I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17636

Examiner and Judge Designs in Economics: A Practitioner's Guide

Eric Chyn
Brigham Frandsen
Emily Leslie

JANUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17636

Examiner and Judge Designs in Economics: A Practitioner's Guide

Eric Chyn
University of Texas at Austin and NBER

Brigham Frandsen
Brigham Young University and IZA

Emily Leslie
Brigham Young University and IZA

JANUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Examiner and Judge Designs in Economics: A Practitioner's Guide*

This article provides empirical researchers with an introduction and guide to research designs based on variation in judge and examiner tendencies to administer treatments or other interventions. We review the basic theory behind the research design, outline the assumptions under which the design identifies causal effects, describe empirical tests of the conditions for identification, and discuss tradeoffs associated with choices researchers must make for estimation. We demonstrate concepts and best practices concretely in an empirical case study that uses an examiner tendency research design to study the effects of pre-trial detention.

JEL Classification: C21, C26, C31, K14, C54
Keywords: judges, examiners, judge fixed effects, applied econometrics

Corresponding author:
Emily Leslie
Department of Economics
Brigham Young University
Provo, UT 84602
USA
E-mail: emily.leslie@byu.edu

* Programs and data for the empirical case study are provided online. We are grateful for insightful comments from David Romer and four anonymous referees. In addition, we received constructive feedback from Amanda Agan, Elior Cohen, Rob Collinson, Jason Cook, Gordon Dahl, John Eric Humphries, Lawrence Katz, Magne Mogstad, Samuel Norris, Aurelie Ouss, Roman Rivera, Henrik Sigstad, Kamelia Stavreva, Megan Stevenson, and Winnie van Dijk.

1 Introduction

In 1932, criminologists in New Jersey documented wide disparities in the sentencing tendencies of trial judges in their state. The most severe judge imprisoned 57.7% of the convicted defendants randomly assigned to their courtroom, and the most lenient only 33.6% (Gaudet et al., 1932). The same study also described an early experiment in which over a hundred mathematics teachers were asked to grade the same exam, producing scores that ranged from 28 to 92. While variation in decision-maker tendencies raises issues of fairness, it also provides a convincing empirical strategy.

In the past two decades, researchers have begun using disparities among judges and other decision makers like those documented nearly a century ago to identify causal effects in non-experimental settings. For example, in pioneering work, Kling (2006) estimated the impact of incarceration length on post-release labor market earnings by leveraging plausibly exogenous variation in sentencing arising from the rules that assign offenders to judges.[1] This strategy is sometimes called the "judge fixed effects" or "judge leniency" design. We will generally refer to this strategy as the "examiner tendency" design, since it has been applied in at least 136 studies across a range of settings that feature various types of decision-makers (see Table A.1 in the online appendix). Recent examples include studies of the effects of pre-trial detention, consumer bankruptcy, foster care, disability benefits, patents, medical diagnoses, and health treatments.

The key ingredient in this research design is that examiners with different tendencies will expose comparable individuals to different treatments or interventions. In the ideal scenario, administrative procedures ensure that assignment to examiners is independent of other factors that determine the outcome besides the treatment. In addition, examiners ideally should affect outcomes only through the treatment of interest. Quasi-random assignment to examiners mimics random assignment to treatment and control groups in a randomized control trial.

This paper aims to provide an up-to-date overview of examiner tendency designs and create a guide for researchers interested in applying this method. Our overview is motivated by recent methodological work and the fact that there is no single comprehensive summary of what is understood about examiner tendency designs. We aim to clarify the conditions under which examiner tendency designs will succeed or fail. Moreover, we hope to provide a guide for common implementation decisions that are not (currently) covered in standard econometric texts.

To set the stage for the rest of the paper, the following overview highlights our key points about examiner tendency designs:

- **The validity of the examiner tendency design rests not only on random assignment (or conditional random assignment), but also on the plausibility that exclusion and**

[1] In earlier work, Waldfogel (1995) leveraged variation across judges to calibrate a structural model and study the selection of cases for trial.

monotonicity conditions hold. Exclusion requires that examiners influence outcomes only through the treatment of interest, and monotonicity requires that an individual treated by an examiner with a lower propensity to treat would surely be treated by an examiner with a higher propensity. When these conditions hold, instrumental variables (IV) estimation identifies a proper weighted average (i.e., one that uses nonnegative weights) of local average treatment effects (LATEs). This weighted average reflects causal effects for individuals who would have received a different treatment status if they had been assigned to a different examiner. In addition, the strictest version of monotonicity—dubbed *pairwise monotonicity*—also allows identification of marginal treatment effects (MTEs). Under a weaker version, *average monotonicity*, which allows for violations of pairwise monotonicity, the IV estimand still has a causal interpretation, but MTEs are no longer identified.

- **When examiners affect outcomes through multiple treatments, the design fails to identify causal effects when strong conditions on how outcomes respond or how examiners decide on treatments do not hold.** If outcomes respond to treatments linearly and with constant effects, linear IV identifies them, as long as the number of treatments does not exceed the number of examiners and examiners vary sufficiently in their propensities. Outside of the constant treatment effects framework, however, linear IV only identifies proper weighted average effects under stringent—and, in many cases, difficult to motivate—conditions on how examiners allocate individuals to treatments (Humphries et al., 2023; Bhuller and Sigstad, 2022).

- **Jackknife instrumental variables estimation (JIVE) and related approaches eliminate many-instruments bias that could distort IV estimation when there are many examiners.** In the simple case with no additional covariates, JIVE is equivalent to the common practice of IV estimation using a leave-out mean as the instrument. When there are additional exogenous covariates, the improved jackknife procedure (IJIVE) proposed by Ackerberg and Devereux (2009) or the unbiased jackknife estimator (UJIVE) proposed by Kolesár (2013) ensures that covariates are handled consistently in the first and second stages and eliminates additional biases that can accompany the estimation of covariate effects. When individuals are assigned to examiners in clusters or groups, the jackknife leave-out procedure should be implemented at the cluster level (Frandsen et al., 2023b).

- **Whether clustering is necessary and, if so, the appropriate level at which to compute standard errors, depends on how individuals are assigned to examiners.** For example, if each individual is separately randomized to an examiner, no clustering is necessary. If individuals are assigned to examiners in batches or shifts (and individuals are not randomly

assigned to the batches), inference should be clustered at the batch or shift level (Abadie et al., 2022).

- **While the conditions for identification ultimately rest on institutional and economic foundations, specification tests can probe their empirical plausibility.** Familiar balance tests from the RCT methodology can be used to assess random assignment to examiners. Classical overidentification tests (e.g., Sargan, 1958) probe the exclusion restriction in a linear framework. A battery of recent procedures test whether exclusion and monotonicity conditions hold when effects are heterogeneous, including Kitagawa (2015), Norris et al. (2018), and Frandsen et al. (2023a). We do not recommend the common practice of screening based on whether the first-stage F-statistic exceeds a threshold value. Such screening can exacerbate distortions from weak instruments. A valid alternative is to screen on the sign of the first-stage in-sample correlation between the JIVE instrument and treatment (Angrist and Kolesár, 2024). Below we provide simulation-based evidence to support this recommendation.

The remaining sections of the paper are organized as follows. Section 2 formally introduces an econometric framework based on constant treatment effects. Our initial focus on the case of constant treatment effects provides a foundation to discuss basic issues surrounding the examiner research design. To accompany our econometric framework, we introduce a conceptual model of examiner behavior to show the relationship between basic econometric conditions and examiner decision making. In Section 3, we discuss estimation in the case of constant treatment effects. Our discussion highlights the importance of jackknife instrumental variables (JIVE) to address bias that can arise when attempting to use variation in examiner tendencies in two-stage least squares (2SLS) estimation. As previewed above, this section also highlights the need for internally consistent use of covariates in IV models. Section 4 covers inference, including guidance on clustering. Our discussion in Section 5 extends our formal framework to consider heterogeneous treatment effects and highlights the necessity of monotonicity and exclusion restrictions for identification of conventional weighted average treatment effect parameters. We also discuss key assumptions behind the estimation of marginal treatment effects and identification in settings where examiners can influence outcomes through multiple channels. Section 6 reviews empirical tests that shed light on the plausibility that key identifying conditions hold. We provide a detailed guide to implementing examiner tendency research designs by conducting a case-study analysis of the effects of pre-trial detention in Section 7. The code and data for the empirical example are available online. Finally, we conclude in Section 8 with a discussion of recent innovations in the use of examiner research designs as well as areas for future research.

2 Framework

In this section, we lay out a simple econometric and conceptual framework. We begin with a standard linear model with constant treatment effects. Although restrictive, the constant treatment effects framework provides a simple setting to discuss most of the practical issues around identification, estimation, and inference. In Section 5, we consider additional issues that arise when treatment effects are heterogeneous. As it turns out, the estimation and inference approaches we propose for the simpler constant treatment effects case carry over to the more general heterogeneous treatment effects setting.

2.1 Basic econometric model

We seek to estimate the effects of a binary treatment, such as pre-trial detention or placement into foster care, denoted by the indicator D_i. Let $Y_i(0)$ be the potential outcome if individual i is untreated, and let $Y_i(1)$ be the potential outcome if treated. Individual i's realized outcome is $Y_i = Y_i(0) + (Y_i(1) - Y_i(0)) D_i$, and the effect of treatment for individual i is $Y_i(1) - Y_i(0)$. For now, we assume treatment effects to be constant: $Y_i(1) - Y_i(0) = \delta$ for all i. In this case, the realized outcome can be represented as:

$$Y_i = \alpha + \delta D_i + \varepsilon_i, \tag{1}$$

where we define $\varepsilon_i = Y_i(0) - E[Y_i(0)]$ and $\alpha = E[Y_i(0)]$ and use $E[Y_i(0)]$ to denote the expected value of outcomes in the untreated state.

Despite the simplicity of the model, estimation of δ poses a challenge. In many settings, treatment status D_i will be related to other determinants of the outcome, here captured by ε_i. As a result, D_i will be endogenous, and ordinary least squares estimates of $\hat{\delta}$ will be biased.

We now assume that an examiner such as a judge determines each individual's treatment status and examiners may differ in their decisions. Let $J_i \in \{1, \ldots, k\}$ denote the judge to whom individual i is assigned. Let $D_i(j)$ be individual i's potential treatment status if assigned to judge j, and define judge j's propensity as $p(j) = E[D_i(j)]$. In our notation, j indexes specific judges and J_i is a random variable corresponding to the judge to whom individual i is assigned.

While examiners affect treatment status, we assume they have no other effects on outcomes—i.e., an exclusion restriction assumption. To be precise about what this means, we expand the potential outcome notation above to reflect examiner assignment. Let $Y_i(0, j)$ and $Y_i(1, j)$ be individual i's untreated and treated potential outcomes if assigned to examiner j, respectively. In words, the exclusion restriction assumption requires that changing examiner assignment from examiner j to j' does not change either of an individual's potential outcomes. Formally, this

condition can be expressed as follows:

Assumption 1 (Exclusion restriction)**.** $Y_i(d,j) = Y_i(d,j') = Y_i(d)$ *for* $d \in \{0,1\}$ *and all* $j, j' \in \{1,\ldots,k\}$ *and for all* i*.*

To identify δ in equation 1, we assume random assignment of individual i is to one of the k-many examiners. This ensures that examiners receive comparable case mixes and any differences in the probability of treatment between judges are due to differences in examiner propensities rather than differences in the individuals assigned to the examiners.[2] The random assignment assumption is formally expressed as:

Assumption 2 (Examiner random assignment)**.** $\left(Y_i(0), Y_i(1), \{D_i(j)\}_{j=1}^{k}\right)$ *are jointly independent of* J_i*.*

This assumption means that judge assignment is unrelated to an individual's potential outcomes or potential treatment status.

Random assignment to examiners means that we can identify examiner propensities as simply the average treatment status among individuals assigned to each examiner: $p(j) = E[D_i|J_i = j]$. Equivalently, if we define Z_i to be a $k \times 1$ vector of examiner indicators, we can express propensities in terms of the following regression equation:

$$D_i = Z_i'\pi + \nu_i, \tag{2}$$

where $E[\nu_i|Z_i] = 0$ by definition. The propensity of the examiner to whom individual i is assigned is given by $p(J_i) = E[D_i|Z_i] = Z_i'\pi$. The treatment residual, ν_i, captures everything that determines treatment status besides the assigned examiner. For example, if D_i were an indicator for pre-trial release, ν_i might include factors like prior criminal history, the severity of the charge, and other characteristics of the defendant that bail judges might take into consideration when deciding on release or detention. These other factors may also influence the outcome—that is, ν_i and ε_i may be correlated. For example, defendants with a prior criminal history may be more likely to be detained prior to trial, and more likely to be ultimately convicted. This correlation is why an ordinary least squares regression based on equation (1) is likely to obtain biased estimates.

The outcome equation (1) and treatment equation (2) fit into the standard linear instrumental variables framework. Given the exclusion restriction and examiner random assignment, instrumental variables estimators can consistently estimate the parameter δ provided examiners vary in their

[2] In some contexts, examiners or judges may be *conditionally* randomly assigned. For example, defendants charged with felonies might be assigned to a different set of judges from those charged with misdemeanors. In this case, the analysis should control for the covariates conditional on which judges are randomly assigned. Section 3.3 discusses how covariates may be incorporated.

treatment propensity. At a minimum, this requires that there exists at least one pair of examiners whose propensities differ from each other's, as the following assumption makes precise:

Assumption 3 (Nontrivial variation in propensities)**.** *For some* $\mu > 0$ *there exist examiners* $j, j' \in \{1, \ldots, k\}$ *such that* $|p(j) - p(j')| \geq \mu$ *and* $\min\{\Pr(J_i = j), \Pr(J_i = j')\} \geq \mu$*.*

The exclusion restriction, examiner random assignment, and nontrivial variation in propensities satisfy the traditional instrumental variables requirements of exogeneity and relevance.[3] As a result, the treatment effect δ is identified by the usual instrumental variables estimand:

$$\delta = \frac{Cov(Y_i, p(J_i))}{Cov(D_i, p(J_i))}. \tag{3}$$

Equation (3) shows that δ is identified. Note that it is not an estimator because the expression involves population covariances and true judge propensities—neither of which are observed. Section 3 covers estimation in this baseline case when the treatment of interest has constant effects. We subsequently discuss causal inference when treatment effects are heterogeneous and introduce monotonicity assumptions (which become necessary for identification when effects are not constant) in Section 5.

2.2 Conceptual model of examiner decision-making

In this section, we lay out a simple conceptual framework that models examiner decisions as a cost-benefit problem.[4] The solution to the decision problem is a threshold crossing rule that compares the probability that treatment has positive net benefit to a cutoff value. This cutoff value may vary across examiners because of differences in preferences or information. For concreteness, we frame the model in the context of judges deciding over pre-trial detention.

Let $D_i(j)$ denote judge j's decision for defendant i: $D_i(j) = 1$ when the decision is to detain, and $D_i(j) = 0$ when the decision is to release. Judges value preventing defendants from engaging in misconduct prior to trial, such as failing to show up for the trial or committing crimes between the arrest and trial. Let θ_i be a binary indicator for whether defendant i would engage in misconduct if released. Of course, not all defendants would engage in misconduct if released, and judges also value allowing defendants their freedom while they await trial. We represent judge j's preferences

[3] The non-trivial variation in propensity condition in Assumption 3 is equivalent to the standard instrumental variables relevance condition. For instance, in Imbens and Angrist (1994), the condition is defined as the assumption that the conditional expectation of treatment is a non-trivial function of the instrument. That is, $\mathbb{E}[D_i|Z_i = w]$ is non-trivial function with respect to values w in the support of Z_i.

[4] See Canay et al. (2023) for an alternative model of examiner decision-making that is based on a generalized Roy model (Heckman and Vytlacil, 2005).

over these competing values using the following utility function:

$$U_j\left(d;\theta_i\right)=\begin{cases}0 & , \quad \theta_i=0, d=0\\ -a_j & , \quad \theta_i=1, d=0\\ -b_j & , \quad \theta_i=0, d=1\\ c_j & , \quad \theta_i=1, d=1\end{cases}, \; a_j\geq 0, b_j\geq 0, \; c_j\geq \max\left\{-a_j,-b_j\right\}.$$

This utility function means that judge j incurs a cost of a_j if a defendant who would engage in misconduct is released, a cost b_j if a defendant who would not have engaged in misconduct is detained, and a benefit c_j if a defendant who would have engaged in misconduct is detained. The requirement that $c_j \geq \max\left\{-a_j,-b_j\right\}$ reflects the intuition that judges prefer correct decisions to incorrect ones.[5] We normalize the utility of releasing a defendant who would not have engaged in misconduct to zero.

