P^*(D(z,a)=1, D(z',a')=1 \mid Y(0)=y_0, Y(0)=y_1) = \min\{\alpha_z(a,y_0,y_1), \alpha_{z'}(a',y_0,y_1)\}.$$

Hence, unconditionally,

$$P^*(D(z,a)=1, D(z',a')=1) = \sum_{y_0,y_1} \min\{\alpha_z(a,y_0,y_1), \alpha_{z'}(a',y_0,y_1)\}\pi(y_0,y_1)$$
$$= \sum_{(y_0,y_1)\in\mathcal{Y}^2} \min\left\{\pi_z(y_0,y_1,1,1) + a\pi_z(y_0,y_1,0,1), \pi_{z'}(y_0,y_1,1,1) + a'\pi_{z'}(y_0,y_1,0,1)\right\}. \quad (20)$$

We now claim that for any $(z,a)\in\mathcal{Z}\times\{0,1\}$,

$$\pi_z(y_0,y_1,1,1) + a\pi_z(y_0,y_1,0,1) = \sum_{d_a=1, d_{1-a}\in\{0,1\}} \pi_z(y_0,y_1,d_0,d_1). \quad (21)$$

If $a=1$, eq. (21) is immediate. If $a=0$, eq. (21) may be written $\pi_z(y_0,y_1,1,1) = \pi_z(y_0,y_1,1,0) + \pi_z(y_0,y_1,1,1)$. But, by assumption, the densities π_z are consistent with

Assumption 2, and thus $\pi_z(y_0, y_1, 1, 0) = 0$, and hence we see eq. (21) holds. Substituting eq. (21) into eq. (20), we then obtain that

$$\begin{aligned} &P^*(D(z,a) = 1, D(z',a') = 1) \\ &\qquad = \sum_{(y_0,y_1)\in\mathcal{Y}^2} \min\left\{\pi(y_0, y_1, D(z,a) = 1), \pi(y_0, y_1, D(z',a') = 1)\right\} \\ &\qquad\qquad \geq (1 - \delta_{z,z,a,a'})P^*(D(z,a) = 1), \end{aligned}$$

where the inequality is by condition 4. We have thus verified that the disagreement bound in eq. (5) is satisfied under P^*.

A.6 Proof of Corollary 2

The proof is completely analogous to that for Corollary 1, except appealing to Proposition 4 in place of Proposition 1.

A.7 Auxiliary lemmas

For the results in this section, let $z_{\max} \in \operatorname{argmax}_{z\in\mathcal{Z}} E[D \mid Z = z]$ denote the most lenient decision-maker, and let $\alpha_{\max} = E[D \mid Z = z_{\max}]$ denote their leniency. Let $\alpha_0 = P(D = 1)$ denote the treatment rate under the status quo.

Lemma 1. *Suppose $\alpha_0 < \alpha$. Let F_{P^*} denote the cdf of $Y(1) \mid D(Z,1) > D(Z,0)$ under P^*. Then for any $P^* \in \mathcal{P}_I^*(P; \mathcal{P}_\alpha^*)$ and $t \in \mathbb{R}$,*

$$\begin{aligned} F^{lb}(t) := \max\left\{\frac{1-\alpha_0}{\alpha-\alpha_0}F^{lb}_{Y(1)|D(Z,0)=0}(t) - \frac{1-\alpha}{\alpha-\alpha_0}, 0\right\} \\ \leq F_{P^*}(t) \leq F^{ub}(t) := \min\left\{\frac{1-\alpha_0}{\alpha-\alpha_0}F^{ub}_{Y(1)|D(Z,0)=0}(t), 1\right\}, \end{aligned}$$

where

$$F^{lb}_{Y(1)|D(Z,0)=0}(t) := \max\left\{\frac{1}{1-\alpha_0}\left(F^{lb}_{Y(1)}(t) - \alpha_0 P(Y \leq t \mid D = 1)\right), 0\right\}, \tag{22}$$

$$F^{ub}_{Y(1)|D(Z,0)=0}(t) := \min\left\{\frac{1}{1-\alpha_0}\left(F^{ub}_{Y(1)}(t) - \alpha_0 P(Y \leq t \mid D = 1)\right), 1\right\} \tag{23}$$

$$F^{lb}_{Y(1)}(t) := P(Y \leq t \mid D = 1, Z = z_{\max})\alpha_{\max}, \tag{24}$$

$$F^{ub}_{Y(1)}(t) := P(Y \leq t \mid D = 1, Z = z_{\max})\alpha_{\max} + 1 - \alpha_{\max}. \tag{25}$$

Proof. Since $P^*(D(Z,1) = 1 \mid D(Z,0) = 0) = (\alpha - \alpha_0)/(1 - \alpha_0)$, by the law of total probability, we have

$$P^*(Y(1) \leq t \mid D(Z,0) = 0) = P^*(Y(1) \leq t \mid D(Z,1) > D(Z,0))\frac{\alpha - \alpha_0}{1 - \alpha_0} + P^*(Y(1) \leq t \mid D(Z,1) = D(Z,0) = 0)\frac{1 - \alpha}{1 - \alpha_0}. \quad (26)$$

Rearranging the expression and using the fact that $P^*(Y(1) \leq t \mid D(Z,1) = 0, D(Z,0) = 0)$ and $F_{P^*}(t)$ both lie in the unit interval yields

$$\max\left\{\frac{1 - \alpha_0}{\alpha - \alpha_0}P^*(Y(1) \leq t \mid D(Z,0) = 0) - \frac{1 - \alpha}{\alpha - \alpha_0}, 0\right\} \leq F_{P^*}(t) \leq \min\left\{\frac{1 - \alpha_0}{\alpha - \alpha_0}P^*(Y(1) \leq t \mid D(Z,0) = 0), 1\right\}.$$