If judges knew θ_i, the optimal decision rule would be clear: release if $\theta_i = 0$ and detain if $\theta_i = 1$. But judges have no crystal ball and must make do with the information they have. We denote the information that judge j has about defendant i at the time of the arraignment hearing by v_{ij}. The index j allows for the possibility that judges may differ in the information available to them or their skill at eliciting and interpreting the relevant information. We assume judges choose detention status by maximizing expected utility conditional on their observed information:

$$D_i\left(j\right)=\arg\max_{d\in\{0,1\}} E\left[U_j\left(d;\theta_i\right)|v_{ij}\right].$$

A little algebra shows that judge j will detain defendant i if the defendant's probability of misconduct, $q\left(v_{ij}\right):=\Pr\left(\theta_i=1|v_{ij}\right)$, exceeds a threshold, τ_j:

$$D_i\left(j\right)=1\left(q\left(v_{ij}\right)\geq\tau_j\right),$$

where the threshold depends on the judge's preferences:

$$\tau_j=\frac{b_j}{a_j+b_j+c_j}.$$

The threshold rule captures the intuition that judges will be more hesitant to detain defendants (i.e., they will apply a higher threshold) when they weigh the costs of detaining a defendant who would not engage in misconduct more heavily—that is, when b_j is larger. Judges who weigh the cost of releasing a defendant who engages in misconduct more heavily (larger a_j) or who value detaining

[5] In the case that $c_j < \max\left\{-a_j,-b_j\right\}$, it would mean that the judge prefers to either wrongly release or wrongly detain a defendant relative to correctly detaining a defendant.

a defendant who would have engaged in misconduct more strongly (larger c_j) will be more likely to detain defendants. A judge's propensity in this framework is

$$p(j) = \Pr(q(v_{ij}) \geq \tau_j).$$

Let's now consider the interpretation of the basic identifying assumptions in this conceptual framework of judge decision making. The exclusion restriction in this setting means that the judge's detention decision, $D_i(j)$, is the only way in which defendant i's outcomes depend on the judge assignment. It requires that judges differ in no other decision or characteristic that affects defendant outcomes. For example, if arraignment judges not only make detention decisions, but also make decisions regarding court-appointed legal representation, then the exclusion restriction would be violated if court-appointed legal representation affects outcomes.[6]

Examiner random assignment in this setting means that defendants who have particular characteristics or potential outcomes have the same likelihood of being assigned to any particular judge as defendants who have other characteristics or potential outcomes. Judge random assignment would be violated if, for example, certain judges take cases at specific times of day or days in the week, or if certain judges "specialize" in particular kinds of cases.

Finally, nontrivial variation in propensities means that judges differ in their preferences (i.e., the relative costs of releasing a defendant who engages in misconduct or detaining a defendant who would not have), information, or skill in eliciting and interpreting the relevant information. Judges must also have some degree of discretion in the treatment decision. A setting in which all judges see the same information about a given defendant and where their decisions are dictated by rules or formulas may not give rise to nontrivial variation in propensities across judges.

3 Estimation

3.1 Two-stage least squares

A natural starting place for estimation is to use two-stage least squares (2SLS) to compute the sample counterpart to the instrumental variables estimand in equation (3). The first stage, given by equation (2), can estimated by OLS:

$$D_i = Z_i'\pi + \nu_i.$$

[6]Similarly, if judges differ in their tendency to warn or verbally admonish defendants, then there could be violations of exclusion if these types of judicial behavior matter for defendant outcomes.

The resulting first-stage fitted values are $\hat{p}(J_i) = Z_i'\hat{\pi}$ and serve as an instrument for D_i in the structural equation:

$$Y_i = \alpha + \hat{\delta}^{2SLS} D_i + \hat{\varepsilon}_i,$$

where $\hat{\varepsilon}_i$ is the 2SLS residual and

$$\hat{\delta}^{2SLS} = \frac{\widehat{Cov}\left(\hat{p}(J_i), Y_i\right)}{\widehat{Cov}\left(\hat{p}(J_i), D_i\right)}.$$

As long as the conditions in a given empirical setting satisfy the identification assumptions discussed above as well as standard textbook conditions, such as independence across observations and a large number of observations per examiner, then $\hat{\delta}^{2SLS}$ will be approximately normally distributed with a mean centered on δ and standard errors that common statistical packages will readily produce.

However, an important consideration is that many applications of the examiner tendency design feature a large number of examiners and relatively few cases per examiner. The textbook approximation fails in such settings: 2SLS is no longer centered on the true causal effect δ, but is biased towards the OLS estimand (i.e., $Cov(Y_i, D_i)/Var(D_i)$). The bias of 2SLS in this case is a manifestation of the many-instruments bias documented by Bekker (1994). Under an asymptotic approximation where the ratio of the number of examiners, k, to the sample size converges to a constant, κ, the probability limit of 2SLS is

$$\hat{\delta}^{2SLS} \to \delta + \kappa \left(\frac{\sigma_{\varepsilon\nu}}{\sigma_D^2 - (1-\kappa)\sigma_\nu^2} \right),$$

where $\sigma_{\varepsilon\nu}$ is the covariance between ε_i (the error term in the outcome equation) and ν_i (the error term in the first-stage equation), and the terms σ_D^2 and σ_ν^2 are the variances of D_i and ν_i. As the number of examiners gets larger relative to the sample size (i.e., as κ approaches one) the bias of 2SLS approaches $\sigma_{\varepsilon\nu}/\sigma_D^2$, which is the bias of OLS. The approximation that $k/n \to \kappa$ is not meant to be a description of the actual data collection process or a promise about future data collection; rather, it's meant to capture better the behavior of the estimator in finite samples.

The bias of 2SLS arises with many examiners even if the conditions in a setting satisfy the standard IV assumptions (i.e., random assignment, exclusion, relevance). The bias comes from the outsized influence D_i has on $\hat{p}(J_i)$ when there are few cases per examiner. Recall that Z_i is a set of indicator variables, and the estimate $\hat{p}(J_i) = Z_i'\hat{\pi}$ is the sample average treatment status among individuals assigned to examiner J_i. Importantly, this sample average includes individual i which implies that this sample average will be correlated with D_i. This correlation will be stronger if there are fewer cases assigned to that examiner. When there are few examiners relative

to the sample size—equivalently, when there are many cases per examiner—we can safely ignore this extra correlation between D_i and $\hat{p}(J_i)$. When there are many examiners, the endogenous variation in D_i—the reason for employing an IV strategy in the first place—contaminates $\hat{p}(J_i)$.

3.2 The case for JIVE

A solution to the many-instruments bias of 2SLS in settings with many examiners is jackknife instrumental variables (JIVE, Angrist et al., 1999). JIVE cleans up the contamination in $\hat{p}(J_i)$ due to the influence of D_i by replacing it with $\hat{p}_i^{JIVE} = Z_i'\hat{\pi}_{-i}$, where

$$\hat{\pi}_{-i} = \left(\sum_{l\neq i} Z_l Z_l'\right)^{-1} \sum_{l\neq i}^{n} Z_l D_l. \tag{4}$$

In the simplest case with no covariates, $\hat{p}_i^{JIVE}$ is simply the sample average treatment status among individuals assigned to examiner J_i *besides* individual i. The JIVE estimate of the treatment effect is then the usual just-identified IV formula, using $\hat{p}_i^{JIVE}$ as a single instrument:

$$\hat{\delta}^{JIVE} = \frac{\widehat{Cov}\left(Y_i, \hat{p}_i^{JIVE}\right)}{\widehat{Cov}\left(D_i, \hat{p}_i^{JIVE}\right)}.$$

The jackknife remedy for IV bias now appears in nearly every published study using the examiner tendency design, although it usually goes by the name "leave-out mean" rather than jackknife.[7] For example, Dahl et al. (2014) estimate the leniency of the disability insurance examiner assigned to each case by calculating the examiner's tendency among all other cases assigned to the examiner. This "leave-out mean examiner propensity measure" is identical to JIVE's version of the first stage fitted value when no additional covariates are involved. More care is required when there are additional covariates (see Section 3.3).

The jackknife or leave-out procedure must be modified when individuals are assigned to examiners in clusters, such as batches or work shifts. The reason is that unobserved determinants of outcomes and treatment status—that is, ε_i and ν_i—may be correlated within clusters. If individuals i and j share a cluster, then endogenous variation from individual j's treatment status, D_j, contaminates individual i's fitted value, $\hat{p}_i$, in the usual observation-level jackknife procedure. This contamination biases JIVE towards OLS for the same reasons that 2SLS is biased. The solution is to estimate i's fitted value, $\hat{p}_i$, leaving out observation i's entire cluster, not just observation i itself, a procedure called CJIVE which is explored further in Frandsen et al. (2023b). Denoting the set of observations in individual i's cluster as $\mathcal{C}_i$, the CJIVE fitted value is defined as: $\hat{p}_i^{CJIVE} = Z_i'\hat{\pi}_{-\mathcal{C}_i}$,

[7] Over 90 percent of the studies we survey in Table A.1 used a jackknife or leave-out procedure for calculating the examiner propensity measure.

where

$$\hat{\pi}_{-\mathcal{C}_i} = \left(\sum_{l \notin \mathcal{C}_i} Z_l Z_l' \right)^{-1} \sum_{l \notin \mathcal{C}_i}^{n} Z_l D_l.$$

and the CJIVE estimator is

$$\hat{\delta}^{CJIVE} = \frac{\widehat{Cov}\left(Y_i, \hat{p}_i^{CJIVE}\right)}{\widehat{Cov}\left(D_i, \hat{p}_i^{CJIVE}\right)}.$$

Note that the CJIVE estimator requires several clusters per examiner as an examiner with only one assigned cluster would have no observations from which to estimate a cluster-jackknifed propensity. Clustered assignment to examiners also affects inference, an issue explored in detail in Section 4.1.

3.3 Covariates

It is often helpful to control for a set of covariates X_i because of the belief that conditioning is required for identification in a given setting or a desire to increase precision. For example, suppose one set of rotating judges presides over weekend arraignments, and another set over weekday arraignments. Because judges are randomly assigned conditional on weekend or weekday, the vector X_i should include a weekend indicator. Similarly, suppose that prior criminal history strongly predicts an outcome of interest. The inclusion of criminal history in X_i could improve the precision of 2SLS estimates. While these considerations motivate the use of covariates, it may be desirable to omit some factors that predict the outcome (but are not needed to ensure conditional random assignment) from X_i in order to use these in balance tests to assess whether the data are consistent with random assignment (see Section 6). In addition, note that factors that themselves may be affected by treatment or judge assignment should not be included in X_i because doing so may introduce bias into the estimator.

Researchers must make a modeling choice for covariates. One possibility is to condition nonparametrically on covariates by performing estimation separately for each covariate value. This approach spares the researcher from taking a stand on functional form, but it is only feasible for discrete covariates that take on few values and have many observations per cell. The more standard approach is to assume additive separability between the treatment and covariates. Formally, one assumes that the realized outcome satisfies:

$$Y_i = \delta D_i + X_i'\beta + \varepsilon_i, \tag{5}$$

where we re-define $\varepsilon_i = Y_i(0) - \mathbb{E}[Y_i(0) | X_i]$.

The presence of covariates complicates the leave-out or jackknife remedy for many-instruments biased discussed above. Two recent estimators adapt JIVE to the case with covariates: the unbi-

ased jackknife estimator (UJIVE) proposed by Kolesár (2013) and the improved jackknife (IJIVE) procedure proposed by Ackerberg and Devereux (2009). UJIVE proceeds as JIVE but features an important modification: the jackknifed first stage regression in equation (4) now includes covariates. Following the jackknifed first stage regression, the covariates are partialled out of the fitted values for D_i, also using jackknifed regressions. IJIVE, on the other hand, partials out covariates from the outcome, treatment, and examiner dummies prior to the jackknifed first-stage estimation of equation (4). Notably, UJIVE remains consistent even when the number of covariates is large (Kolesár, 2013), while IJIVE may not be. This theoretical edge suggests UJIVE should be considered the default estimator.[8] With either approach, researchers who employ these methods ensure that covariates are handled consistently in the first and second stages. A researcher who conditions on one set of covariates in constructing the examiner propensities and a different set of covariates when estimating effects in a second stage can unwittingly impose spurious exclusion restrictions, biasing the estimates. Both UJIVE and IJIVE adapt to the case with clustering naturally by simply replacing the jackknife regressions in both procedures with cluster-level jackknife regressions.

4 Inference

Standard errors, hypothesis tests, and confidence intervals based on the usual heteroskedasticity-robust IV variance formula provide reliable inference for standard cross-sectional data under conditions that should be satisfied in most empirical settings with examiner-based designs (Ackerberg and Devereux, 2009). The conditions include that there are a sufficient number of cases per examiner, individuals are assigned independently to examiners (as opposed to batches of individuals assigned as a group to an examiner), and examiners vary sufficiently in their propensities. Heteroskedasticity-robust variance formulas accommodate binary outcomes such as conviction or recidivism which are inherently heteroskedastic measures and appear commonly in examiner-based designs. The IV procedures built into statistical software applications like Stata produce estimates of these variances (provided the user has constructed $\hat{p}_i^{JIVE}$ as above) or the variants such as IJIVE or UJIVE for the case of designs that rely on covariates.[9]

Occasionally, however, an empirical setting may violate these conditions and inference requires more care. One concern is that the usual standard errors can mislead when there are few cases per examiner. Intuitively, the reason for this is that $\hat{p}_i^{JIVE}$ will be very noisily estimated for observations assigned to examiners with few cases, and the estimation error will be correlated across observations. One effective remedy is to restrict the sample to examiners with a sufficiently large

[8] At the same time, our empirical example described in Section 7 shows that UJIVE and IJIVE give similar results (see Table 4).

[9] This is true even though $\hat{p}_i^{JIVE}$ is estimated. Wooldridge (2010) outlines fairly general conditions under which generated instruments do not affect inference.

number of cases. For example, the case study in Section 7 restricts to bail judges with at least 200 cases. We recommend showing robustness to alternative choices of the cutoff, as we do in Table 5. Studies that do not enjoy the luxury of a large number of cases per examiner may need to employ the many-instrument adjustments to jackknife instrumental variables standard errors suggested by Evdokimov and Kolesár (2019). The next subsections discuss how to approach violations to the two other standard conditions for inference: independent examiner assignment and strong identification.

4.1 Clustering

Many applications, however, depart from the standard cross-sectional setting with independent assignment to examiners. In these cases, inference based on the usual heteroskedasticity-robust formulas could be misleading. Instead, it may be necessary to use cluster-robust inference.

Cluster-robust inference requires deciding the level at which to cluster. In the design-based framework described in Abadie et al. (2022), the level at which to cluster is dictated by the level at which assignment to examiners occurs.[10] From this perspective, the randomness that generates sampling variation in the estimates stems from the examiner assignment mechanism. That is, in hypothetical repeated samples, the estimates of the treatment effect vary because a given individual's assigned examiner can change, thereby affecting the potential outcomes that are revealed for each individual. The cluster-robust standard error formula captures the sampling variation arising from clustered assignment to examiners. For example, if all individuals in a batch or a work shift are randomly assigned to the same examiner, then inference should be clustered at the batch or shift level.[11]

By contrast, many practitioners cluster at the examiner level,[12] perhaps out of a desire to be conservative by clustering at a coarse level, or because they are positing that error terms are correlated among observations assigned to the same examiner. In the design-based approach to inference recommended by Abadie et al. (2020) and Abadie et al. (2022), the correlation structure of unobserved determinants of the outcome is irrelevant for the clustering decision. The clustering level is determined by an institutional fact: the level at which individuals were assigned to judges.

4.2 Inference and weak identification

Weak identification is another potential concern for inference in examiner tendency designs. In this setting, weak identification means examiners vary little in their propensities to assign individuals

[10]The design-based approach to inference is distinct from model-based inference. In the latter, sampling variation in estimates is governed by an assumed joint distribution of the errors terms specified by the model.

[11]Note that if inference is clustered, then the jackknife estimation should also be clustered at the same level.

[12]Notable examples include Dobbie et al. (2018) and Bald et al. (2022).

to treatment. In some IV settings, the conventional asymptotic approximations break down under weak identification and the usual standard errors can yield misleading inference (Andrews et al., 2019; Mikusheva and Sun, 2021). This section discusses when weak identification is likely to cause practical problems, and how to address weak identification in such problematic cases.

The weak identification problem is distinct from the many-instruments problem discussed in Section 3. Even if identification is strong (i.e., examiners vary substantially in their propensities), 2SLS using examiner dummies as instruments suffers from many-instruments bias. JIVE eliminates the many-instruments bias, but does not necessarily solve the weak-identification problem. It does, however, allow us to apply recent econometrics findings on how to deal with weak identification in single-instrument settings.[13]

Recent research has clarified that in single-instrument IV settings, like examiner designs using a JIVE instrument, weak identification substantially distorts estimation and inference only when the degree of endogeneity—here, the correlation between ν_i and ε_i[14]—is very high. Angrist and Kolesár (2024) show that the coverage of 95-percent confidence intervals is distorted by at most 5 percentage points no matter how weak the instrument when the degree of endogeneity is less than about 0.76. The reason is that although weaker instruments lead to more bias, they also lead to larger standard errors and wider confidence intervals. When the degree of endogeneity is high enough, however, weak identification can substantially distort inference.

The large majority of IV specifications in recently published studies exhibit degrees of endogeneity below the danger zone of 0.76. The largest estimated degree of endogeneity encountered in the studies examined by Angrist and Kolesár (2024) was 0.47. Lee et al. (2023) analyzed a broader set of studies—every single-variable just-identified IV specification published in the *American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics,* and the *Review of Economic Studies* in 2021. Out of 89 such published specifications for which they could calculate the required statistics, 75 (84 percent) had an estimated degree of endogeneity below the 0.76 benchmark.[15]

The results in Angrist and Kolesár (2024) suggest, therefore, that in most empirical settings, the usual IV standard errors and associated confidence intervals should be reliable, even when identification is weak. However, there are certainly empirically relevant scenarios where weak identification should not be ignored. What should a researcher to do in these cases? The recent econometrics literature suggests two strategies. First, Angrist and Kolesár (2024) suggest to screen on the sign of the estimated first stage. In our case, this means to proceed with the analysis only if the covariance

[13] As discussed above, and as noted in Bhuller et al. (2020), although the underlying examiner dummies are many, under certain conditions the JIVE fitted value (i.e., $\hat{p}_i^{JIVE}$) can be treated like a single instrument.

[14] This expression assumes independent data and homoskedasticity. The expression is a little more involved with heteroskedasticity or dependent data, like clustering.

[15] They estimate the degree of endogeneity via the sample correlation between first and second stage residuals.

between treatment status and the JIVE instrument is positive, i.e., $\widehat{Cov}\left(D_i, \hat{p}_i^{JIVE}\right) > 0$. This intuitive requirement cuts the weak instruments bias roughly in half. This differs from the older rule of thumb to proceed only if the first stage F-statistic exceeds 10—a point that we discuss in detail in our simulation exercises below.[16]

Second, in cases where the degree of endogeneity is very high, Lee et al. (2023) offer adjusted critical values (i.e., different from 1.96) that will ensure confidence intervals maintain their advertised coverage. The adjustments depend on the first-stage F-statistic of the single instrument and the estimated degree of endogeneity. For example, if the first-stage F were 24 and the estimated degree of endogeneity were 0.8, their adjustment delivers a critical value of 4.017 for the interval's lower bound, and 2.56 for the upper bound.[17] Alternatively, Mikusheva and Sun (2021) propose a first-stage test statistic specifically for jackknife instrumental variables estimators that can be used to determined if weak identification is a problem.

The recommendations above are supported by the theoretical analysis in Angrist and Kolesár (2024) and Lee et al. (2023). We now illustrate via simulations their empirical relevance for examiner designs. In our simulations, we create 100 judges who each assign 100 defendants to a binary treatment. We generate individual treatment status D_i and outcome Y_i variables via a simplified and parameterized version of the conceptual model in Section 2.2. Specifically, in the simulations, individual i's treatment status when assigned to judge j is generated as $D_i = 1\left(\Phi\left(v_i\right) \geq \tau_j\right)$, where Φ is the standard normal CDF, and v_i is a standard normal random variable. In terms of the conceptual model in Section 2.2, v_i represents the examiners' information about individual i's suitability for treatment, we set the function q to be Φ, and judge thresholds τ_j are evenly distributed over a range of width h centered on 0.5. Judge j's propensity to assign treatment is $p_j = 1 - \tau_j$, and thus judge propensities are also centered on 0.5 with range h. The simulations explore the consequences of weak identification by varying h. The case when h is near zero corresponds to weak identification (as there is little variation between judges). The case of $h = 1$ corresponds to very strong identification (where the least and most strict judges have propensities of 0 and 1, respectively). Defendants are randomly assigned to each of the $k = 100$ judges with equal probability. Defendant i's outcome is $Y_i = \delta D_i + \varepsilon_i$, where ε_i is a standard normal random variable. We generate ε_i to have a correlation with v_i equal to ρ, which determines the degree of endogeneity of D_i. Across all simulations, we hold the treatment effect constant at $\delta = 0.3$.

[16]Note that screening on the sign instead of the magnitude of the first-stage estimate could have implications if there are multiple screening criteria imposed in the publication process. For example, screening based on the sign alone implies that studies with less precision will "pass" an initial review. This could have implications for publication bias if reviewers also prioritize studies that reject the null at conventional statistical significance levels. Studies that produce empirical results with large standard errors will reject the null only when their estimated effects are large in magnitude.

[17]The adjustment, dubbed "VtF" by Lee et al. (2023) can be implemented in Stata by following instructions on David Lee's website: `https://irs.princeton.edu/davidlee-supplementVTF`.

Our exercise varies h from 0 to 1 in increments of 0.05 and simulates $1,000$ samples for a given set of model parameters. In each sample, we construct a confidence interval for δ based on the point estimate and standard error from each the following four procedures: (1) 2SLS using judge dummies; (2) JIVE; (3) JIVE, screening on the first-stage F-statistic exceeding 10, a common benchmark (where the F-statistic is from regressing treatment on the JIVE fitted value); (4) JIVE, screening on having a positive first-stage coefficient, an approach recommended for IV from Angrist and Kolesár (2024).

Figure 1 provides results that show how inference depends on instrument strength as well as the endogeneity specified in the data generating process. Panel A sets $\rho = 0.30$, a low degree of endogeneity, and Panel B sets $\rho = 0.60$, a high degree of endogeneity. The y-axis measures our main statistic of interest: the fraction of samples associated with each value of h in which the confidence intervals exclude the true treatment effect. The x-axis corresponds to our measure of instrument strength, the propensity range across judges.

The main result from this analysis is that JIVE, whose rejection rate is plotted with a dashed line, never over-rejects, no matter how weak the instrument. This is consistent with similar examiner tendency design simulation results provided in Bhuller et al. (2020, Appendix D) and with the theoretical analysis in Angrist and Kolesár (2024). In contrast, a naive approach of using 2SLS with judge dummies (solid line) rejects the truth at a high rate when identification is weak, an illustration of the well-known inference distortion with weak instruments (Andrews et al., 2019).

What do we observe when using the common practice of screening on the first-stage F-statistic? The short-dashed line plots the rejection rate conditional on the JIVE first-stage F-statistic exceeding 10, a standard approach to avoid weak-instrument distortion.[18] Conditioning on the F-statistic leads to a rejection rate near 50 percent when the instrument is weak even when there is a low degree of endogeneity. Recent work suggests an alternative: researchers can use the usual standard errors provided that one conditions on the sample correlation between treatment and the jackknifed instrument being positive. Angrist and Kolesár (2024) provide evidence that, after screening on the sign of the estimated first stage, inference based on the usual standard errors is reliable. Our simulations bear this out: the dash-dotted curve shows that the rejection rate when conditioning on a positive first stage stays near the nominal level, no matter how weak the instrument.

Our recommendation for practice is therefore not to screen on the first-stage F-statistic. As demonstrated above, the common practice of screening on the first-stage F-statistic exceeding 10, or any other level, is unnecessary, and can even be harmful. That said, in line with recommendations from Angrist and Kolesár (2024), there is little harm in checking that the JIVE instrument's

[18] Staiger and Stock (1997) propose a rule-of-thumb cutoff of 10 for weak instruments.

first stage goes in the expected direction.[19]

5 Extensions to the basic framework

5.1 Heterogeneous treatment effects

The recommendations for estimation and inference thus far have all been in the context of a model with a constant treatment effect. While such a model is a natural starting place, constant effects may be unrealistic in many empirical settings. In this section, we focus on the case of heterogeneous treatment effects and show that the recommendations for estimation and inference above carry through to this more realistic case. Let the treatment effect for person i be denoted by $\delta_i = Y_i(1) - Y_i(0)$. In the case of heterogeneous treatment effects, a common parameter of interest in the literature is a weighted average of treatment effects: $\mathbb{E}(w_i\delta_i)/\mathbb{E}[w_i]$, for non-negative weights w_i.

Heterogeneous treatment effects have important implications for the interpretation of the IV estimand. Recall that the IV estimand is the covariance between assigned examiner propensity and individual outcomes divided by the variance of the examiner propensity:

$$\delta_{2SLS} = \frac{\mathbb{E}\left[(Y_i - \mathbb{E}[Y_i])(\mathbb{E}[D_i|J_i] - \mathbb{E}[D_i])\right]}{\mathbb{E}\left[(\mathbb{E}[D_i|J_i] - \mathbb{E}[D_i])^2\right]}. \tag{6}$$

As discussed in Frandsen et al. (2023a), a setting which features random assignment and satisfies the exclusion restriction implies that the expression in equation 6 can be written in terms of individual level treatment effects as:

$$\delta_{2SLS} = \frac{\mathbb{E}\left[\left(\sum_{j=1}^{k} \lambda_j \left(p(j) - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)\right)\delta_i\right]}{\mathbb{E}\left[\sum_{j=1}^{k} \lambda_j \left(p(j) - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)\right]}, \tag{7}$$

where λ_j is the probability of being assigned to examiner j, $p\left(j\right)$ is the examiner propensity to treat, p is the average propensity across all examiners ($p = \sum_{j=1}^{k} \lambda_j p_j$), and $\bar{D}_i$ is person i's expected treatment status across examiners ($\bar{D}_i = \sum_{j=1}^{k} \lambda_j D_i(j)$).

From this expression, we can see that the IV estimand is a weighted average of individual treatment effects. The weight for person i is equal to the following sum across all examiners: $\sum_{j=1}^{k} \lambda_j \left(p(j) - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)$, which is proportional to the correlation across examiners be-

[19] Checking the sign of the first stage in the full sample serves a different purpose from checking that the sign of the first stage is positive in subsamples. The latter is a test of average monotonicity, discussed in more detail in Section 6. In contrast, the sign of the first stage in the full sample can only be negative if the variation in estimated jackknifed propensities is entirely driven by statistical noise, rather than differences in true propensities across judges.

tween an individual's potential treatment status and examiner propensity. As a result, the weight is largest for people whose potential treatment status is highly correlated with examiner propensity. Of course, some individuals can have a weight of zero: for example, those whom all examiners would assign to treatment (always-takers) have $\bar{D}_i = 1$ and $D_i(j) = 1$ for all j. Similarly, those who would not be assigned to treatment by any examiner (never-takers) have $\bar{D}_i = 0$ and $D_i(j) = 0$ for all j, and these individuals will again receive zero weight. In general, the possibility that some individuals will have weights equal to zero implies that the IV estimand may not capture the effects most relevant to certain policy changes (Heckman and Vytlacil, 2005).[20]

The only individuals who can receive nonzero weight are those whose treatment status is the subject of disagreement: some examiners would assign to treatment and others would not. An examiner with an above average treatment propensity ($p(j) > p$) who would assign a person to treatment ($D_i(j) = 1$) would have a positive term in the person's weight summation, as would an examiner with a below average treatment propensity ($p(j) < p$) who would not assign the person to treatment ($D_i(j) = 0$).

The IV estimand has a reasonable causal interpretation when the weights are all nonnegative. When might some weights be negative? A simple example with two examiners 1 and 2 illustrates when this could occur. Suppose that these two examiners have equal caseloads (i.e., $\lambda_1 = \lambda_2 = .5$) and the treatment propensities for examiners 1 and 2 are $p\,(1) = 0.75$ and $p\,(2) = 0.25$, respectively. This implies $p\,(1) - p = 0.25$ and $p\,(2) - p = -0.25$. Consider an individual who would be treated only by the lower propensity examiner (i.e., $D_i(1) = 0$ and $D_i(2) = 1$). In this individual's case, $D_i(1) - \bar{D}_i = -0.5$ and $D_i(2) - \bar{D}_i = 0.5$. In this scenario $\lambda_j\,(p(j) - p)\left(D_i\,(j) - \bar{D}_i\right) < 0$ for both judges, and individual i is weighted negatively in the IV estimand. This is a problem since the weighted average in equation (7) can yield values outside of the set of convex combinations of individual treatment effects if it includes some negative weights. For example, it could produce a negative value even if all individual treatment effects are positive.

A pairwise monotonicity assumption addresses exactly this kind of situation by requiring that anyone who is treated by one examiner would also have been treated if assigned to an examiner of equal or greater propensity to treat. Formally, we represent this idea as:

Assumption 4. *Pairwise Monotonicity: For all $j, \ell \in \{0, ..., k\}$, either $D_i(j) \geq D_i(\ell)$ or $D_i(j) \leq D_i(\ell)$ for each individual i.*

Pairwise monotonicity is sufficient to ensure that each person receives nonnegative weight. When pairwise monotonicity holds, all individuals who are not always- or never-takers can be divided into groups corresponding to each propensity value p. We say an individual is a p-complier

[20]For example, consider a judicial context where a large policy reform eliminates convictions or incarceration. The IV estimand from an examiner-based research design will not reflect effects for many important types of individuals affected by these policies (e.g., those for whom all examiners would always incarcerate).

if they are treated when assigned to an examiner with $p(j) \geq p$ and not otherwise. Imbens and Angrist (1994) show that identifying a weighted average of treatment effects (with nonnegative weights) among complier groups is possible under the above conditions. Imbens and Rubin (1997) extend this result a step further to show that when the exclusion restriction, examiner random assignment, and pairwise monotonicity conditions all hold, marginal effects for every p-complier group are identified.

What does the pairwise monotonicity assumption imply for the basic conceptual framework introduced in Section 2.2? Pairwise monotonicity is implied when all examiners have the same beliefs or skills at eliciting information: $v_{ij} = v_i$. Notably, this common information condition implies that all examiners have a shared ranking of individuals in terms of their likelihood of committing misconduct. In a setting with many examiners, if any two examiners disagree on where a single individual should fall in the ranking, this individual (a defier) could generate a failure of monotonicity. Practically speaking, violations of monotonicity may occur when examiners who are harsh on average may be lenient on particular groups of individuals or types of crimes due to different underlying beliefs.[21]

5.2 Heterogeneous treatment effects and heterogeneous rankings

Examiners may not always have a shared ranking of individuals in terms of suitability for treatment (e.g., because of differences in bias, information or skill).[22] This condition violates the pairwise monotonicity assumption, but 2SLS may still identify a proper weighted average of treatment effects when weaker alternative conditions hold.

A first alternative condition is "average monotonicity." This condition simply posits that the weights in equation 7 are nonnegative (Frandsen et al., 2023a). Formally, this idea is expressed as:

Assumption 5. *Average Monotonocity: For all i,* $\sum_{j=1}^{k} \lambda_j \left(p(j) - p\right) \left(D_i\left(j\right) - \bar{D}_i\right) \geq 0$.

Intuitively, the assumption is that the examiner-specific treatment status and examiner overall treatment propensity are positively correlated for each person. Equivalently, the average propensity among judges who would treat individual i must be no less than the average propensity among judges who would not. When there are only two examiners, average monotonicity is the same as pairwise monotonicity. With three or more examiners, violations of pairwise monotonicity between a pair of examiners for a given individual can be offset if there is a positive covariance between treatment status and propensity across all examiners for that individual. This condition allows for the possibility that examiners may not entirely share an ordering in terms of suitability for

[21] Consistent with this, a number of studies have documented that examiners differ in their severity behavior with respect to certain types of crimes or racial groups (Abrams et al., 2012).

[22] Imbens and Angrist (1994) pointed out that examiners may differ in their rankings if treatment decisions are based on several criteria.

treatment (i.e., v_{ij} can vary across examiners), as long as these disagreements are not so extensive as to make anyone's treatment status correlate negatively with examiner propensity. Note that the "average" in average monotonicity refers to the average relationship between potential treatment status and the propensity across examiners *for a given individual*. It is important to highlight that it is not an average across individuals.

Several models of examiner decision-making violate pairwise monotonicity, but are consistent with average monotonicity. One such model is a variant of the single-index threshold-crossing model from Section 2.2 that features some examiners engaging in taste-based discrimination by shifting their cutoffs (i.e., being less lenient) for members of a minority group. In Appendix A, we provide examples that illustrate how average monotonicity may or may not hold when there are violations of pairwise monotonicity.

While average monotonicity is plausible in more settings than pairwise monotonicity, there are limitations in terms of what parameters are identified when this condition holds. Under pairwise monotonicity, IV can identify marginal treatment effects that can be aggregated to answer a variety of policy questions (Mogstad et al., 2018). When average monotonicity holds alone, marginal treatment effects are no longer identified.

Chan et al. (2022) provide a second approach to identification that departs from pairwise monotonicity, but relies on assumptions that are more restrictive than average monotonicity. Their approach specifies a framework that features both differences in preferences (or skills) across examiners and randomness in the signal examiners receive about each individual. The latter implies there is uncertainty about the treatment status any examiner j would assign to each person i. In this framework, they define two conditions that together are stricter than the average monotonicity condition. Specifically, they define "probabilistic monotonicity": for each pair of examiners, one must have a weakly higher probability of treating all people than the other. In addition, they also define "skill-propensity independence" which requires that skill is independent of treatment propensity across examiners and probabilistic monotonicity holds for examiners with equal skill. In their empirical application, they find evidence that violations of these conditions lead to misleading 2SLS estimates, an illustration of the potential for heterogeneous treatment effects to interfere with identification.

Finally, a third weakening of the conventional monotonicity assumption is the "compliers-defiers" condition described in de Chaisemartin (2017). When this condition holds, within any pair of examiners there may exist some defiers (individuals whom the low propensity examiner would treat, but not the high propensity examiner), as long as there are at least as many compliers (individuals who would be treated by the high propensity examiner, but not the low propensity examiner) with the same local average treatment effect as the defiers. In other words, defiers can be offset by compliers with the same treatment effect. Because the compliers-defiers condition rests

on the existence of compliers with the same average treatment effect as the defiers, it may hold for some outcomes and not for others. The set of compliers whose treatment effects are captured in the 2SLS estimate ("surviving compliers") is not necessarily unique, making it potentially impossible to characterize which individuals drive the estimated effect. The compliers-defiers condition is not equivalent to conventional monotonicity in the two-examiner case, nor does it nest average monotonicity. This is because the condition allows for the existence of some people whose treatment status is negatively correlated with examiner propensity.

We expect that conditions in most applications are more likely to satisfy the average monotonicity assumption relative to the three alternatives to pairwise monotonicity described above.[23] All three approaches allow for the presence of some defiance (i.e., low propensity examiners treating people who are not treated by high propensity examiners). In a setting where skill is well-defined, a framework similar to the one adopted by Chan et al. (2022) may be useful. However, in many settings it is difficult, if not impossible, to label examiner decisions as being correct or incorrect. The compliers-defiers condition, while weaker than the conventional pairwise monotonicity condition, is still a condition that restricts the pattern of behavior between every pair of judges. Motivating the existence of this granular pattern based on contextual or institutional details may be challenging in many cases.