The claim follows if we can show that the cdf of $Y(1) \mid D(Z,0) = 0$ can be bounded by $F^{lb}_{Y(1)|D(Z,0)=0}(t)$ and $F^{ub}_{Y(1)|D(Z,0)=0}(t)$. To this end, note that by the law of total probability,

$$P^*(Y(1) \leq t) = P^*(Y(1) \leq t \mid D(Z,0) = 1)\alpha_0 \; + \; P^*(Y(1) \leq t \mid D(Z,0) = 0)(1 - \alpha_0).$$

Rearranging, and using the fact that $P^*(Y(1) \leq t \mid D(Z,0) = 1) = P(Y \leq t \mid D = 1)$ gives

$$P^*(Y(1) \leq t \mid D(Z,0) = 0) = (P^*(Y(1) \leq t) - \alpha_0 P(Y \leq t \mid D = 1))/(1 - \alpha_0). \quad (27)$$

We now bound $P^*(Y(1) \leq t)$. Note that since $P^*(Y(1) \leq t \mid D(z_{\max}, 0) = 1) = P(Y \leq t \mid D = 1, Z = z_{\max})$, applying the law of total probability again, we have

$$P^*(Y(1) \leq t) = P(Y \leq t \mid D = 1, Z = z_{\max})\alpha_{\max} + P^*(Y(1) \leq t \mid D = 0, Z = z_{\max})(1 - \alpha_{\max}).$$

Using the trivial bounds $P^*(Y(1) \leq t \mid D = 0, Z = z_{\max}) \in [0,1]$ yields $F^{lb}_{Y(1)}(t) \leq P^*(Y(1) \leq t) \leq F^{ub}_{Y(1)}(t)$. Plugging these bounds into eq. (27) then yields the bounds $F^{lb}_{Y(1)|D(Z,0)=0}(t) \leq P^*(Y(1) \leq t \mid D(Z,0) = 0) \leq F^{ub}_{Y(1)|D(Z,0)=0}(t)$, as claimed. □

Lemma 2. *Suppose $\alpha_0 < \alpha$. If $\mathcal{P}^*_I(P; \mathcal{P}^*_{\alpha,Mon}) \neq \emptyset$, then there exist distributions $P^{lb}, P^{ub} \in \mathcal{P}^*_I(P; \mathcal{P}^*_{\alpha,Mon})$ such that $P^{lb}(Y(1) \leq t \mid D(Z,1) > D(Z,0)) = F^{lb}(t)$ and $P^{ub}(Y(1) \leq t \mid D(Z,1) > D(Z,0)) = F^{ub}(t)$, with $F^{ub}(t)$ and $F^{lb}(t)$ defined as in Lemma 1.*

Proof. Fix an arbitrary $P^* \in \mathcal{P}^*_I(P; \mathcal{P}^*_{\alpha,Mon})$. By definition of the set $\mathcal{P}^*_I(P; \mathcal{P}^*_{\alpha,Mon})$, P^* is an element of this set if and only if it satisfies the following properties: (a) $P^*(D(Z,1) = 1) = \alpha$,

(b) P^* generates P, (c) Assumption 2 holds (d) Assumption 3 holds; (e) IV validity (eq. (2)) holds, and (f) $P^*(Y(0) = 0) = 1$.

Property (f) implies that it is sufficient to think of P^* as a distribution of the random vector $(Y(1), D(\cdot, 0), D(\cdot, 1), Z)$. Furthermore, under property (b), property (e) is equivalent to Z being independent of $(Y(1), D(\cdot, 0), D(\cdot, 1))$, with marginal distribution $P^*(Z = z) = P(Z = z)$. To construct P^{ub}, we tweak the distribution of $(Y(1), D(\cdot, 0), D(\cdot, 1))$ under P^*, so that the resulting distribution satisfies properties (a)–(d), as well as $P^{ub}(Y(1) \leq t \mid D(Z, 1) > D(Z, 0)) = F^{ub}_{Y(1)}(t)$.

Since $(D(\cdot, 0), D(\cdot, 1))$ are discrete, to specify the distribution of $(Y(1), D(\cdot, 0), D(\cdot, 1))$, it suffices to specify the probabilities $P^{ub}(Y(1) \leq t, D(\cdot, 0) = G_0, D(\cdot, 1) = G_1)$ for any $t \in \mathcal{Y}$, with $G_d = (G^1_d, \ldots, G^K_d)$ and $G^k_d \in \{0, 1\}$ for $d = 0, 1$. We set

$$
\begin{aligned}
&P^{ub}(Y(1) \leq t, D(\cdot, 0) = G_0, D(\cdot, 1) = G_1) = \\
&P^{ub}(D(\cdot, 1) = G_1 \mid Y(1) \leq t, D(\cdot, 0) = G_0) P^{ub}(Y(1) \leq t \mid D(\cdot, 0) = G_0) P^*(D(\cdot, 0) = G_0),
\end{aligned}
$$

where, writing $G_0 = 0$ as a shorthand for the vector of K zeros, $(0, \ldots, 0)$,

$$
P^{ub}(Y(1) \leq t \mid D(\cdot, 0) = G_0) = \begin{cases} 1 & \text{if } G_0 = 0, \\ P^*(Y(1) \leq t \mid D(\cdot, 0) = G_0) & \text{otherwise,} \end{cases}
$$

and

$$
P^{ub}(D(\cdot, 1) = G_1 \mid Y(1) \leq t, D(\cdot, 0) = G_0) = \prod_{k=1}^{K} \phi(G^k_1 \mid G^k_0, t),
$$

with the function ϕ specified below. In other words, P^{ub} matches P^* except that the conditional distribution of $Y(1) \mid D(\cdot, 0) = 0$ is degenerate and equal to the smallest value in the support $\mathcal{Y}$ (as we shall see below, this ensures that the marginal distribution of $Y(1)$ matches $F^{ub}_{Y(1)}$ in eq. (25)), and the conditional distribution of $D(z, 1) \mid D(z, 0), Y(1)$ is the same for all z, and conditional on $Y(1)$, the vectors $\{(D(z, 0), D(z, 1))\}_z$ are independent across z.

We first verify that $P^{ub} \in \mathcal{P}^*_I(P; \mathcal{P}^*_{\alpha, Mon})$ so long as ϕ is chosen so that

$$
E_{P^{ub}}[D(Z, 1)] = \alpha, \quad \text{and} \tag{28}
$$

$$
\phi(0 \mid 1, t) = 0. \tag{29}
$$

Second, we show that the distribution of $Y(1) \mid D(Z, 0) = 0$ under P^{ub} matches $F^{ub}_{Y(1)|D(Z,0)=0}$ in eq. (23). Finally, we choose ϕ so that the conditional cdf of $Y(1) \mid D(Z, 1) > D(Z, 0)$

matches F^{ub}. Note eq. (28) implies that property (a) holds. Since $Y(0) = 0$, the data distribution depends only on the marginal distribution of $D(\cdot, 0)$ and the conditional distribution of $Y(1) \mid D(\cdot, 0) \neq 0$. Since these coincide with P^*, P^{ub} generates P, so that property (b) also holds. Likewise, since IV monotonicity depends only on the distribution of $D(\cdot, 0)$, and this matches P^*, property (d) also holds. Finally, eq. (29) implies policy monotonicity, so that property (c) holds also. Hence, under eqs. (28) and (29) $P^{ub} \in \mathcal{P}_I^*(P; \mathcal{P}^*_{\alpha,Mon})$ as claimed.

Next, we verify that $Y(1) \mid D(Z, 0) = 0$ under P^{ub} matches $F^{ub}_{Y(1)|D(Z,0)=0}$ in eq. (23). Since $P^{ub} \in \mathcal{P}_I^*(P; \mathcal{P}^*_\alpha)$, replacing P^* with P^{ub} in eq. (27) yields

$$P^{ub}(Y(1) \leq t \mid D(Z, 0) = 0) = (P^{ub}(Y(1) \leq t) - \alpha_0 P(Y \leq t \mid D = 1))/(1 - \alpha_0). \tag{30}$$

Under IV monotonicity, the event that $D(\cdot, 0) = 0$ is equivalent to $D(z_{\max}, 0) = 0$. Hence, by the law of total probability, and the fact that P^{ub} generates P,

$$\begin{aligned} &P^{ub}(Y(1) \leq t) \\ &\quad = P^{ub}(Y(1) \leq t \mid D(\cdot, 0) = 0)(1 - \alpha_{\max}) + P(Y \leq t \mid D = 1, Z = z_{\max})\alpha_{\max} = F^{ub}_{Y(1)}(t), \end{aligned}$$

with $F^{ub}_{Y(1)}$ defined in eq. (25). Replacing $P^{ub}(Y(1) \leq t)$ with $F^{ub}_{Y(1)}(t)$ in eq. (30) yields

$$P^{ub}(Y(1) \leq t \mid D(Z, 0) = 0) = \frac{1}{1 - \alpha_0}(F^{ub}_{Y(1)}(t) - \alpha_0 P(Y \leq t \mid D = 1)).$$

The right-hand side must be smaller than 1, since P^{ub} is a valid probability measure (because $P^{ub} \in \mathcal{P}_I^*(P; \mathcal{P}^*_\alpha)$). Hence, the right-hand side matches $F^{ub}_{Y(1)|D(Z,0)=0}$ in eq. (23) as claimed.

Finally, we choose ϕ so that the conditional cdf of $Y(1) \mid D(Z, 1) > D(Z, 0)$ matches F^{ub}. Given eq. (29), it suffices to specify the conditional distribution of $D(z, 1) \mid D(z, 0) = 0, Y(1) = y$ (the conditional distribution exists by Proposition 10.2.8 in Dudley (2002)). Let y_η denote the η quantile of the conditional distribution of $Y(1) \mid D(Z, 0) = 0$ under P^{ub}, with $\eta := (\alpha - \alpha_0)/(1 - \alpha_0)$. We set

$$P^{ub}(D(z, 1) = 1 \mid D(z, 0) = 0, Y(1) = y) = \begin{cases} 1 & \text{if } y < y_\eta, \\ \frac{\eta - P^{ub}_{Y(1)|D(Z,0)=0}(Y(1) < y_\eta)}{P^{ub}_{Y(1)|D(Z,0)=0}(Y(1) = y_\eta)} & \text{if } y = y_\eta, \\ 0 & \text{if } y > y_\eta, \end{cases} \tag{31}$$

where we interpret $0/0$ as 0 if $P^{ub}_{Y(1)|D(Z,0)=0}(Y(1) = y_\eta) = 0$. The intuition for this choice is that by Bayes rule, the density of $Y(1) \mid D(Z, 1) > D(Z, 0)$ is proportional to $P^{ub}(D(z, 1) =$