5.3 Marginal treatment effects

In the case of heterogeneous treatment effects, researchers are often interested in estimating marginal treatment effects (MTEs). In this section, we describe what MTEs mean in the examiner tendency setting and why they are useful. The identification of MTEs requires a monotonicity condition that holds for every pair of examiners (e.g. conventional pairwise monotonicity or the compliers-defiers condition). Therefore, as we will show, MTEs are not identified when pairwise monotonicity fails to hold.

Under pairwise monotonicity, examiners agree on the ordering of individuals in terms of suitability for treatment. MTEs describe how treatment effects vary along the suitability spectrum. Under pairwise monotonicity, we can without loss of generality assign each individual an index value U_i, distributed uniformly over $(0, 1)$ corresponding to their location on the suitability spectrum. A p-complier, defined above as someone who would be treated by any judge with $p(j) \geq p$ and not otherwise, would have $U_i = p$. The marginal treatment effect at p, defined in Heckman et al. (2001, 2006) and Heckman and Vytlacil (2007), is defined as the average treatment effect

[23] Sigstad (2023) studies judicial panels in several settings and provides empirical evidence that suggests that average monotonicity is a more realistic condition even in settings where pairwise monotonicity is violated frequently. As we note in the conclusion, further research assessing the plausibility of monotonicity conditions and magnitude of bias due to violations of this conditions remains an on-going topic for future research.

among p-compliers:

$$\delta^{MTE}(p) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,|U_i = p\right].$$

MTEs are often of interest in their own right. In the pre-trial detention example, the MTEs give the effects pre-trial detention for defendants who would always be detained ($\delta^{MTE}(0)$), for defendants who would never be detained ($\delta^{MTE}(1)$), and all defendants in between.

MTEs are also of interest because other policy-relevant parameters can be estimated as a function of MTEs. For example, integrating MTEs over the propensity range from zero to one delivers the overall average treatment effect (ATE), a parameter often coveted by researchers. Researchers following this route to the ATE should be mindful both that it relies on pairwise monotonicity and exclusion while also requiring that judge propensities span the range from zero to one. If the range of observed propensities is narrower, the estimate for ATE will implicitly extrapolate beyond the support of observed propensities.[24]

Heckman and Vytlacil (2005) show that, under pairwise monotonicity and strict exclusion, these marginal treatment effects are identified provided sufficient variation in examiner propensities. That is, the MTE is the limit of the LATE parameter as the difference in probability of treatment between two examiners goes to zero, or equivalently, the slope of the reduced form relationship between outcomes and judge propensities. To see this, consider a case in which there are just two examiners, one with a lower propensity, p, and one with propensity $p' > p$. Let Z_i be a binary indicator taking a value of 1 when an individual is assigned to the examiner with a higher propensity and zero otherwise. In this case, the local average treatment effect is identified by the Wald ratio between the two examiners:

$$\begin{aligned}
\delta^{LATE}(p', p) &= \frac{\mathbb{E}(Y_i|Z_i = 1) - \mathbb{E}(Y_i|Z_i = 0)}{\mathbb{P}(D_i = 1|Z_i = 1) - \mathbb{P}(D_i = 1|Z_i = 0)} \\
&= \frac{\mathbb{E}(Y_i|p(J_i) = p') - \mathbb{E}(Y_i|p(J_i) = p)}{p' - p}.
\end{aligned}$$

With this expression for LATE in mind, the MTE is intuitively identified by a comparison of outcomes for individuals assigned to examiners whose propensities to administer treatment are

[24] When the compliers-defiers condition in de Chaisemartin (2017) holds, marginal treatment effects for surviving compliers can be recovered, but these cannot be integrated over to estimate an average treatment effect. Because the difference in average outcomes between any two examiners reflects only treatment effects for surviving compliers, the MTEs for surviving compliers are identified. However, the defiers and the compliers that functionally cancel out the negatively weighted defiers in the estimand will never be represented in the MTE estimation, making it impossible to estimate a population average treatment effect.

close together. More formally, the marginal treatment effect is identified by:

$$\delta^{MTE}(p) = \lim_{p' \to p} \delta^{LATE}(p', p) = \frac{\partial \mathbb{E}(Y|p(J_i) = p)}{\partial p}$$

For visual intuition, consider the top left panel of Figure 2. Each point on this figure corresponds to a hypothetical examiner who assigns a binary treatment that affects a binary outcome. The horizontal axis measures $p(j)$, and the vertical axis measures the average outcomes for individuals assigned to each examiner. As discussed in Frandsen et al. (2023a), when pairwise monotonicity and strict exclusion hold, the slope of the function connecting these points is the MTE at each point. The function plotted in the bottom left panel illustrates the MTEs at each point $p(j)$. Since the outcome is binary, each individual's treatment effect and the associated MTEs must fall between -1 and 1.[25]

When pairwise monotonicity does *not* hold, the LATE estimand between a pair of neighboring examiners in terms of propensity no longer identifies a marginal treatment effect.[26] For intuition, consider the top right panel of Figure 2, which also plots propensity and average outcome values for a set of hypothetical examiners. In contrast to the top left panel, the set of points is inconsistent with pairwise monotonicity for two reasons. First, in the area of the graph marked with an "A", there are two examiners with identical propensities, but different average outcomes. If we take these to be population points (rather than estimates from a sample) and assume that random assignment to examiners and strict exclusion hold, the pattern at A implies these two examiners differ in the set of individuals that they assign to treatment, a violation of monotonicity. Similarly, the area of the graph marked with a "B" shows examiners for whom the estimated LATE between them would take on impossible values, i.e., outside the interval from negative one to one (as shown in the bottom right panel).[27] Again, under random assignment and strict exclusion, this pattern is only possible if examiners disagree on the ordering of people in terms of suitability for treatment and thereby treat non-nested sets of individuals.

As noted above, the researcher can still estimate a proper weighted average of treatment effects across all examiners as long as average monotonicity holds. However, without pairwise monotonicity, the Wald estimator between any particular pair of examiners is no longer interpretable as

[25]Continuous outcomes will only have bounds on the range of possible treatment effects if the outcome itself is bounded. For example, potential earnings may be unbounded, in which case there would be no mathematical limit to the change in earnings that a person could experience as a result of treatment. On the other hand, effects on a bounded continuous outcome will be bounded by the size of the range of the outcome. For example, if defendants charged with a certain class of crimes can only receive up to 365 days in jail, then treatment effects on jail sentence must lie between -365 and 365.

[26]Similarly, violations of strict exclusion preclude identification of MTEs.

[27]Note that the issue is not that the slope of $E(Y|p)$ takes on both positive and negative slopes; pairwise monotonicity does not imply that $E(Y|p)$ must be a monotonic function. Rather, it implies that the slope of the expected value function stay within the interval of possible treatment effect values based on the range of the outcome variable.

a causal treatment effect.

5.4 Multi-valued treatments

The canonical examiner tendency design reveals the effects of a single binary treatment. However, treatment often takes on more than two values, or examiners affect outcomes through multiple channels. For example, an arraignment judge may decide whether to assign individuals to one of three pretrial statuses: detention, supervised release, or unsupervised release (three distinct treatment categories).[28] In some contexts, the researcher may have interest in a particular channel—the *focal* treatment—while all other treatments are considered secondary. In others, several treatment channels may be observed and of interest to the researcher. In this section, we review what examiner tendency designs identify when treatment takes on more than two values or examiners affect outcomes through several channels, drawing from recent work on IV in the presence of multiple treatments.

5.4.1 Variable treatment intensity

In some instances, treatment takes on several ordered values, corresponding to variable treatment intensity or "dosage." For example, a judge may choose the amount of bail a defendant must post. Under exclusion and monotonicity conditions similar to the basic framework above, IV identifies a weighted average of individual-level responses to a one-unit increase in treatment, or the *average causal response* (Angrist and Imbens, 1995). The monotonicity condition adapted to this setting means that for any pair of examiners, one examiner always assigns individuals to at least as high a treatment level as the other. The average causal response identified by IV puts positive weight on individuals whose treatment level would vary across examiners—a generalization of compliers.

Estimation and inference proceeds much like that for the effects of a binary treatment. 2SLS, or its jackknife variants described above if there are many examiners, where examiner dummies serve as excluded instruments for the endogenous treatment intensity variable, produces consistent and asymptotically normal estimates for the average causal response. The "propensities" estimated in the first stage would no longer be judge-level probabilities of treatment, but judge-level expected values of treatment.

[28] Examiners may also assign individuals to overlapping treatment categories. In the arraignment context, the judge may decide whether to assign individuals in criminal cases to pretrial detention as well as determining whether they are eligible to be represented by a public defender. If treatments are overlapping, we can always define exclusive treatment categories (e.g., detained without a public defender, detained with a public defender, released without a public defender, and released with a public defender).

5.4.2 Multiple channels

In the cleanest applications, examiners influence outcomes through a single channel. However, in some settings, examiners make multiple decisions that could impact individual outcomes (e.g., a judge setting bail and deciding whether to appoint a public defender, as discussed above). This section highlights how the presence of such multiple channels threatens the validity of examiner designs and may render credible causal inference impossible. We also discuss conditions under which additional channels do not bias IV estimates for a focal treatment of interest, as well as the conditions under which additional channels can be accounted for in the estimation. The required conditions that we highlight are stringent, however, and may not hold in many settings.

When multiple treatment categories exist, researchers can take one of two approaches. One strategy is to define treatment using a single binary category. Returning to our pre-trial example, a researcher could solely define their treatment as an indicator for being detained pretrial and ignore public defender assignment. This approach effectively collapses the data into two groups although there are four distinct categories of defendants based on whether or not individuals are detained pre-trial or receive a public defender. We stress that researchers should keep in mind that this decision may have consequences. Most importantly, multiple channels can cause violations in the exclusion restriction that bias IV estimates if the judge decisions across multiple channels are systematically correlated. Concretely, if judges who are more likely to release defendants pre-trial are also more likely to appoint a public defender, then differences in average outcomes across judges with high and low propensities to release defendants are potentially contaminated by the additional effects of appointing a public defender.

When can researchers safely estimate the effects of a single binary treatment despite the presence of other channels or treatment categories? If examiners' influence on outcomes through any additional channels is uncorrelated with their propensity to assign the focal treatment—a condition dubbed *average exclusion* in Frandsen et al. (2023a)—then IV estimates still identify the effect of interest.[29] Average exclusion is a strong condition and needs justification on a case by case basis. If the additional channels—for example, appointment of a public defender—are observed, then researchers can provide empirical support for average exclusion by checking if examiners' propensities for the additional channels are uncorrelated with their propensity for the focal treatment.

Another approach to causal inference in settings with multiple examiner decisions is to explicitly define each channel as a distinct treatment. This may be necessary in the absence of a compelling argument for average exclusion *or* when the effects of all channels are directly of interest. Doing so requires that additional identifying conditions hold. The classical approach posits

[29] Kolesár et al. (2015) discussed a similar condition and showed identification in a constant effects framework.

that the outcome depends on treatments linearly with constant effects. Linear 2SLS using examiner indicators as instruments for multiple endogenous variables can identify those effects relative to the omitted treatment category. This is possible as long as examiners are randomly assigned and there is sufficient variation in examiners' propensities to assign the various treatments.[30]

As discussed in Section 5, the constant effects condition can be relaxed in the case of a single binary treatment, as long as a monotonicity condition holds, and IV identifies a local weighted average treatment effect among compliers (Imbens and Angrist, 1994). Is the same true for multiple treatments? That is, can linear IV with multiple endogenous variables identify proper weighted averages of heterogeneous treatment effects?[31] Ongoing work shows that the answer is yes, but the additional conditions required may be difficult to justify in most examiner design settings. In particular, Bhuller and Sigstad (2022) give the conditions under which linear 2SLS with several endogenous treatments recovers proper weighted averages of treatment effects, and Humphries et al. (2023) discuss contexts when a conventional 2SLS approach that controls for non-focal propensities can identify causal effects of the treatment of interest.

Identification in settings with multiple treatments places stringent conditions on examiner decision making. Appendix B describes the conditions required in the Bhuller and Sigstad (2022) and Humphries et al. (2023) frameworks. In general, the two frameworks are distinct, but we illustrate the stringency of the conditions in a special case where they are equivalent: that of three mutually exclusive unordered treatments, indexed by $\{0, 1, 2\}$, and three examiners, also indexed by $\{0, 1, 2\}$. For example, judges in some settings may choose between assigning criminal defendants to probation, paying a fine, or rendering community service. One can define treatment effects for each "margin" of interest based on comparing potential outcomes under each treatment d relative to a reference treatment, which we denote by 0. Formally, we represent this quantity as: $\delta_i^{0\to d} := Y_i(d) - Y_i(0)$, where $Y_i(d)$ is individual i's potential outcome under treatment d. The conditions proposed in Bhuller and Sigstad (2022) and Humphries et al. (2023) both restrict how each examiner's treatment assignment decisions may differ from a reference examiner, whom we also index by 0.[32] The reference examiner may assign individuals to any of the three treatment categories. Examiner 1, however, may differ from the reference examiner only in that some of the individuals assigned to treatment 0 by the reference examiner may be assigned to treatment 1 by

[30] Equivalently, one can add examiner propensities for non-focal treatments as controls; the IV coefficient on the treatment of interest will be the same as if one instrumented simultaneously for all treatments using examiner indicators.

[31] IV methods beyond linear 2SLS can identify treatment effects in the discrete choice models discussed by Heckman et al. (2006), Heckman and Pinto (2018) and Lee and Salanié (2018). We focus on what 2SLS can identify in the examiners design.

[32] Humphries et al. (2023) does not explicitly define a reference examiner in its framework. However, in the just identified case, the conditions there imply the existence of a reference examiner. We demonstrate this point formally in Appendix B.

examiner 1. Similarly, examiner 2 may differ from the reference examiner only in that some of the individuals assigned to treatment 0 by the reference examiner may be assigned to treatment 2 by examiner 2.

Intuitively, when the above restrictions on examiner treatment assignment hold, any difference between the average outcomes of individuals assigned to examiners 1 and 0 reflects only the fact that some individuals receive treatment 1 rather than treatment 0; similarly, any difference in outcomes between individuals assigned to examiners 2 and 0 is due to the fact some individuals receive treatment 2 rather than treatment 0. In this way, the researcher can identify proper weighted averages of $\delta_i^{0\to d}$ using 2SLS by defining indicators D_{di} for each treatment category d that are equal to one if treatment assignment is equal to d (and zero otherwise), and instrumenting for these indicators using the examiner dummies (omitting a reference examiner).

Finally, while the results from Bhuller and Sigstad (2022) and Humphries et al. (2023) are both helpful for understanding multiple treatments and examiner tendency designs, it is worth noting two limitations highlighted by their discussions. First, as our example above demonstrates, the requirements can be limiting in terms of implications for examiner decision-making patterns. In line with this, only restrictive models of examiner decision-making can satisfy their identifying assumptions. For example, threshold crossing models with a single unobservable to determine treatment can be sufficient for identification (see Bhuller and Sigstad (2022) for a more detailed discussion); however, these models are restrictive since they assume judges share a common ranking of individuals in terms of their suitability to receive the treatments being considered. Second, the condition that one treatment category serves as the reference treatment should not be viewed as an arbitrary choice. For example, consider the sentencing judge choosing between probation, fines, or community service. In our simplified just-identified setting, the researcher must identify one of these three punishments as a reference treatment—meaning that, for every defendant about whom judges disagree over the appropriate punishment, the disagreement can only be between two treatments, with one of the two preferred options always being the reference treatment.

6 Specification testing

Causal interpretation of estimates from an examiner tendency design relies on several identifying conditions holding. As detailed above, these include the (conditional) random assignment of individuals to judges or examiners, meaningful variation in the propensity of examiners to assign individuals to treatment and exclusion restrictions whereby examiners only influence outcomes through treatment assignment. In addition, when there are heterogeneous treatment effects, the design also relies on monotonicity conditions that place restrictions on how individual treatment assignment varies across examiners. When any of these conditions are violated, IV may fail to

identify causal effects and estimates may be misleading. For example, if examiners are not randomly assigned, then IV estimates may reflect selection differences across examiners that are correlated with treatment propensity. If the monotonicity conditions are violated, IV may identify an improper weighted average of treatment effects where some weights are negative. In some cases, this may imply that the IV estimand is the *opposite sign* of the true causal effects.

The primary identification arguments for examiner tendency designs should be based on institutional and economic reasoning. At the same time, recent advances in the literature provide a range of empirical tests that can shed light on violations of the identifying conditions in a given setting. In this section, we describe four approaches to testing identifying conditions in examiner tendency designs.

- **Assessing random assignment:** Researchers can use conventional balance tests from the RCT literature to assess the plausibility of random assignment of examiners. One approach is to regress the examiners' treatment propensities on a vector of observed characteristics and test for their joint significance. Another is to run a series of regressions with observed characteristics on the left-hand side and examiner indicators on the right-hand side and test for the joint significance of the examiner indicators. These two approaches differ in the violations of random assignment they have statistical power to detect. For example, a regression of an observed individual characteristic on the examiner propensity (instead of examiner indicators) will have greater statistical power to detect violations of random assignment that are correlated with the examiner propensities. However, one may want a test that also has power to detect violations that are uncorrelated with examiner propensities if the analysis will be leaning on the stronger strict exclusion and pairwise monotonicity conditions. In this case, one should regress pre-treatment characteristics on the set of examiner dummies.