$1 \mid D(z,0) = 0, Y(1) = y)P^{ub}(Y(1) = y)$, so that this choice shifts the density as far left as possible while satisfying the requirement that $E[D(Z,1)] = \alpha$, thus maximizing F_{P^*}. Under this choice, by iterated expectations,

$$\begin{aligned} P^{ub}(D(Z,1) = 1) &= P^{ub}(D(Z,1) = 1 \mid D(Z,0) = 0)(1-\alpha_0) + \alpha_0 \\ &= E_{P^{ub}}[P^{ub}(D(z,1) = 1 \mid D(z,0) = 0, Y(1)) \mid D(z,0) = 0](1-\alpha_0) + \alpha_0 \\ &= \alpha, \end{aligned}$$

so that eq. (28) holds. Furthermore, since $P^{ub} \in \mathcal{P}_I^*(P; \mathcal{P}_\alpha^*)$, rearranging eq. (26) with P^* replaced by P^{ub} yields

$$\begin{aligned} &P^{ub}(Y(1) \le t \mid D(Z,1) > D(Z,0)) \\ &= \frac{1-\alpha_0}{\alpha - \alpha_0} P^{ub}(Y(1) \le t \mid D(Z,0) = 0) - \frac{1-\alpha}{\alpha-\alpha_0} P^{ub}(Y(1) \le t \mid D(Z,1) = D(Z,0) = 0) \\ &\qquad = \frac{1-\alpha_0}{\alpha-\alpha_0} F^{ub}_{Y(1)|D(Z,0)=0}(t) \left(1 - P^{ub}(D(Z,1) = 0 \mid Y(1) \le t, D(Z,0) = 0)\right) \end{aligned}$$

where the second equality uses Bayes rule and the identity $P^{ub}(D(Z,1) = 0 \mid D(Z,0) = 0) = (1-\alpha)/(1-\alpha_0)$. By iterated expectations and eq. (31), $1 - P^{ub}(D(Z,1) = 0 \mid Y(1) \le t, D(Z,0) = 0) = 1$ if $t < y_\eta$, and equals $\eta / P_{Y(1)|D(Z,0)=0}(Y(1) \le t)$ otherwise. Since $\eta / P_{Y(1)|D(Z,0)=0}(Y(1) \le t) = \frac{\alpha-\alpha_0}{1-\alpha_0} F^{ub}_{Y(1)|D(Z,0)=0}(t)$, it follows that

$$P^{ub}(Y(1) \le t \mid D(Z,1) > D(Z,0)) = \min\left\{ \frac{1-\alpha_0}{\alpha-\alpha_0} F^{ub}_{Y(1)|D(Z,0)=0}(t), 1 \right\},$$

which matches eq. (23) as claimed.

The construction of P^{lb} is identical, except we set the distribution of $Y(1) \mid D(\cdot,0) = 0$ to be degenerate and equal to the largest value in the support $\mathcal{Y}$, and set

$$P^{lb}(D(z,1) = 1 \mid D(z,0) = 0, Y(1) = y) = \begin{cases} 0 & \text{if } y < y_{1-\eta}, \\ \frac{\eta - P^{lb}_{Y(1)|D(Z,0)=0}(Y(1) > y_{1-\eta})}{P^{lb}_{Y(1)|D(Z,0)=0}(Y(1) = y_\eta)} & \text{if } y = y_{1-\eta}, \\ 1 & \text{if } y > y_{1-\eta}, \end{cases}$$

where $y_{1-\eta}$ denotes the $1-\eta$ quantile of the conditional distribution of $Y(1) \mid D(Z,0) = 0$ (and we again interpret $0/0$ as 0 if $P^{lb}_{Y(1)|D(Z,0)=0}(Y(1) = y_\eta) = 0$). □

Appendix B Additional details

B.1 Example where IV monotonicity helps

Consider a setting with a binary instrument and binary outcomes, such that the observed data probabilities are given by:

$P(Y=y, D=d \mid Z=1)$	$Y=1$	$Y=0$
$D=1$	1/3	1/3
$D=0$	1/3	0

$P(Y=y, D=d \mid Z=0)$	$Y=1$	$Y=0$
$D=1$	1/3	0
$D=0$	1/3	1/3

Assume IV monotonicity holds. Then there are three response types under the status quo, instrument compliers (C), instrument always takers (A), and instrument never-takers (N), and it follows from the data probabilities that these all have population share 1/3. Furthermore, since the only untreated types when $Z=1$ are N, and $P(Y=0, D=0 \mid Z=1)=0$, it follows that $P(Y(0)=1 \mid N)=1$. But since $P(Y=0, D=0 \mid Z=0)=1/3$, it follows that $P(Y(0)=0 \mid C)=1$. Similarly, since the only treated types when $Z=0$ are A, and $P(Y=0, D=1 \mid Z=0)=0$, it follows that $P(Y(1)=1 \mid A)=1$, which, combined with the fact that $P(Y=0, D=1 \mid Z=1)=1/3$ implies that $P(Y(1)=0 \mid C)=1$. Hence, instrument compliers have zero treatment effect, while that for instrument never takers lies between -1 and 0. Consider a counterfactual policy satisfying policy monotonicity that increases treatment take-up by some small amount ϵ. Since policy compliers are drawn from the pool of instrument never-takers and instrument compliers, it follows that the bounds for policy compliers are given by $E[Y(1)-Y(0) \mid D(Z,1) > D(Z,0)] \in [-1, 0]$. In particular, IV monotonicity allows us to conclude that the policy must have weakly negative impacts.

Without IV monotonicity, however, the data is also compatible with the population comprising only instrument defiers (D) with population share 1/3 such that $P(Y(1) = Y(0) = 1 \mid D) = 1$, with the remainder of the population being instrument compliers, half of whom have treatment effect equal to -1 and half of whom have a treatment effect equal to 1. Thus, without IV monotonicity, we only obtain trivial bounds on the effect for policy compliers, $E[Y(1)-Y(0) \mid D(Z,1) > D(Z,0)] \in [-1, -1]$.

This example is quite extreme, however, in that we were able to rule out a positive treatment effect for instrument never-takers under monotonicity. So long as we cannot rule out that mass ϵ of instrument never-takers has a treatment effect equal to 1, assuming IV monotonicity will also lead to trivial bounds on the effect for policy compliers for policies that only marginally increase take-up.

B.2 Implementation details

Calibration of disagreement probability bounds To calibrate the parameter $\overline{\mathsf{DP}}$ in eq. (11), we proceed in two steps. First, we relate it to a correlation parameter in a Gaussian signal model with correlated signals. Second, we match the correlation parameter in this model to the data from Sigstad (2026). We then calculate the implied value of $\overline{\mathsf{DP}}$.

We consider a model in which judges obtain noisy signals of the criminal misconduct risk of defendants, V_z, similar to the model in Chan et al. (2022, Section V). We assume that for each pair of judges $z \in \mathcal{Q}$ and $z' \in \mathcal{Q}^c$, the signals are jointly normal, with means normalized to zero and variances normalized to one,

$$\begin{pmatrix} V_z \\ V_{z'} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

where ρ governs the correlation between the signals. Each judge then releases the defendant if the signal is low enough, $D(z,1) = \mathrm{I}\{V_z \leq \alpha_z\}$, where the cutoffs α_z are calibrated to match the release rate of judge z, $\alpha_z = \Phi^{-1}(q)$ for $z \in \mathcal{Q}$, where q is the release quota and Φ standard normal cdf, and $\alpha_{z'} = \Phi^{-1}(E[D(z',0)])$ if $z \in \mathcal{Q}^c$, since judges not subject to the quota are assumed not to change their behavior under the counterfactual. Under this model, the average disagreement probability equals

$$\begin{aligned} P(D(Z_{\mathcal{Q}},1)=0, D(Z_{\mathcal{Q}^c},1)=1) &= \frac{1}{|\mathcal{Q}||\mathcal{Q}^c|} \sum_{z\in\mathcal{Q}, z'\in\mathcal{Q}^c} P(D(z,1)=0, D(z',1)=1) \\ &= \frac{1}{|\mathcal{Q}||\mathcal{Q}^c|} \sum_{z\in\mathcal{Q}, z'\in\mathcal{Q}^c} P(V_z > \alpha_z, V_{z'} < \alpha_{z'}). \end{aligned}$$

Thus, given a value of ρ, we can compute the probability on the right-hand side, and calculate the implied value of the disagreement probability as $\overline{\mathsf{DP}} = \frac{1}{|\mathcal{Q}^c|}\sum_{z'\in\mathcal{Q}^c} P(V_z > \Phi^{-1}(q), V_{z'} < \alpha_{z'})/q$, using the fact that the marginal release rates equal the quota q, and that judges below quota have the same release rate. In our application, the number of cases handled by each judge differs a lot, so instead of a simple average, we use a weighted average of the probabilities $P(V_z > \Phi^{-1}(q), V_{z'} < \alpha_{z'})$, weighted by the number of cases handled by z and z'.

As a benchmark, we use the average value of ρ across pairs of judges from the top decile and bottom nine deciles of leniency in the data from Sigstad (2026). This data has information on panels of judges ruling on criminal cases in the São Paulo Appeal Court. As described in the main text, we focus on three judge panels. To calculate judge leniencies, we

regress judge decisions on judge and case fixed effects, and form leniency deciles. To reduce estimation noise, we then restrict the data pairs of judges who see at least a thousand cases together. Then, for each pair of judges z and z' such that z is in the bottom 90% and z' in the top decile, we calculate the value of correlation $\rho_{z,z'}$ implied by the joint distribution of their votes. Averaging across all such pairs then yields the value $\rho = 0.989$.

Details on inference For the universal release policy, we first form a joint upper 95% confidence band for the probabilities $\{P^*(Y(1) = 1, D(z,0) = 1 \mid R = r), P^*(Y(1) = 0, D(z,0) = 1 \mid R = r)\}_{z \in \mathcal{Z}}$, separately for each race. We do this assuming that the sample analogs are approximated by a Gaussian distribution with covariance matrix calculated under the assumption that the data is independent across judges (thus ignoring covariate adjustment). The sup-t critical value for the confidence band is then calculated by simulation. We then form confidence intervals for the judge-specific intervals $\mathcal{I}_z$ in eq. (4), as shown in Figure 2. The confidence interval reported in Table 1 are then formed by intersecting these judge-specific bounds.