- **First-stage diagnostics:** When examiners vary little in their treatment propensities, IV estimates from an examiner-tendency design can be biased and confidence intervals misleading. Researchers traditionally gauge the strength of the instruments by the partial F-statistic from a regression of treatment on the instruments, often following the $F > 10$ rule of thumb (Staiger and Stock, 1997). As detailed in Section 4, there are pitfalls to this approach in applications of examiner tendency designs. First, when there are many examiners and the instruments are taken to be examiner indicators in a 2SLS procedure, the F-statistic can be a misleading guide to instrument strength (Hansen et al., 2008). Second, conditioning on the first-stage F-statistic—that is, some researchers may be tempted to discard results that do not pass the $F > 10$ test—distorts inference on the second-stage treatment effects, as we showed in Section 4. We therefore do not recommend that researchers condition on the first-stage F-statistic. Instead, we recommend using a jackknife-IV estimator (IJIVE or CJIVE)

and applying the recently proposed approach by Angrist and Kolesár (2024) that suggests conditioning on the sign of the estimated first stage relationship between treatment and the jackknifed instrument. They show that conditioning on a right-signed estimated first stage reduces weak-instrument bias without distorting inference—a pattern that our simulation evidence presented earlier also bears out.

- **Testing exclusion and monotonicity conditions:** Thus far, the literature recommends two types of tests. First, researchers can jointly test whether the conventional strict exclusion and pairwise monotonicity conditions hold using the test described in Frandsen et al. (2023a). As discussed in Section 5.3, this test relies on the fact that these conventional assumptions imply that individual outcomes averaged at the examiner level should be a continuous function with bounded slope of the examiner-level treatment probability ("propensity"). Intuitively, the test asks whether the sample examiner-level mean outcomes and propensities are consistent with population examiner-level average outcomes and propensities that satisfy the bounded slope condition for each pair of examiners. Second, the weaker average monotonicity condition can also be tested using a procedure suggested in Frandsen et al. (2023a), which amounts to checking whether first stages within observable subgroups are positive.

- **Estimating effects of multiple channels:** Researchers must be careful to account for the presence of multiple treatments in some settings. In a setting that features constant treatment effects, Section 5.4 notes that it is possible to instrument for multiple treatments simultaneously using examiner indicators and recover an estimate of the effect of each treatment relative to the omitted treatment category. If researchers believe constant treatment effects may be plausible in their setting, Sargan's (1958) test of overidentifying restrictions can be helpful. The Sargan overidentification test can be implemented using pre-existing statistical software packages that estimate an IV model where all observed treatments are endogenous variables and the set of examiner indicators are instruments. The testing procedure is based on a regression of the second-stage residuals on examiner indicators, and assessing the joint significance of the examiner indicators. Rejections of the null hypothesis are consistent with violations of constant treatment effects for the endogenous treatments. In the case of three treatment categories, one can see this test as being equivalent to testing whether the sample examiner-level average outcomes and propensities are consistent with the population examiner-level average outcomes and propensities lying on a plane—an implication that holds with constant treatment effects. Testing the conditions required for identifying the effects of multiple channels when those effects may be heterogeneous is still an active area of research (e.g. Bhuller and Sigstad, 2022; Humphries et al., 2023) and established best practices have not yet emerged.

7 Case study: Effects of pre-trial detention

In this section, we provide a concrete guide to implementing our suggested best practices using an empirical example that analyzes the effects of pre-trial detention on conviction. This exercise uses an examiner tendency design in which the decision-makers of interest are bail judges. The code and data for the example are available online. As in Dobbie et al. (2018), we use a sample of court records from misdemeanor and felony cases in Miami-Dade County, Florida over the period 2006-2014. Following arrest, defendants in Miami-Dade were brought to a police station where they could secure pre-trial release by posting bail that varied based on the seriousness of their offense. The 70 percent of defendants who do not immediately post bail appeared at bail hearings. The bail judge at the hearing could change the bail amount or impose additional conditions.

As described in Dobbie et al. (2018), multiple bail judges preside over cases that appear throughout the week in Miami-Dade. Judge assignment typically occurs within 24 hours of arrest, and varies based on the crime category (misdemeanor or felony) and whether hearings are scheduled during weekdays or weekends. While weekday cases are handled by a single judge, weekend cases are handled by a rotating cadre of judges. As a result, defendants scheduled during the same court "shift" (i.e., all cases in a crime category on a given calendar date) would appear before the same bail judge. There is little scope for manipulating judge assignment given the short window between arrest and hearings. Bail hearings are unrelated to the process of trial judge assignment so there is no mechanical relationship between the pretrial hearing process and later stages of a case.

We use data from court records, which include information on arrest charges, the identities of bail judges, bail amount and type, if and when bail was posted, as well as defendant characteristics such as name, gender, race, date of birth, and address. The identifying information for defendants allows us to link records and observe whether an individual has a prior criminal case during the sample period ("prior offenders"). The data also indicate whether the defendant is ultimately convicted for their case, the main outcome of interest.

For our analysis, we follow Dobbie et al. (2018) and restrict our attention to cases assigned to a weekend bail hearing because these are cases where bail judges are assigned based on a rotating schedule. In the main analysis, we restrict the sample to cases which have a bail judge who presided over at least 200 bail hearings during our sample period. Examiners with a small number of observations have noisily estimated propensities. Removing examiners who make decisions for only a small number of cases can therefore increase the precision of the estimates since this removes observations for which the first stage is relatively weak.[33] These restrictions leave 91,282

[33]There is no objective criteria for choosing the minimum number of cases per examiner, so we recommend (1) choosing a minimum caseload that is not excessively restrictive in the study setting and (2) demonstrating that changing the minimum caseload does not alter the results substantially as a robustness check.

cases, presided over by 146 unique judges.

Table 1 reports summary statistics for our analysis sample. The first column shows that the sample is mostly male and split roughly evenly between Black and non-Black defendants. Columns 2 and 3 show that defendants who are released prior to trial are more likely to be white and less likely to have a prior offense (in the past year) relative to those who are detained.[34] After their bail decisions have been made and their case is heard, the released defendants are also less likely to be convicted. These differences are consistent with the hypothesis that pretrial release affects case outcomes, although the differences in demographics and previous criminal histories motivate the need to go beyond simple comparisons between released and detained defendants.

We are interested in studying the causal effects of pretrial release, as represented in the following model:

$$Y_i = \delta Released_i + X_i'\beta + \varepsilon_i, \tag{8}$$

where Y_i is a post bail hearing outcome for individual i, $Released_i$ is an indicator for whether the individual was "treated" by being released within three days of the bail hearing, and X_i is a vector of court-by-year-by-day-of-week fixed effects, which we refer to as "court-by-time fixed effects." The court indicator distinguishes between felony and misdemeanor cases. The inclusion of X_i helps bolster the plausibility of the judge random assignment condition. Other case characteristics are omitted from X_i initially in order to use them for balance tests, as discussed in Section 3.3.

A key concern is that OLS estimates from equation (8) may be biased if there are unobserved factors that are correlated with both pretrial release and post treatment outcomes such as whether the defendant was convicted or commits a new crime in the future. For example, one possibility is that bail judges may be more likely to release more advantaged defendants who may have the lowest likelihood of committing a new crime in the future. In this case, OLS estimates will be biased toward a finding that pretrial release lowers future criminal activity.

To credibly estimate the causal effects of pretrial release, we employ an IV strategy based on bail hearing judge assignment. As noted above, our setting implies that defendants are *conditionally* randomly assigned to judges. Specifically, we assume that within covariate (in our case, court-by-time) cells, shifts were randomly allocated among the set of judges who had the potential to be assigned to that cell. Our court-by-time fixed effects allow for the possibility that some judges may not be available for shifts in all years, may not be present on particular weekend days, or may work primarily in one court. Because defendants are assigned to judges in shifts, not individually,

[34] Defendants released before trial are also more likely to have a felony or violent offense. This finding that released are associated with more severe offenses is potentially due to the fact that the likelihood of failing to appear at one's trial court date is a key judicial criteria for pre-trial release. Given this objective, released defendants may not necessarily be a group of defendants who are associated with more severe charges.

we cluster at the shift level and use the CJIVE estimator, which constructs a judge leniency measure excluding all defendants in the same cluster and handles covariates appropriately as described in Section 3.

To construct the CJIVE instrument, we first compute two sets of residuals from regressing the treatment variable ($Released_i$) and judge indicator variables on the vector of covariates X_i. We then regress the residualized treatment variable on the residualized judge indicators, leaving out one cluster (shift) at a time.[35] We form our estimated leniency meausure, $\hat{p}_i$, from the predicted values of residualized treatment for defendants in the omitted cluster in each regression.

The histogram in Figure 3 shows that there is meaningful variation in this leniency measure. In addition, the shape of the figure demonstrates that there is a substantive first-stage relationship between the instrument ($\hat{p}_i$) and the likelihood of pretrial release ($Released_i$). A simple linear regression shows that defendants are 3 percentage points more likely to be released pretrial if they were assigned to a judge whose estimated release rate was 10 percentage points higher.[36]

As detailed in Section 5.2, IV estimates of the parameter δ can be interpreted as a weighted average of the causal effects of pre-trial release when there is treatment effect heterogeneity and the conditions of instrument exogeneity, exclusion, and monotonicity hold. Notably, this parameter represents causal impacts among the subset of complier defendants who would be released by lenient judges but not by strict judges. As discussed above, we next undertake a series of exercises to shed light on the plausibility of the usual identifying conditions invoked in applications of judge research designs.

Balance tests support the idea that defendants in this setting are conditionally randomly assigned to judges working a given shift. As a benchmark, Table 2, column 1, reports results from a linear probability model with pre-trial release, the endogenous "treatment" variable of interest, specified as the dependent variable and the independent variables include defendant and case characteristics as well as court-by-time fixed effects. These statistically significant estimates demonstrate that defendants who do and do not receive pre-trial release still have observable differences in baseline characteristics even after controlling for court-by-time fixed effects. In column 2, the dependent variable is the cluster jackknifed measure of judge leniency; in contrast to column 1, these results show that the vector of defendant and case characteristics (which, crucially, were *not* included in the covariates used during the construction of the CJIVE instrument) have no significant joint predictive power for the leniency instrument's value.

[35] It is straightforward to use standard statistical program such as Stata to construct the CJIVE estimator. When assignment to judges is not clustered, researchers can run UJIVE or IJIVE using the "manyiv" package available at `https://github.com/gphk-metrics/stata-manyiv`. Note that none of the variables should be residualized prior to running any of these programs.

[36] The first-stage slope on a 2SLS (i.e., non-jackknifed) fitted value would be one mechanically; the smaller slope here on the jackknifed fitted reflects sampling uncertainty given the finite number of shifts assigned to each judge and the fact that $\hat{p}_i$ is an out-of-sample prediction.

Next, we assess the exclusion restriction and pairwise monotonicity conditions. The exclusion restriction in our setting may be violated if bail judges influence case outcomes through secondary channels like appointment of a public defender. Pairwise monotonicity may be violated if bail judges who are stricter overall would nevertheless release some defendants whom more lenient judges would detain, perhaps because more lenient judges may be stricter for particular groups of defendants. The joint test proposed by Frandsen et al. (2023a) can detect these types of violations. As noted in Section 6, the test examines slope restrictions on the relationship between the judge-level expected values of the outcome and treatment. When we implement the test using the Stata package *testjfe* and specifying conviction as the post bail hearing outcome of interest, we reject the null that strict exclusion and pairwise monotonicity both hold at the one percent level.

Figure 4, generated using the *graph* option on the *testjfe* command, provides intuition for the results of the test of slope restrictions in our sample. Each point corresponds to a judge and shows the share of defendants that they see in bail hearings who go on to be convicted (y-axis) along with their estimated propensity to release defendants pretrial (x-axis). After fitting a flexible function to these points, the test checks two conditions implied by strict exclusion and pairwise monotonicity: (1) whether the slope of the fitted function is impossibly large because it exceeds the range of possible treatment effects sizes, and (2) whether the judge assignment has significant explanatory power for the outcome after accounting for each judge's predicted point on the fitted function. Intuitively, we can think of the fitted function as mapping out a set of candidate *population* points—combinations of true propensity to treat and true average outcomes across judges—and the testing procedure as assessing whether the candidate population points imply impossibly large treatment effects and if the distance from the empirical points to the fitted function is consistent with sampling variation.

In our example, the test rejects the null hypothesis that strict exclusion and pairwise monotonicity conditions both hold because judge assignment has significant explanatory power for outcomes even after accounting for the judge's treatment propensity. For comparison, Figure 4 also shows a set of simulated points, generated by assuming the estimated function (solid maroon line) is the true data generating process and adding sampling variation to generate each data point; these points show how such a graph might appear when exclusion and monotonicity are satisfied. Since sampling variation will be larger in settings with fewer cases per judge and smaller with more cases per judge, both the distance of the points from the line and the underlying sample sizes are relevant to whether the test will reject the null.

Given these results, either the strict exclusion condition or pairwise monotonicity (or both) are likely to be violated in our setting. As discussed in Sections 5.2 and 5.4, the weaker average exclusion and average monotonicity conditions may be more likely to be satisfied in this setting. These alternative conditions also mean IV estimates using our judge leniency instrument identify

a proper weighted average of complier causal effects.

How plausible are these alternative identifying conditions? As noted in Frandsen et al. (2023a), two exercises can provide evidence on the validity of both the average exclusion and average monotonicity conditions. First, average exclusion can be assessed by examining the correlation between the judge-level propensity for pre-trial release and the alternative judge-level channels that are observed. Average exclusion implies these correlations should be zero. Second, the average monotonicity condition requires that the covariance between judges' covariate-specific treatment propensity and the judges' overall propensity is nonnegative. This implies that the first-stage coefficient on the jackknifed fitted value is positive within each group defined by baseline characteristics.

While we lack data on alternative judge-level channels to test average exclusion, Table 3 provides results from our assessment of average monotonicity. We report first-stage results for a variety of subgroups of defendants and find that release status is consistently positively correlated with the judge leniency instrument. As we find no evidence in violation of the condition of average monotonicity, we move forward and interpret IV estimates using our CJIVE instrument as a local average treatment effect of pretrial release on conviction.

Our main results on the effects of pre-trial release are reported in Table 4. Columns 1-4 provide a set of benchmark results. We begin with OLS estimates of equation (8) in column 1. This descriptive result indicates that being released is associated with a 23.2 percentage point reduction in the probability of conviction. The next three columns turn to the IV results: Column 2 reports 2SLS using the vector of judge dummies as excluded instruments, and Columns 3 and 4 report IJIVE and UJIVE, jackknifing at the individual level. These point estimates are notably larger in magnitude than the OLS results. The final two columns report our preferred results which use the CJIVE estimator to leave out each defendant's cluster (shift) in the calculation of the judge leniency measure. The point estimate in column 5 indicates that pretrial release reduces conviction rates by 43.1 percentage points. The 95-percent confidence interval around the point estimate is wide, stretching from -83 to -3 percentage points. The inclusion of additional covariates in column 6 yields a point estimate of -0.498 with a somewhat more narrow confidence interval.[37]

As expected, the results in Table 4 demonstrate that the choice of estimator matters. The OLS result appears to substantially understate the impact of release on defendant convictions, as does 2SLS which is biased towards OLS when there are many judges. In addition, the pattern of results shows the bias of IJIVE and UJIVE toward OLS in the presence of clustered treatment assignment. As noted in Section 3, when defendants are assigned to judges in groups or shifts, it is possible for each defendant's characteristics, both observed and unobserved, to be systematically related

[37] As one point of comparison, Dobbie et al. (2018) also find that pretrial release has a significant negative impact on the likelihood of conviction, although the magnitude of their estimate is smaller.

to the characteristics of other defendants in their cluster. This implies that a defendant's potential outcomes may be correlated with the treatment status of other defendants within the same cluster. In our setting, one possibility is that a group of defendants arraigned in the same weekday hearing shift could have correlated characteristics because the non-random deployment of police across the city over time leads to defendants with similar backgrounds being arrested on the same day.

Why does the choice of estimator matter? In addition to endogeneity, the IV estimates likely differ from OLS estimates due to the fact that our preferred IV estimates represent causal impacts among the subset of complier defendants. Following standard practice, Column 4 of Table 1 summarizes compliers in our sample in terms of their average case and defendant characteristics. As noted in Abadie (2003), complier weighted averages for characteristics or potential outcomes can be estimated using an IV model where the interaction between the characteristic of interest and the treatment indicator is specified as the dependent variable of interest.[38]

The key finding from this descriptive exercise is that compliers have cases that are typically less severe and involve lower-level offenses relative to average. Relative to the sample average, Table 1 shows that compliers are charged with fewer offenses (1.01 vs. 1.63), have a much lower likelihood of being charged with a felony offense (0.28 vs. 0.52), and are much less likely to be charged with a violent crime (0.05 vs. 0.21).[39] In addition, the last row of Table 1 reports the estimated share of compliers who would be convicted if they had not been released, revealing that 97 percent would be convicted in this "untreated" state. The fact that nearly all compliers would be convicted when they are not released is consistent with the idea that many defendants prefer a plea deal (which results in conviction) for their low-level crime to a stay behind bars of indeterminate length while they await trial.[40]

As a final exercise, we conduct sensitivity analysis in our pre-trial release setting. Virtually all researchers using an examiner tendency design will choose to exclude observations from examiners who see relatively few cases. In our main analysis sample, we exclude cases assigned to judges who held fewer than 200 bail hearings. Of course, the decision of exactly what cutoff to specify is subject to discretion, so we recommend demonstrating that results are robust to varying the

[38] In such a model, the resulting IV estimate for the coefficient on the treatment variable is the complier-weighted average of the variable in the treated state. Note that it is also possible to estimate complier weighted averages using the interaction between the characteristic of interest and an indicator for not being treated as the dependent variable in an IV model. Table 1 uses both approaches with our preferred IV specification and averages the results.