For inference on the disparate impact parameter Δ, we adopt a two-step approach. We first form a joint confidence set for $(E[Y(1) \mid R = w], E[Y(1) \mid R = b])$ with nominal size 4.5%, as described in the previous paragraph, except that we form a joint upper 95.5% confidence band for the probabilities $\{P^*(Y(1) = 1, D(z,0) = 1 \mid R = r), P^*(Y(1) = 0, D(z,0) = 1 \mid R = r)\}_{z \in \mathcal{Z}}$ *jointly* for both races. Intersecting the judge-specific probabilities for each race then yields a joint 95.5% confidence interval for the race-specific universal release parameters. In the second step, for each value of $(E[Y(1) \mid R = w], E[Y(1) \mid R = b])$ in the confidence set, we form a delta-method confidence interval with level 99.5% for Δ that accounts for the statistical uncertainty in the aggregate moments (more precisely, we do this for each value within a 100×100 grid of points within the confidence band). Our final confidence interval is formed by taking unions of all the second-step intervals.

For inference on the quota policy, we take advantage of the fact that the data only enters the program through the data-compatibility constraints (condition 1 in Proposition 1), which we may write as $A\pi = \mu$, where μ stacks the $4K$ probabilities $P(Y = y, D = d \mid Z = z)$, $y, d \in \{0,1\}, z \in \mathcal{Z}$, and the matrix A is not data-dependent. We further augment the data vector with three aggregate moments, μ_A. These correspond to $E[Y \mid Z \in \mathcal{Q}] - E[Y \mid Z \in \mathcal{Q}^c]$ (up to scaling, this is the estimate of the policy effect under the reallocation policy), and the aggregate release rates $E[D \mid Z \in \mathcal{Q}]$ and $E[D \mid Z \in \mathcal{Q}^c]$. We augment μ with this vector and augment A so that the last three rows of $A\pi$ match μ_A. The addition of these aggregate moments doesn't impact point estimates, but it helps tighten inference.

We form a joint two-sided confidence band for μ of the form $\hat{\mu} \pm cv_1 se(\hat{\mu})$, with the

	$E[Y(1) \mid R = b]$		$E[Y(1) \mid R = w]$		Δ	
	Orig.	Repl.	Orig.	Repl.	Orig.	Repl.
Specification	(1)	(2)	(3)	(4)	(5)	(6)
Linear extrap.	40.0	40.0	33.8	33.5	5.4	5.4
	(0.6)	(0.5)	(0.7)	(0.6)	(0.2)	(0.2)
Quadratic extrap.	39.4	42.8	31.9	36.2	5.4	4.8
	(2.1)	(1.8)	(2.1)	(1.8)	(0.7)	(0.7)
Local linear extr.	43.6	43.6	34.6	35.0	4.2	4.3
	(1.6)	(2.3)	(1.4)	(2.0)	(0.6)	(0.7)

Notes: This table compares the estimates and standard errors (in parentheses) from Table 3 of ADH22 to our replication, as reported in Table 1.

Appendix Table B.1: Comparison of estimates and standard errors in ADH22 and our Replication.

critical value cv_1 computed by simulating from a Gaussian distribution with correlation matrix matching that of $\hat{\mu}$, with level 99%. Here $se(\hat{\mu})$ is a vector of standard errors for μ. We form an analogous joint confidence band for μ_A, $\hat{\mu}_A \pm cv_2 se(\hat{\mu}_A)$ with level 96%. We then replace the constraint $A\pi = (\hat{\mu}, \hat{\mu}_A)$ with the constraints $A\pi \geq (\hat{\mu} - cv_1 se(\hat{\mu}), \hat{\mu}_A - cv_2 se(\hat{\mu}_A))$ and $A\pi \leq (\hat{\mu} + cv_1 se(\hat{\mu}), \hat{\mu}_A + cv_2 se(\hat{\mu}_A))$.

Finally, to compute the bounds and confidence intervals for policy compliers, we use the fact that the parameter of interest is a ratio, $\theta / \frac{n_Q}{n}(E[D(Z_Q, 0)] - q)$, where n_Q is the number of cases handled by the bottom 90% of judges. The denominator is an affine function of π, so this is a linear-fractional program. We transform it to a linear program using the Charnes–Cooper transformation, and form confidence intervals by projection as described above.

B.3 Additional tables and figures

Table B.1 compares our estimates and standard errors for the parametric extrapolation methods, as reported in Table 1, to the original estimates in Table 3 in ADH22. The CIs in ADH22 account for the fact that the judge-specific means are covariate-adjusted, whereas since we do not have the microdata, we conduct inference assuming the judge-specific means are sample means from i.i.d. draws. There are also some differences in weighting: ADH22 weight by the inverse precision of the judge fixed effects (accounting for covariate estimation), whereas we weight by the number of released defendants. In spite of these differences, the

	Whites		Blacks		
	$E[Y(1) \mid R]$	PC TE	$E[Y(1) \mid R]$	PC TE	Δ
Specification	(1)	(2)	(3)	(4)	(5)
Pooled	$(29.2, 49.1)$	$(20.9, 86.3)$	$(23.4, 41.9)$	$(12.0, 90.7)$	$(1.0, 9.6)$
Disaggregated	$(29.1, 49.3)$	$(20.5, 86.9)$	$(22.6, 42.8)$	$(8.7, 94.3)$	$(0.9, 9.9)$

Notes: See Table 1. 95% confidence intervals in parentheses; these do not account for covariate adjustment. "Pooled" refers to the main specification that pools judges with 300 or fewer cases; it is identical to the first row of Table 1. "Disaggregated" refers to estimates without pooling.

Appendix Table B.2: Estimates for universal release policy the disparate impact parameter using NYC bail judge data with and without pooling.

	no PC bound		PC bounds	
	Policy effect	PC TE	Policy effect	PC TE
Specification	(1)	(2)	(3)	(4)
A: Blacks (Pooled)				
Valid IV only	$(0.0, 10.7)$	$(0.0, 100.0)$	$(1.8, 10.7)$	$(16.4, 100.0)$
$\overline{DP} = 0.025$	$(2.8, 7.9)$	$(26.3, 76.7)$	$(2.8, 7.9)$	$(26.3, 76.7)$
B: Blacks (Disaggregated)				
Valid IV only	$(0.0, 10.7)$	$(0.0, 100.0)$	$(1.6, 10.7)$	$(15.2, 100.0)$
$\overline{DP} = 0.025$	$(2.8, 7.9)$	$(26.3, 76.7)$	$(2.8, 7.9)$	$(26.3, 76.7)$
C: Whites (Pooled)				
Valid IV only	$(0.0, 6.9)$	$(0.0, 100.0)$	$(0.7, 6.9)$	$(9.9, 100.0)$
$\overline{DP} = 0.021$	$(1.2, 5.9)$	$(17.0, 91.1)$	$(1.2, 5.9)$	$(17.0, 91.1)$
D: Whites (Disaggregated)				
Valid IV only	$(0.0, 6.9)$	$(0.0, 100.0)$	$(0.6, 6.9)$	$(8.8, 100.0)$
$\overline{DP} = 0.021$	$(1.2, 5.9)$	$(17.0, 91.1)$	$(1.2, 5.9)$	$(17.0, 91.1)$

Notes: See Table 2. 95% confidence intervals in parentheses; these do not account for covariate adjustment. "Pooled" refers to the main specification that pools judges with 300 or fewer cases; Panels A and C are identical to Panels A and B in Table 2. "Disaggregated" refers to estimates without pooling.

Appendix Table B.3: Estimates for quota policy using NYC bail judge data with and without pooling.

Specification	Policy effect (1)	PC TE (2)
A: Pooled		
Valid IV only	$(-9.4, 9.4)$	$(-100.0, 100.0)$
$\overline{OD} = 0.019$	$(-5.4, 2.0)$	$(-62.6, 23.2)$
B: Disaggregated		
Valid IV only	$(-9.4, 9.4)$	$(-100.0, 100.0)$
$\overline{OD} = 0.019$	$(-5.4, 2.0)$	$(-62.6, 23.2)$

Notes: See Table 3. 95% confidence intervals reported in parentheses. "Pooled" refers to the main specification that pools ADAs with 300 or fewer cases; Panels A is identical to Panels A and B in Table 3. "Disaggregated" refers to estimates without pooling.

Appendix Table B.4: Estimates for quota policy using Suffolk County prosecutor data with and without pooling.

estimates and standard errors are very similar for all specifications.

Table B.2 and Table B.3 show that inference for the NYC bail judge data is virually identical if we do not pool judges with a few cases; in some instances, the confidence intervals are wider, since the projection-based confidence intervals use larger critical values. Table B.4 shows that aggregation has no impact on inference in the Suffolk County prosecutor data. In all cases, the point estimates without pooling are empty, since the estimated probabilities $P(Y(d) = y, D(z, 0) = d)$ are not compatible with eq. (2). Pooling reduces the estimation noise in these probabilities for decision-makers with a few cases; it also helps ensure that the estimated probabilities are approximately Gaussian, as assumed by our projection inference method.

References

Agan, A., Doleac, J. L., & Harvey, A. (2023). Misdemeanor prosecution. *The Quarterly Journal of Economics*, *138*(3), 1453–1505. https://doi.org/10.1093/qje/qjad005

Aizer, A., & Doyle, J., Joseph J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, *130*(2), 759–803. https://doi.org/10.1093/qje/qjv003

Albright, A. (2022, July). *No money bail, no problems? Trade-offs in a pretrial automatic release program*. https://doi.org/10.31235/osf.io/42pbz

Angelova, V., Dobbie, W., & Yang, C. (2025). Algorithmic recommendations and human discretion. *Review of Economic Studies*. https://doi.org/10.1093/restud/rdaf084

Arnold, D., Dobbie, W., & Hull, P. (2022). Measuring racial discrimination in bail decisions. *American Economic Review*, *112*(9), 2992–3038. https://doi.org/10.1257/aer.20201653

Bai, Y., Huang, S., Moon, S., Santos, A., Shaikh, A. M., & Vytlacil, E. J. (2025, November). *Inference for treatment effects conditional on generalized principal strata using instrumental variables*. arXiv: 2411.05220.

Bai, Y., Huang, S., Moon, S., Shaikh, A. M., & Vytlacil, E. J. (2025, November). *On the identifying power of monotonicity for average treatment effects*. arXiv: 2405.14104.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176. https://doi.org/10.1080/01621459.1997.10474074

Baron, E. J., Doyle Jr, J. J., Emanuel, N., Hull, P., & Ryan, J. (2024). Discrimination in multiphase systems: Evidence from child protection. *The Quarterly Journal of Economics*, *139*(3), 1611–1664. https://doi.org/10.1093/qje/qjae007

Chan, D. C., Gentzkow, M., & Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, *137*(2), 729–783. https://doi.org/10.1093/qje/qjab048

Chen, X., & Flores, C. A. (2015). Bounds on treatment effects in the presence of sample selection and noncompliance: The wage effects of job corps. *Journal of Business & Economic Statistics*, *33*(4), 523–540. https://doi.org/10.1080/07350015.2014.975229

de Chaisemartin, C. (2017). Tolerating defiance? Local average treatment effects without monotonicity. *Quantitative Economics*, *8*(2), 367–396. https://doi.org/10.3982/QE601

Dobbie, W., Goldin, J., & Yang, C. S. (2018). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, *108*(2), 201–240. https://doi.org/10.1257/aer.20161503

Dobbie, W., & Song, J. (2015). Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. *American Economic Review*, *105*(3), 1272–1311. https://doi.org/10.1257/aer.20130612

Doyle, J. J., Jr. (2007). Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, *97*(5), 1583–1610. https://doi.org/10.1257/aer.97.5.1583

Dudley, R. M. (2002). *Real analysis and probability*. Cambridge University Press.