[39] Dobbie et al. (2018) use data from Miami and Philadelphia and present similar estimates of complier characteristics. As we noted above with the statistics on released and detained defendants, the finding that compliers are associated with relatively low-level offenses is potentially due to the fact that the probability of failure to appear in court is a key judicial criteria for pre-trial release.

[40] For defendants charged with misdemeanors in our sample, cases where the defendant was released take about three times as long to resolve as those where the defendants were detained (152 days vs. 49 days), consistent with the possibility that detaining people faced with low-level charges pretrial induces them to accept plea deals relatively quickly.

minimum allowable cases per judge. In Table 5, we estimate our preferred specification (see column 5 in Table 4) using various minimum numbers of cases per judge to construct the sample. The estimate in column (2) of Table 5 is somewhat larger than the estimate from our preferred specification, but overall the results are not sensitive to varying this analysis sample inclusion criterion.

8 Concluding remarks

Random assignment to examiners who vary in their tendency to administer treatments or other interventions provides researchers with opportunities to evaluate policies in a range of contexts. The credibility of an examiner-based research design hinges on the institutional and contextual features that assign individuals to the examiner. Moreover, the interpretation of results from examiner-tendency approaches rests on a number of supplemental identifying conditions holding and the appropriateness of various implementation decisions. In this review article, we highlight best practices regarding estimation and inference in examiner-based IV strategies and motivate these choices in an econometric framework.

We conclude by highlighting areas where active methodological research on examiner tendency designs will continue to refine best practices. One such area of active research quantifies violations of the monotonicity conditions that are key to examiner designs and assesses the magnitude of any resulting bias. Sigstad (2023) is one recent study that makes progress in this direction. Specifically, he provides novel large-scale evidence on the extent of monotonicity violations by studying four judicial settings where it is possible to observe panels of judges making decisions over the same case. Intuitively, he tests for violations of monotonicity by examining disagreements when judges serve on panels. To illustrate, imagine that one judge is more strict than another in an initial case where they are both assigned, but the reverse is true in a subsequent case. In this scenario, the decisions in one of the cases must violate monotonicity. His analysis finds that pairwise monotonicity is frequently violated in all the settings that he considers and is difficult to detect using the standard monotonicity tests described in this guide. However, his analysis also shows that violations of the less stringent average monotonicity condition are much less frequent and the negative IV weights associated with cases violating average monotonicity are small. These results provide some reassuring evidence that the bias in 2SLS estimates due to violations of the traditional monotonicity condition may be small, at least in some settings.

Finally, the thorny problem of multiple treatments with heterogeneous effects is a focus of active econometric research. In such settings, recent research has highlighted that linear 2SLS with multiple endogenous variables can identify a positively weighted average of treatment effects only when relatively strong conditions on examiner decision-making hold. Recognizing the limitations

of conventional 2SLS approaches in settings with multiple treatments, several frontier empirical studies such as Humphries et al. (2023), Rivera (2023) and Kamat et al. (2023) combine examiner-based variation in tendencies with novel estimation approaches—often based on structural models of examiner-decision making—to estimate the causal effects of multiple treatments. A useful avenue for future research is the development of empirical tests of the validity of their identifying assumptions.

9 Tables and figures

Table 1: Pre-trial Detention Case Study: Defendant-level Summary Statistics

	Unweighted			Estimate
	(1) Full sample	(2) Released	(3) Detained	(4) Complier-Weighted
Male	0.84	0.79	0.87	0.69
Black	0.52	0.50	0.54	0.56
Age at Bail Decision	35.67	33.98	36.49	35.56
Prior offender	0.56	0.43	0.63	0.54
Number of Offenses	1.63	1.59	1.66	1.01
Felony Offense	0.52	0.56	0.51	0.28
Any Drug Offense	0.29	0.30	0.28	0.23
Any Violent Offense	0.21	0.32	0.16	0.05
Any Property Offense	0.36	0.25	0.42	0.24
Any Guilty Offense	0.59	0.41	0.67	0.96
N	91,282	29,870	61,412	91,282

Notes: This table provides summary statistics for defendants included in our analysis sample. The first column reports overall means for the listed variables described in each row. The second column reports means for the subsample of defendants released pretrial, and the third column shows means for the subsample of defendants detained pretrial. The fourth column reports estimates of complier-weighted means. We follow the approach from Abadie (2003) and detailed in Section 7 to estimate complier weighted averages using our CJIVE measure of judge leniency and our preferred IV specification. The last row presents complier-weighted averages for conviction, which is a post-treatment outcome variable. For this measure, Column 4 reports the estimated complier-weighted mean in the untreated state (i.e., the share of compliers who are convicted when they are not released pretrial). Standard errors clustered at the shift level are presented in parentheses.

Table 2: Assessing Balance

	(1) Treatment	(2) Leniency measure
Male	-10.100	-0.019
	(0.404)	(0.033)
Black	-2.254	-0.027
	(0.296)	(0.025)
Age	-0.310	-0.002
	(0.012)	(0.001)
Prior offender	-17.359	-0.004
	(0.308)	(0.024)
Number of counts	-2.137	0.015
	(0.131)	(0.011)
Felony charge	24.597	-0.586
	(9.818)	(0.643)
Drug charge	2.210	0.055
	(0.426)	(0.038)
Violent charge	14.587	0.012
	(0.419)	(0.033)
Property charge	-11.424	0.052
	(0.374)	(0.029)
Joint F-stat	1,160.001	1.348
p-value	0.000	0.207
N	91,282	91,282
Mean of dep. var.	0.327	0.000

Notes: This table reports results from a balance test analysis using the sample constructed to study the effects of pre-trial release. Column 1 reports results from a linear probability model with pre-trial release as the dependent variable. The independent variables include defendant and case characteristics as well as court-by-time fixed effects. Column 2 reports results using our preferred judge leniency instrument (CJIVE) as the dependent variable in the linear probability model. Note that the independent variables have been rescaled (divided by 100) for readability of the coefficients and standard errors.

Table 3: First Stage Analysis for Pre-trial Release

	(1) Full sample	(2) Male	(3) Black	(4) Prior offender	(5) Any drug	(6) Any violent	(7) Any property	(8) Felony case
Z	0.314	0.248	0.326	0.286	0.302	0.069	0.269	0.161
	(0.067)	(0.069)	(0.083)	(0.078)	(0.120)	(0.108)	(0.090)	(0.088)
N	91,282	77,087	47,861	51,392	26,189	19,312	33,047	47,927

Notes: This table is an analysis of the first-stage impact of judge leniency on pre-trial release. Each column reports the results of a first stage regression where the instrument is defined as the CJIVE measure of judge leniency. The first column reports the results from regressing the indicator for pretrial release on the CJIVE measure for the full sample and the vector of court-by-time fixed effects. Columns 2 through 8 show results from repeating this regression for subsamples of defendants. Standard errors clustered at the shift level are presented in parentheses.

Table 4: Second Stage Analysis for Pre-trial Release

	OLS	Judge dummies	IJIVE	UJIVE	CJIVE	
	(1)	(2)	(3)	(4)	(5)	(6)
Released	-0.232	-0.271	-0.293	-0.293	-0.431	-0.498
	(0.007)	(0.066)	(0.107)	(0.107)	(0.203)	(0.168)
Jackknife	No	No	Individual	Individual	Cluster	Cluster
Additional covariates	No	No	No	No	No	Yes

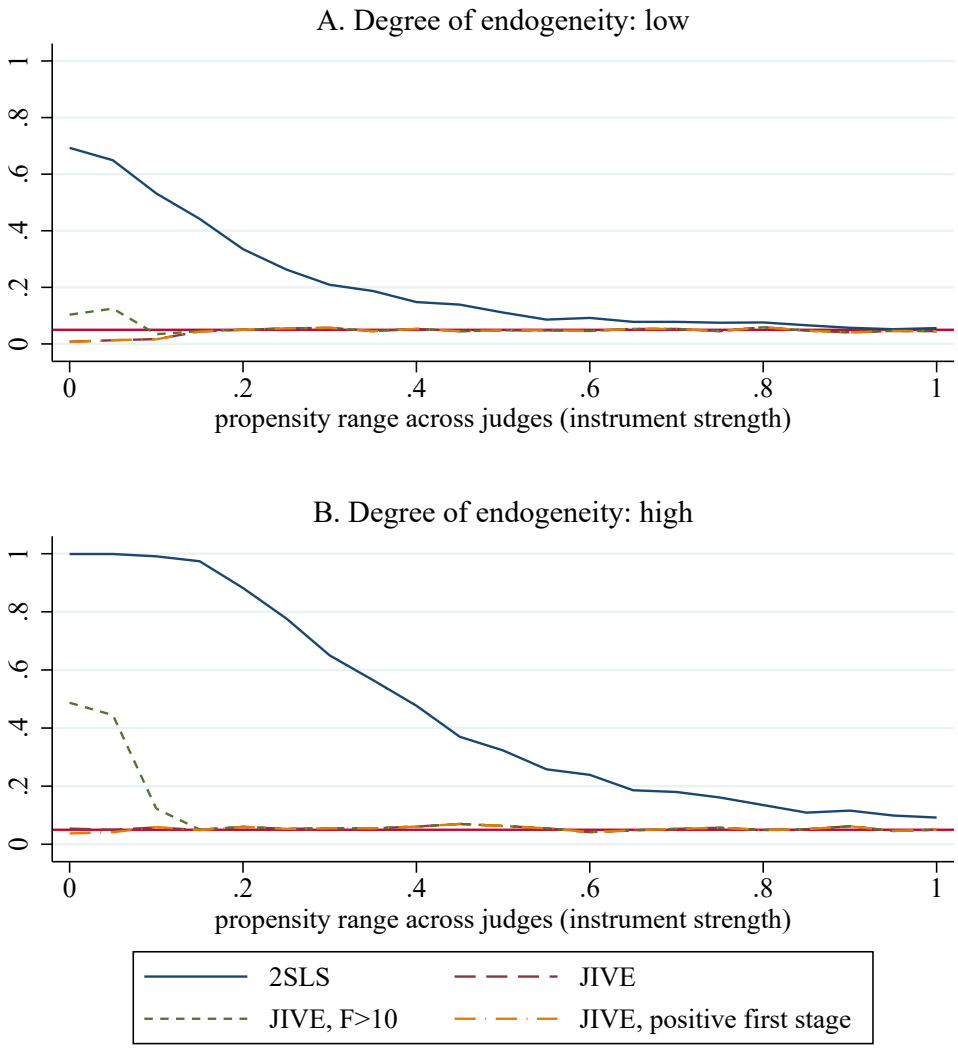
Notes: This table reports estimates of the effects of pre-trial release. The sample size for all specifications is 91,282. The mean of the indicator for being convicted, the dependent variable in all specifications, is 0.585. For comparison, Column 1 shows results from an OLS regression of an indicator for being convicted of any charge on an indicator for being released pretrial. Column 2 shows estimates from a 2SLS regression of the conviction indicator on the pretrial release indicator, where a vector of judge dummies instruments for the pretrial release indicator. The IV estimates in Columns 3-6 use jackknife estimators rather than simply instrumenting using judge dummies. In Columns 3 and 4, the jackknifing is done at the individual level using the IJIVE and UJIVE estimators. In Column 5, the jackknifing is done at the cluster level. In Column 6, the jackknifing is done at the cluster level and an additional vector of demographic and case characteristic controls is included. All specifications include a vector of court-by-time fixed effects.

Table 5: Robustness to Varying Sample Restriction

	(1) 50 cases/judge	(2) 100 cases/judge	(3) 200 cases/judge	(4) 300 cases/judge
Released	-0.427	-0.485	-0.431	-0.424
	(0.198)	(0.207)	(0.203)	(0.203)
N	93,909	93,413	91,282	86,375
Mean of dep. var.	0.583	0.583	0.585	0.584

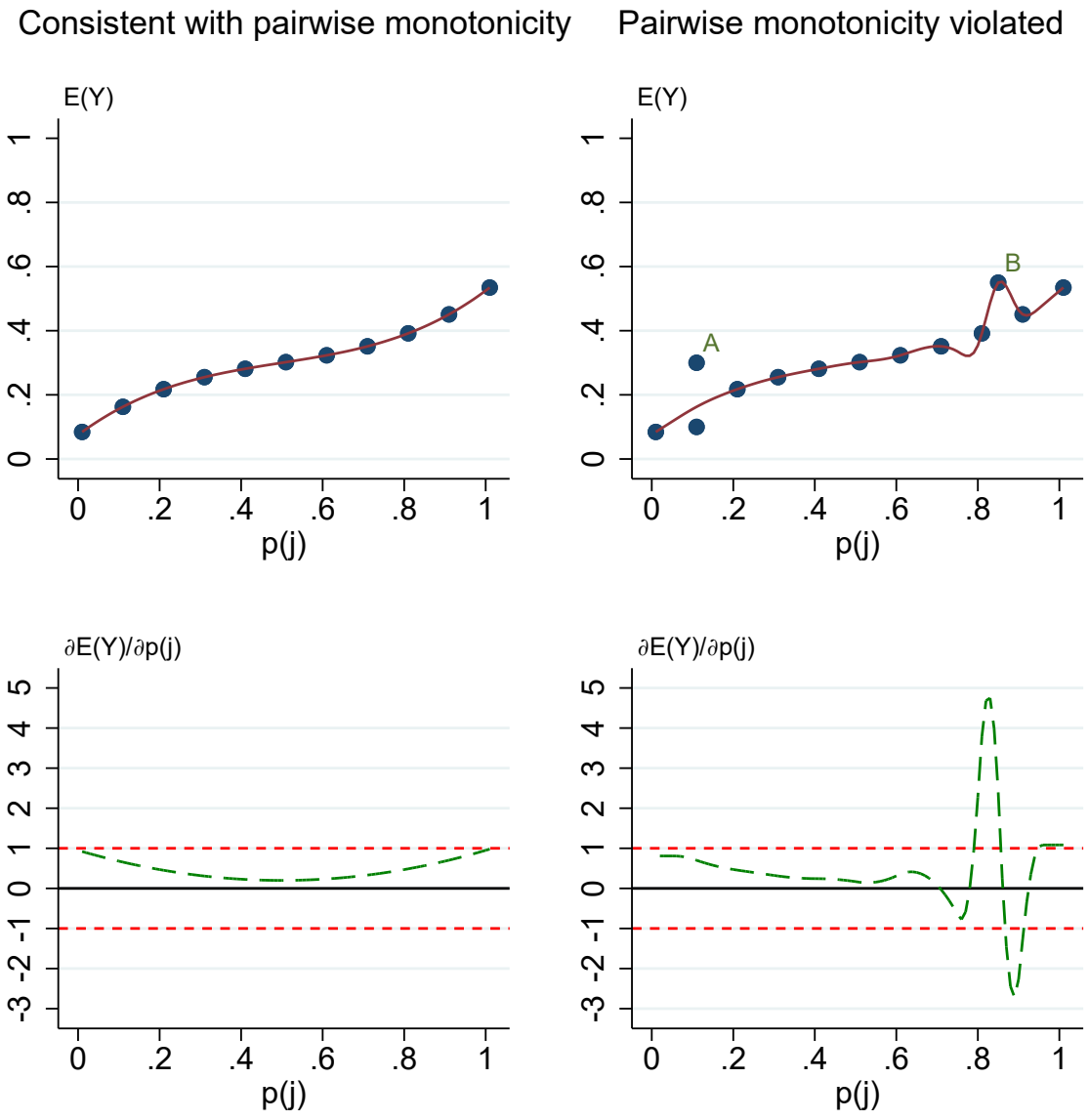
Notes: This table provides a sensitivity analysis based on varying the sample inclusion criteria for number of cases per judge. Each column reports the IV estimated effects of pre-trial release from our preferred specification from samples that use alternative criteria. Column 1 begins with the least restrictive criteria of including cases assigned to judges who see at least 50 cases. Columns 2, 3 and 4 report results by increasing the threshold number of cases to 100, 200 and 300, respectively. The main sample for our analysis is based on the threshold of 200 cases per judge. Standard errors clustered at the shift level are presented in parentheses.

Figure 1: Weak Instrument Simulation Exercise:
IV Rejection Rate, Nominal 5-percent Test



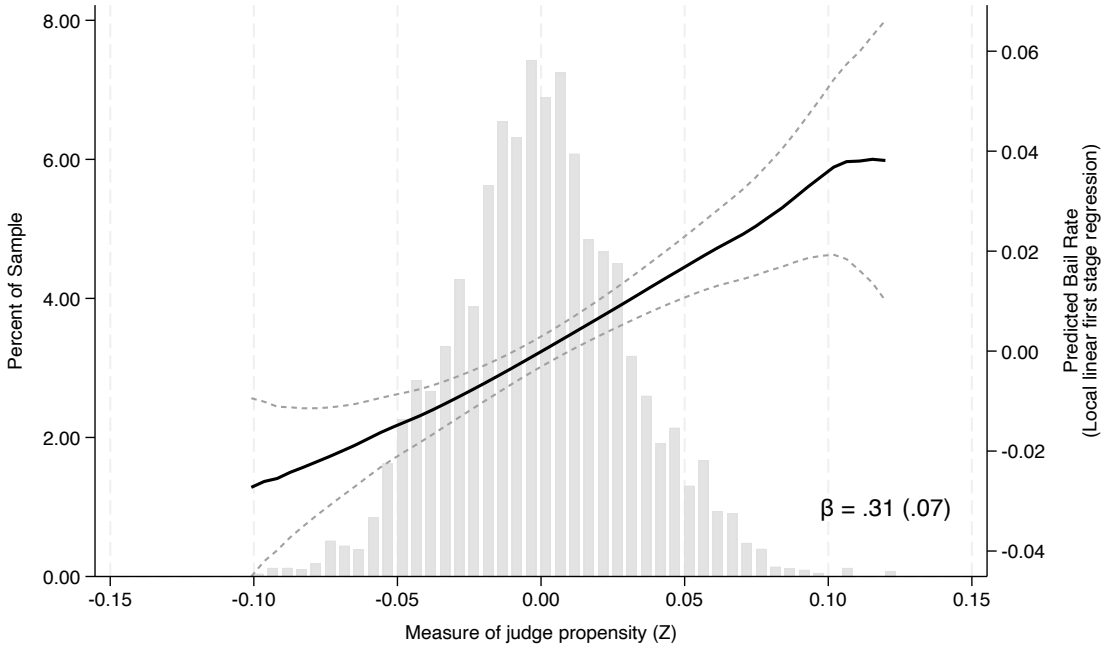
Notes: This figure plots simulated rejection rates as a function of instrument strength based on estimates and robust standard errors from the four estimation procedures indicated in the legend. The solid horizontal line indicates the nominal level of the test (0.05). The data are generated according to the simulation design described in the text. The sample size is 10,000 with 100 examiners and 100 cases per examiner. The simulations with low degree of endogeneity set $\rho = 0.3$ and those with high degree of endogeneity set $\rho = 0.6$.