Frandsen, B., Lefgren, L., & Leslie, E. (2023). Judging judge fixed effects. *American Economic Review*, *113*(1), 253–277. https://doi.org/10.1257/aer.20201860

Heckman, J. J., Urzúa, S., & Vytlacil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, *88*(3), 389–432. https://doi.org/10.1162/rest.88.3.389

Heckman, J. J., & Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, *96*(8), 4730–4734. https://doi.org/10.1073/pnas.96.8.4730

Heckman, J. J., & Vytlacil, E. J. (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica*, *73*(3), 669–738. https://doi.org/10.1111/j.1468-0262.2005.00594.x

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (pp. 4779–4874, Vol. 6B). Elsevier. https://doi.org/10.1016/S1573-4412(07)06070-9

Iaryczower, M., & Shum, M. (2012). The value of information in the court: Get it right, keep it tight. *American Economic Review*, *102*(1), 202–237. https://doi.org/10.1257/aer.102.1.202

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475. https://doi.org/10.2307/2951620

Kamat, V. (2019, October). *On the identifying content of instrument monotonicity*. arXiv: 1807.01661.

Kitagawa, T. (2021). The identification region of the potential outcome distributions under instrument independence. *Journal of Econometrics*, *225*(2), 231–253. https://doi.org/10.1016/j.jeconom.2021.03.006

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, *96*(3), 863–876. https://doi.org/10.1257/aer.96.3.863

Locher, L., Stensrud, M. J., & Sarvet, A. L. (2025, September). *Interpretational errors with instrumental variables*. arXiv: 2509.02045.

Machado, C., Shaikh, A. M., & Vytlacil, E. J. (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics*, *212*(2), 522–555. https://doi.org/10.1016/j.jeconom.2018.04.007

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, *80*(2), 319–323.

Manski, C. F. (1997). Monotone treatment response. *Econometrica*, *65*(6), 1311. https://doi.org/10.2307/2171738

Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of majorization and its applications*. Springer New York. https://doi.org/10.1007/978-0-387-68276-1

Mogstad, M., & Torgovitsky, A. (2024, January). Instrumental variables with unobserved heterogeneity in treatment effects. In C. Dustmann & T. Lemieux (Eds.), *Handbook of labor economics* (pp. 1–114, Vol. 5). Elsevier. https://doi.org/10.1016/bs.heslab.2024.11.003

Mogstad, M., Torgovitsky, A., & Walters, C. R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, *111*(11), 3663–3698. https://doi.org/10.1257/aer.20190221

Mueller-Smith, M. (2015, August). *The criminal and labor market impacts of incarceration* [Working paper, University of Michigan].

Norris, S., Pecenco, M., & Weaver, J. (2021). The effects of parental and sibling incarceration: Evidence from Ohio. *American Economic Review*, *111*(9), 2926–2963. https://doi.org/10.1257/aer.20190415

Richardson, T. S., & Robins, J. M. (2014). ACE bounds; SEMs with equilibrium conditions. *Statistical Science*, *29*(3), 363–366. https://doi.org/10.1214/14-STS485

Sampat, B., & Williams, H. L. (2019). How do patents affect follow-on innovation? Evidence from the human genome. *American Economic Review*, *109*(1), 203–236. https://doi.org/10.1257/aer.20151398

Sigstad, H. (2024, April). *Marginal treatment effects and monotonicity*. arXiv: 2404.03235.

Sigstad, H. (2026). Monotonicity among judges: Evidence from judicial panels and consequences for judge IV designs. *American Economic Review*, *116*(1), 189–208. https://doi.org/10.1257/aer.20231104

Skemer, M., Redcross, C., & Bloom, H. (2020, September). *Pursuing pretrial justice through an alternative to bail: Findings from an evaluation of new york city's supervised release program* (tech. rep.). MRDC.

Small, D. S., Tan, Z., Ramsahai, R. R., Lorch, S. A., & Brookhart, M. A. (2017). Instrumental variable estimation with a stochastic monotonicity assumption. *Statistical Science*, *32*(4), 561–579. https://doi.org/10.1214/17-STS623

Song, Y., Guo, F. R., Chan, K. C. G., & Richardson, T. S. (2025, November). The categorical instrumental variable model: Characterization, partial identification, and statistical inference.

Stevenson, M. (2018). Assessing risk assessment in action. *Minnesota Law Review*, *103*, 303–384.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., & Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: Review of

methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, *113*(522), 933–947. https://doi.org/10.1080/01621459.2018.1434530

Ura, T., & Zhang, L. (2025, May). *Policy relevant treatment effects with multidimensional unobserved heterogeneity*. arXiv: 2403.13738.

Vytlacil, E. J. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, *70*(1), 331–341. https://doi.org/10.1111/1468-0262.00277