Figure 2: Marginal Treatment Effects Illustration



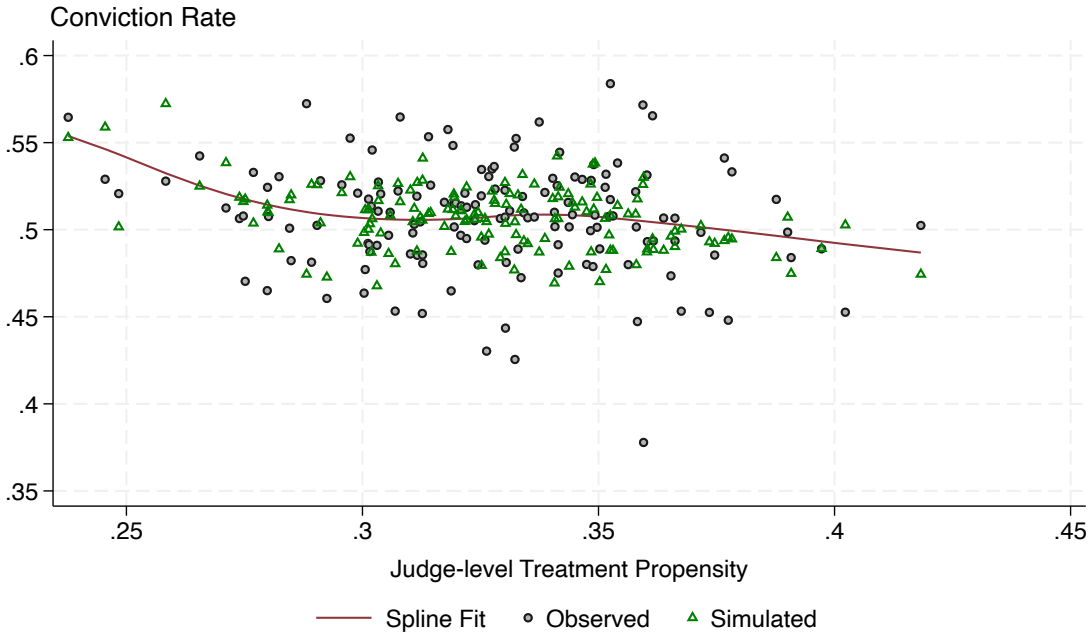
Notes: This figure illustrates hypothetical population-level data from examiner contexts with a binary treatment and binary outcome that would be consistent (left panels) and inconsistent (right panels) with the pairwise monotonicity condition holding. The first row illustrates the relationship between the average outcomes of individuals (e.g., defendants), $E(Y)$, and examiner propensities to administer treatment, p_j. The second row reports the derivative of expected outcomes given examiner propensities, $\partial E(Y)/\partial p_j$. Note that the area of the graph marked with an "A" in the upper right subfigure shows two examiners that have the same propensity p_j but differ in the average outcomes. These population points are inconsistent with pairwise monotonicity. If random assignment and strict exclusion hold, two examiners with the same propensity to treat can only have different average individual outcomes if they differ in the set of individuals whom they assign to treatment (a violation of pairwise monotonicity). The area marked "B" in the upper left graph is inconsistent with pairwise monotonicity because the slope of $E(Y)$ takes on values outside the interval of possible treatment effects for a binary outcome (-1 to 1).

Figure 3: Distribution of Judge Leniency Measure



Notes: This histogram shows the distribution of the CJIVE measure of judge leniency detailed in Section 7. The black line shows a local linear regression of the instrument on the residualized treatment measure. The residuals are based on a model that removes court-by-time fixed effects. For comparison, the figure also reports the estimated coefficient on the CJIVE measure from a linear first-stage regression and the associated standard error clustered at the shift level.

Figure 4: Illustration of Test of Pairwise Monotonicity and Strict Exclusion



Notes: This figure provides an illustration of the test of joint test of strict exclusion and pairwise monotonicity recommended in the Frandsen et al. (2023a). The y-axis reports conviction rates while the x-axis reports judge-level treatment propensities. The dots (in grey) correspond to the observed conviction and treatment propensity in our sample after controlling for court-by-time fixed effects. The test proposed in Frandsen et al. (2023a) is based on fitting a flexible spline function to the observed data on conviction rates and treatment propensities. The solid line (in maroon) shows the predicted values of the spline function fit to the observed data. Intuitively, the test examines two conditions: (1) whether the fitted function meets slope restrictions implied by the range of possible treatment effect sizes; (2) if the judge fixed effects have significant explanatory power after accounting for each judge's predicted point on the fitted function (i.e., whether the distance from the observed points to the fitted function are consistent with sampling variation). In this sample, the test rejects the null hypothesis at the one-percent significance level. Triangles (in green) are simulated mean conviction rates that would be "close enough" to the fitted line to fail to reject the null hypothesis. In a given setting, the definition of close is a function of both distance and the number of cases per judge. For this reason, visual assessment is not a substitute for the formal statistical test proposed in proposed by Frandsen et al. (2023a).

References

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003. ISSN 0304-4076. doi: https://doi.org/10.1016/S0304-4076(02)00201-4. URL https://www.sciencedirect.com/science/article/pii/S0304407602002014.

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When Should You Adjust Standard Errors for Clustering?*. *The Quarterly Journal of Economics*, 138(1):1–35, 10 2022. ISSN 0033-5533. doi: 10.1093/qje/qjac038. URL https://doi.org/10.1093/qje/qjac038.

David S. Abrams, Marianne Bertrand, and Sendhil Mullainathan. Do Judges Vary in Their Treatment of Race? *The Journal of Legal Studies*, 41(2):347–383, June 2012. ISSN 0047-2530, 1537-5366. doi: 10.1086/666006. URL https://www.journals.uchicago.edu/doi/10.1086/666006.

Daniel A Ackerberg and Paul J Devereux. Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity. *The Review of Economics and Statistics*, 91(2):351–362, 05 2009. ISSN 0034-6535. doi: 10.1162/rest.91.2.351. URL https://doi.org/10.1162/rest.91.2.351.

Isaiah Andrews, James H. Stock, and Liyang Sun. Weak instruments in iv regression: Theory and practice. *Annual Review of Economics*, 2019.

J. D. Angrist, G. W. Imbens, and A. B. Krueger. Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67, 1999. ISSN 08837252, 10991255. URL http://www.jstor.org/stable/223249.

Joshua Angrist and Michal Kolesár. One instrument to rule them all: The bias and coverage of just-id iv. *Journal of Econometrics*, 240(2):105398, 2024. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2022.12.012. URL https://www.sciencedirect.com/science/article/pii/S0304407623000295.

Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995. doi: 10.1080/01621459.1995.10476535. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476535.

Anthony Bald, Eric Chyn, Justine Hastings, and Margarita Machelett. The causal impact of removing children from abusive and neglectful homes. *Journal of Political Economy*, 130(7):1919–1962, 2022.

Paul A. Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, 62(3):657–681, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951662.

Manudeep Bhuller and Henrik Sigstad. 2sls with multiple treatments, 2022. URL https://arxiv.org/abs/2205.07836.

Manudeep Bhuller, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad. Incarceration, recidivism, and employment. *Journal of Political Economy*, 128(4):1269–1324, 2020. doi: 10.1086/705330. URL https://doi.org/10.1086/705330.

Ivan A Canay, Magne Mogstad, and Jack Mountjoy. On the Use of Outcome Tests for Detecting Bias in Decision Making. 2023.

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

Gordon B. Dahl, Andreas Ravndal Kostøl, and Magne Mogstad. Family welfare cultures *. *The Quarterly Journal of Economics*, 129(4):1711–1752, 2014. doi: 10.1093/qje/qju019. URL http://dx.doi.org/10.1093/qje/qju019.

Clément de Chaisemartin. Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, 8(2):367–396, 2017. doi: 10.3982/QE601. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/QE601.

Will Dobbie, Jacob Goldin, and Crystal S. Yang. The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–40, February 2018. doi: 10.1257/aer.20161503. URL http://www.aeaweb.org/articles?id=10.1257/aer.20161503.

Kirill S. Evdokimov and Michal Kolesár. Inference in instrumental variables analysis with heterogeneous treatment effects. unpublished working paper, 2019. URL https://www.mit.edu/~kevdokim/het_iv_inf.pdf.

Brigham Frandsen, Lars Lefgren, and Emily Leslie. Judging judge fixed effects. *American Economic Review*, 113(1):253–77, January 2023a. doi: 10.1257/aer.20201860. URL https://www.aeaweb.org/articles?id=10.1257/aer.20201860.

Brigham Frandsen, Emily Leslie, and Samuel McIntyre. Cluster jackknife instrumental variables estimation, 2023b. Working paper, Brigham Young University.

Frederick J Gaudet, George S Harris, John St, and W Charles. Individual differences in the sentencing tendencies of judges. *Am. Inst. Crim. L. & Criminology*, 23:811, 1932.

Christian Hansen, Jerry Hausman, and Whitney Newey. Estimation with many instrumental variables. *Journal of Business and Economic Statistics*, 26(4):398–422, 2008. ISSN 07350015. URL http://www.jstor.org/stable/27639001.

James Heckman, Justin L. Tobias, and Edward Vytlacil. Four Parameters of Interest in the Evaluation of Social Programs. *Southern Economic Journal*, 68(2):211–223, 2001. ISSN 0038-4038.

James J. Heckman and Rodrigo Pinto. Unordered monotonicity. *Econometrica*, 86(1):1–35, 2018. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/44955194.

James J. Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation1. *Econometrica*, 73(3):669–738, 2005. doi: https://doi.org/10.1111/j.1468-0262.2005.00594.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00594.x.

James J. Heckman and Edward J. Vytlacil. Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. In James J. Heckman and Edward E. Leamer, editor, *Handbook of Econometrics*, volume 6, Part B, pages 4779–4874. Elsevier, 2007.

James J. Heckman, Sergio Urzua, and Edward Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432, 2006. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/40043006.

John Eric Humphries, Aurelie Ouss, Kamelia Stavreva, Megan Stevenson, and Winnie van Dijk. Conviction, incarceration, and recidivism: Understanding the revolving door, 2023. URL https://ssrn.com/abstract=4507597.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682. URL http://www.jstor.org/stable/2951620.

Guido W Imbens and Donald B Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997.

Vishal Kamat, Samuel Norris, and Matthew Pecenco. Conviction, Incarceration, and Policy Effects in the Criminal Justice System. 2023.

Toru Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015. ISSN 1468-0262. doi: 10.3982/ECTA11974. URL http://dx.doi.org/10.3982/ECTA11974.

Jeffrey R. Kling. Incarceration length, employment, and earnings. *American Economic Review*, 96(3):863–876, June 2006. doi: 10.1257/aer.96.3.863. URL http://www.aeaweb.org/articles?id=10.1257/aer.96.3.863.

Michal Kolesár. Estimation in an instrumental variables model with treatment effect heterogeneity, 2013. unpublished working paper.

Michal Kolesár, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens. Identification and inference with many invalid instruments. *Journal of Business and Economic Statistics*, 33 (4):474–484, 2015. doi: 10.1080/07350015.2014.978175. URL https://doi.org/10.1080/07350015.2014.978175.

David S Lee, Justin McCrary, Marcelo J Moreira, Jack R Porter, and Luther Yap. What to do when you can't use '1.96' confidence intervals for iv. Working Paper 31893, National Bureau of Economic Research, November 2023. URL http://www.nber.org/papers/w31893.

Sokbae Lee and Bernard Salanié. Identifying effects of multivalued treatments. *Econometrica*, 86(6):1939–1963, 2018. doi: https://doi.org/10.3982/ECTA14269. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14269.

Anna Mikusheva and Liyang Sun. Inference with Many Weak Instruments. *The Review of Economic Studies*, 89(5):2663–2686, 12 2021. ISSN 0034-6527. doi: 10.1093/restud/rdab097. URL https://doi.org/10.1093/restud/rdab097.

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619, 2018. doi: https://doi.org/10.3982/ECTA15463. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15463.

Samuel Norris, Matthew Pecenco, and Jeffrey Weaver. The effects of parental and sibling incarceration: Evidence from Ohio. Technical report, University of Southern California, 2018.

Roman Rivera. Release, Detain, or Surveil? 2023.

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1907619.

Henrik Sigstad. Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs. 2023.

Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2171753.

Joel Waldfogel. The selection hypothesis and the relationship between trial and plaintiff victory. *Journal of Political Economy*, 103(2):229–260, 1995.

J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data, second edition*. Econometric Analysis of Cross Section and Panel Data. MIT Press, 2010. ISBN 9780262232586. URL https://books.google.com/books?id=yov6AQAAQBAJ.

Online Appendix

Examiner and Judge Designs in Economics: A Practitioner's Guide

Eric Chyn, Brigham Frandsen, Emily Leslie

Appendix Table A.1: Applications of Examiner Researcher Designs in Economics

Topic	Study	Year	Treatment	Examiner	Outcome(s)
Crime	Kling	2006	Incarceration length	Judge	Employment, earnings
Crime	Abrams	2010	Incarceration Length	Public Defender	Recidivism
Crime	Chen	2010	Sexual Harassment Case Decision	Judge	Gender Inequality
Crime	Green and Winik	2010	Sentence	Judge	Recidivism
Crime	Di Tella and Schargrodsky	2013	Monitoring method	Judge	Recidivism
Crime	Loeffler	2013	Imprisonment	Judge	Recidivism, unemployment
Crime	Aizer and Doyle	2015	Juvenile incarceration	Judge	Educational attainment, adult recidivism
Crime	Andersen and Wildeman	2015	Probation	Parole Officer	Labor Market Outcomes & Recidivism
Crime	Mueller-Smith	2015	Incarceration	Judge	Recidivism, employment outcomes, public assistance
Crime	Roach and Schanzenbach	2015	Additional Prison Time	Judge	Recidivism
Crime	Gupta et al.	2016	Money bail	Bail judge	Conviction, recidivism
Crime	Leslie and Pope	2017	Pretrial detention	Judge	Conviction, recidivism
Crime	Lum et al.	2017	Setting Bail	Judge	Case Outcome
Crime	Arnold et al.	2018	Racial bias	Bail Judge	Misconduct bail rates
Crime	Arteaga	2018	Parental incarceration	Judge	Children's educational attainment
Crime	Bhuller et al.	2018	Parental incarceration	Judge	Criminal justice outcomes, education, employment
Crime	Dobbie et al.	2018	Pretrial detention	Judge	Conviction, recidivism, employment
Crime	Dobbie et al.	2018	Parental incarceration	Judge	Crime, education, employment
Crime	Harding et al.	2018	Imprisonment	Judge	Labor Market Outcomes
Crime	Stevenson	2018	Pretrial detention	Bail magistrates	Conviction
Crime	Cortés et al.	2019	Juvenile Incarceration	Public Attorney	Recidivism
Crime	Harding et al.	2019	Imprisonment	Judge	Violent Crime
Crime	Hjalmarsson and Lindquist	2019	Military Conscription	Officiator	Crime
Crime	Ribeiro and Ferraz	2019	Pretrial detention	Judge	Recidivism
Crime	White	2019	Incarceration	Judge	Voting
Crime	Aneja and Avenancio-León	2020	Incarceration	Judge	Credit Access
Crime	Bhuller et al.	2020	Incarceration	Judge	Recidivism, employment
Crime	Didwania	2020	Pretrial release	Judge	Case outcomes
Crime	Meier et al.	2020	Early Release	Judge	Reincarceration
Crime	Zapryanova	2020	Prison time, parole time	Judge	Recidivism
Crime	Agan et al.	2021	Misdemeanor prosecution	District Attorney	Subsequent Crime involvement, local crime
Crime	Arbour	2021	Reintegration Program	Probation Officer	Recidivism
Crime	Bhuller and Sigstad	2021	Sentence Reversal	Appeal Panel	Judge Behavior
Crime	Bhuller et al.	2021	Incarceration	Judge	Health of prisoners, prisoners' family members
Crime	Eren and Mocan	2021	Juvenile crime punishment	Judge	Adult recidivism, high school completion
Crime	Grau et al.	2021	Pretrial detention	Judge	Post-verdict labor market outcomes
Crime	Klaassen	2021	Fines & Court Fees	Judge	Reoffending
Crime	Norris et al.	2021	Parental/sibling incarceration	Judge	Incarceration, education, teen parenthood, neighborhood
Crime	Alexeev and Weatherburn	2022	Monetary penalty	Judge	Future crime, drug use
Crime	Augustine et al.	2022	Pretrial diversion programs	Judge	Case outcomes, subsequent Crime contact
Crime	Gillooly	2022	Priority 911 Call Classification	911 Call-Takers	Police Perception
Crime	Jordan and Kim	2022	Police Oversight	Supervising Investigator	Officer Behavior
Crime	LaForest	2022	Early Prison Release	Parole Interviewers	Recidivism
Crime	Rateb	2022	Pretrial Detention	Bail Judge	Court & Crime Outcomes

(Continued on next page.)

Appendix Table A.1: Applications of Examiner Researcher Designs in Economics (continued)

Crime	Williams and Weatherburn	2022	Electronic Monitoring	Judge	Recidivism
Crime	Amaral et al.	2023	Arrest for Domestic Violence	Police Officer	Future Domestic Violence Calls
Crime	Ash and Nix	2023	Asylum	Judge	Failure to appear
Crime	Gonçalves and Mello	2023	Sanctions for speeding	Police officer	Recidivism; speeding offenses
Crime	Helland and Yun	2023	Informal Discovery Conferences	Judge	Number of Discovery Motions
Crime	Humphries et al.	2023	Conviction & Incarceration	Judge	Recidivism
Crime	Jordan et al.	2023	Incarceration, sentencing	Judge	Recidivism
Crime	Lee	2023	Residential Imprisonment	Case Manager	Reincarceration
Crime	Nesbit	2023	Mental Health Treatment	Judge	Recidivism
Crime	Stevenson et al.	2023	Incarceration, convictions	Judge	Recidivism
Crime	Arbour and Marchand	2024	Parole	Parole Board Members	Recidivism
Crime	Axbard	2024	Convicting Corrupt Officials	Judge	Local Government Outcomes
Crime	Wermink et al.	2024	Short-Term Imprisonment	Judge	Recidivism
Finance	Chang and Schoar	2006	Chapter 11 bankruptcy	Judge	Re-filing for bankruptcy
Finance	Munroe and Wilse-Samson	2013	Foreclosure	Chancery-Court Judges	Nearby Foreclosures
Finance	Dobbie and Song	2015	Chapter 13 bankruptcy	Judge	Earnings, mortality, foreclosure rates
Finance	Dobbie et al.	2017	Chapter 13 bankruptcy	Judge	Adverse financial events
Finance	Fraisse	2017	Debt Suspension	Manager	Likelihood to Re-Default
Finance	Bernstein et al.	2019	Firm Liquidation & Reorganization	Judge	Asset Utilization
Finance	Bernstein et al.	2019	Firm Liquidation & Reorganization	Judge	Nearby Employment
Finance	Caskurlu	2019	Strength of Patent Rights	Judges	Acquisition of Substitute Patents & Firm Outcomes
Finance	Rieber and Schechinger	2019	Rating change	Credit rating analyst	Herding behavior
Finance	Dippel and Frye	2020	Allotment of land	US government	Employment, educational attainment, location
Finance	Bamieh et al.	2021	Firing Costs	Judge	Employee Turnover
Finance	Bonfim and Nogueira	2021	Corporate Reorganization and Liquidation	Judge	Reallocation of Labor
Finance	Cespedes et al.	2021	Chapter 13 bankruptcy	Judge	Foreclosure rate, propensity to move, crime rate
Finance	Cheng et al.	2021	Out of court settlement	Judge	financial distress (relative to in-court settlement)
Finance	González-Uribe and Reyes	2021	Alleviating Firm Constraints	Judge	Firm Revenue & Growth
Finance	Grindaker et al.	2021	Bankruptcy	Judge	CEO Career Outcomes
Finance	Honigsberg and Jacob	2021	Expungement	Arbitrators	Future misconduct, long-term career prospects
Finance	Manoel and Augusto Brunassi Silva	2021	Court Delay	Judge	Approval Votes & Plan Approval
Finance	Antill	2022	Firm Liquidation	Bankruptcy Judge	Creditor Recovery
Finance	Beaudry and Willems	2022	Growth Forecast Optimism	IMF Mission Chief	Macroeconomic Outcomes
Finance	Huttunen et al.	2022	Harsher Punishment	Judge	Recidivism for Financial Crimes
Finance	Araujo et al.	2023	Firm Continuation	Judge	Employee Outcomes
Finance	Quignon	2023	Crowd Rating	Evaluator	Funding

(Continued on next page.)

Appendix Table A.1: Applications of Examiner Researcher Designs in Economics (continued)

Finance	Bonfim and Nogueira	2024	Firm Reorganization	Judge	Employee Labor Outcomes
Finance	Ivanov and Wang	2024	Firm Supervision	Bank Examiner	Corporate Lending
Health	Abaluck et al.	2016	Anticoagulation Prescription	Physician	Having a Stoke
Health	Chan et al.	2019	Pneumonia diagnosis	Radiologist	Patient health
Health	Doyle et al.	2019	Hospital quality	Ambulance company	Patient health outcomes
Health	Bakx et al.	2020	Nursing home eligibility	Nursing home assessor	Health, Healthcare spending
Health	Blæhr and Søgaard	2021	Psychotherapy	Hospital departments	Suicide attempts; Health and economic outcomes
Health	Bos et al.	2021	Mental-illness diagnosis	Doctor	Health, labor market, wealth, family outcomes
Health	Berggren	2022	Obtaining an Antibiotic Prescription	General Practitioner	Child Health Outcomes
Health	Cooper et al.	2022	Hospital Prices	Ambulance Company	Mortality
Health	Eichmeyer and Zhang	2022	Medical care	Physician	Opiod use, fentanyl use
Health	Mullainathan and Obermeyer	2022	Mental-illness diagnosis	Doctor	Health, labor market, wealth, family outcomes
Health	Silver and Zhang	2022	Benefit Generosity	Medical Examiner	Health and Well-Being
Health	Bos et al.	2023	Mental Illness Diagnosis	Doctor	Health and Employment Outcomes
Health	Chan et al.	2023	VA Care	Ambulance Company	Mortality, Patient Spending
Health	Eichmeyer and Zhang	2023	Opioid Prescription	Provider	Opioid Dependency & Patient Health Outcomes
Health	Grimon	2023	Child Welfare Intervention	Investigators	Mother's Enrollment in Mental Health and Substance Abuse Treatment Services
Health	Helénsdotter	2023	Removal of Child From Home	Judge	Child's Health
Health	Ofek-Shanny et al.	2023	Long-Term Care Subsidy	Evaluators	Health & Labor Outcomes of Children
Health	Huynh and Chan	2024	Disability Claims	Caseworkers & Physicians	Health Outcomes
Patents	Galasso and Schankerman	2015	Removing patent rights	Judge	Subsequent research related to focal patent
Patents	Sampat and Williams	2015	Patent	Investors	Subsequent research and commercial investment
Patent	Farre-Mensa et al.	2016	Patent	Patent Examiner	Subsequent employment growth, sales, innovation
Patents	Galasso and Schankerman	2018	Patent invalidation	Judge	Subsequent innovation and exit by patent holders
Patents	Gaule	2018	Patent Acquisition	Patent Examiner	Firm Outcomes
Patent	Dyevre	2023	R&D Funding Source	Patent Examiner	Aggregate Productivity Growth
Patent	Gupta	2023	Patent Acquisition	Patent Examiner	Monopoly Life
Public	Doyle	2007	Foster care	Child protection worker	Juvenile delinquency, teen motherhood, employment
Public	Doyle	2008	Foster care	Child protection worker	Adult crime
Public	Famulari	2012	Remedial Writing	Grader	Academic Performance
Public	Doyle Jr	2013	Foster care	Child protection worker	Delinquency, health
Public	Maestas et al.	2013	Disability benefits	Disability examiner	Labor supply
Public	Dahl et al.	2014	Receipt of a welfare	Judge	Participation in the next generation
Public	French and Song	2014	Disability benefits	Judge	Labor supply
Public	Autor et al.	2017	Disability benefits	Judge	Household income, consumption, labor supply
Public	Black et al.	2018	Disability benefits	Judge	Mortality
Public	Hyman	2018	Trade Adjustment Assistance	Case investigator	Labor market outcomes
Public	Okudaira	2018	Worker Court Victory	Judge	Job Creation & Destruction

(Continued on next page.)

Appendix Table A.1: Applications of Examiner Researcher Designs in Economics (continued)

Public	Arni and Schiprowski	2019	Job Search Requirements	Case Worker	Effort and Labor Market Outcomes
Public	Galletta et al.	2019	Judge Sentiments in Court Decisions	Judge	Social Attitudes
Public	Hjalmarsson and Lindquist	2019	Mandatory military service	Draft officiator	Criminal behavior, post-service labor market outcomes
Public	Humphries et al.	2019	Eviction	Judge	Financial distress, residential mobility, neighborhood
Public	Baron and Gross	2020	Foster care	Child protection worker	Children's safety and educational outcomes
Public	Diamond et al.	2020	Foreclosure	Judge	Housing stability, homeownership, financial stress
Public	Bald et al.	2022	Foster care	Child protection worker	Child test scores, grade repetition
Public	Baron and Gross	2022	Foster care	Child protection worker	Adult arrests, convictions, incarceration; child safety
Public	Cohen	2022	Housing assistance	Case manager	Homelessness, crime, income, employment
Public	Collinson et al.	2022	Evictions	Judge	Homelessness, earnings, access to credit
Public	van der Klaauw and Vethaak	2022	Broader Job Search Requirements	Case Worker	Job Finding
Public	De Souza	2023	Disabled Worker Quota	Inspector	Labor Market & Welfare
Public	Herbst	2023	IDR Enrollment	Servicing Agent	Student Borrower Outcomes
Public	Jeremy	2023	Vacancy Referrals	Caseworker	Job Interview & Offers
Public	Lee	2023	Residential housing	Case manager	Reincarceration
Public	Moscona	2023	Management of Development Aid	Project Leader	Violent Conflict in Africa
Public	Silver and Zhang	2023	Monthly cash benefits	Mental disorder examiner	Food security, homelessness, debt, health, mortality
Public	Yin et al.	2023	Vocational Rehabilitation Services	Vocational Rehabilitation Counselors	Employment Outcomes

Notes: This table provides a survey of 136 studies in economics that have used examiner tendency research designs.

A Examples with monotonicity violations

When treatment effects are heterogeneous and examiners differ in how they rank subjects for treatment, Section 5.2 highlights that 2SLS identifies a proper weighted average of treatment effects if an average monotonicity condition holds. This section provides stylized examples of cases in which average monotonicity may *or* may not hold in the presence of violations of pairwise monotonicity.

To begin, consider a setting in which bail judges assign defendants to pretrial detention or pretrial release. Each judge observes two characteristics for the defendant: (i) whether the defendant has a criminal history and (ii) whether they belong to a majority racial group. Let c_i and r_i be indicators for having a criminal history or being a majority racial group member, respectively. Each judge j decides whether to detain defendant i by evaluating whether the defendant's probability of misconduct exceeds a judge-specific threshold $\tau_j(r_i)$. Judges may set different thresholds based on racial group status due to taste-based discrimination. Let $D_i(j)$ be a dummy variable indicating whether person i would be assigned to pretrial detention by judge j. The fraction of the population assigned to pretrial detention by judge j is measured by p_j.

In this setting, we specify that the probability of misconduct depends only on criminal history. Specifically, we assume that 50 percent of defendants with a criminal history and 30 percent of those without a criminal history will engage in pretrial misconduct if released. Race could be indirectly informative about misconduct if rates of criminal histories vary across demographic groups (although we do not impose that condition in our examples below). We formalize taste-based discrimination as instances in which judges apply a lower threshold for pretrial detention to defendants in the minority group given a probability of pretrial misconduct. Specifically, their threshold for individuals in the minority group is 0.4 lower than their threshold for people in the majority group. For some possible judges, this setup can lead to patterns that are consistent with average monotonicity but not pairwise monotonicity. It can also lead to violations of average monotonicity. To illustrate the possible monotonicity violations, consider the following scenarios:

Case 1: Average monotonicity holds; pairwise monotonicity is violated. In the population, suppose the following: 45% do not have a criminal record and belong to the majority group; 5% do not have a criminal record and belong to minority group; 45% have a criminal record and belong to the majority group, and 5% have a criminal record and belong to the minority group. In other words, 50% of those in the majority and 50% of those in the minority groups have criminal records. There are four judges, and we assume that judges 1 and 3 discriminate against members of the minority group by having 0.4 lower thresholds for detention.

Appendix Table A.2 (below) summarizes treatment outcomes for defendants assuming that each judge has an equal caseload. Each labelled column reports the potential treatment outcomes for each of the four types of defendants defined by the two observed characteristics. For example, column 1 shows that no judge assigns defendants who are majority group members without a criminal record to treatment. This is because the probability of misconduct is 30% for individuals without a criminal record and this falls below all judges' thresholds for the majority group.

In this example, pairwise monotonicity does not hold because judge 2, whose propensity to treat is 0.5, assigns people in the minority group without a criminal record to pretrial release, while judge 3, whose propensity to treat is 0.1, assigns them to pretrial detention. However, average monotonicity holds because the covariance between potential treatment status and judge propensity

to treat is nonnegative for all types of defendants.

Appendix Table A.2: Pairwise Monotonicity is Violated and Average Monotonicity Holds

				Potential Treatment Status, $D_i(j)$ by Defendant Type				
				(1)	(2)	(3)	(4)	
		τ_j		No criminal history		Criminal history		
Judge	Discriminator?	Majority	Minority	Majority	Minority	Majority	Minority	p_j
1	Yes	0.4	0.0	0	1	1	1	0.55
2	No	0.5	0.5	0	0	1	1	0.50
3	Yes	0.6	0.2	0	1	0	1	0.10
4	No	0.7	0.7	0	0	0	0	0.00
Covariance of $D_i(j)$ and p_j for defendant type				0	0.019	0.119	0.072	

Notes: This table is an example in which judge behavior violates pairwise monotonicity while average monotonicity holds. The four columns labelled to the right indicate the potential treatment status $D_i(j)$ for defendants defined by whether (i) they have observabled criminal backgrounds and (ii) whether they are members of a minority or majority group. The four rows of the table list each judge j, where two of the judges discriminate against members of a minority group by imposing a lower threshold (τ_j). Each row (final column at right) reports the population weighted likelihood of treatment for each judge (p_j). The bottom row of the table reports the covariance of potential treatment status and judge-specific treatment probability across the four judges, conditional on the type of defendant.

Case 2: Average monotonicity is violated. Consider Case 1 while removing judge 1 from the example. Appendix Table A.3 shows that both average monotonicity and pairwise monotonicity are violated in this scenario. This case with the three remaining judges illustrates that the satisfaction of monotonicity conditions can be sensitive to which judges are included in the sample. This is not specific to average monotonicity; if we start with Case 1 and remove judge 3, then pairwise monotonicity (and, therefore, average monotonicity) would hold.

Appendix Table A.3: Both Pairwise Monotonicity and Average Monotonicity Violated

				Potential Treatment Status, $D_i(j)$ by Defendant Type				
				(1)	(2)	(3)	(4)	
		τ_j		No criminal history		Criminal history		
Judge	Discriminator?	Majority	Minority	Majority	Minority	Majority	Minority	p_j
2	No	0.5	0.5	0	0	1	1	0.50
3	Yes	0.6	0.2	0	1	0	1	0.10
4	No	0.7	0.7	0	0	0	0	0.00
Covariance of $D_i(j)$ and p_j for defendant type				0	-0.033	0.100	0.067	

Notes: This table is a stylized example in which judge behavior violates both pairwise and average monotonicity. See Appendix Table A.2 for detailed notes.

B Detailed discussion of multiple treatments frameworks

Bhuller and Sigstad (2023) and Humphries et al. (2023) provide frameworks in which 2SLS can identify positively weighted averages of the effects of multiple treatments in examiner tendency designs. Specifically, Bhuller and Sigstad (2022) show when linear 2SLS with multiple endogenous treatments using examiner propensities for each treatment as excluded instruments identifies proper weighted causal effects. By contrast, Humphries et al. (2023a) shows when 2SLS controlling for non-focal propensities identifies proper weighted average effects of a focal treatment. This appendix describes the conditions in each framework in turn, and then develops an important special case in which they are equivalent: that of three mutually exclusive treatments and three examiners.

First, we establish notation in the case of three mutually exclusive treatment categories and three examiners that will be useful for both frameworks. For individual i, we index the treatment categories by $D_i \in \{0,1,2\}$ and the three possible examiners by $J_i \in \{0,1,2\}$.[41] Let $D_{si} := 1(D_i = s)$ be an indicator for actually receiving treatment s, and $p_s(J_i) = E[D_{si}|J_i]$ be the propensity of the examiner to assign individuals to treatment s. Denote potential treatment status as $D_{si}(j)$ which is an indicator for receipt of treatment s if the individual is assigned to examiner j. There are several "margins" of treatment effects given the multiple treatments in this context. The natural treatment effects of interest compare potential outcomes under treatment s to a reference treatment which is designated by zero: $\delta_i^{0\to s} := Y_i(s) - Y_i(0)$, where $Y_i(s)$ is individual i's potential outcome under treatment s. The goal in the Bhuller and Sigstad (2022) framework is to identify proper weighted averages of $\delta_{1i}^{0\to 1}$ and $\delta_i^{0\to 2}$ as coefficients on the indicators D_{1i} and D_{2i} from 2SLS estimation of the equation:

$$Y_i = \alpha + \delta_1 D_{1i} + \delta_2 D_{2i} + \varepsilon_i,$$

where $p_1(J_i)$ and $p_2(J_i)$ are the excluded instruments. In the Humphries et al. (2023a) framework the goal similar: identify proper weighted averages of a focal treatment, controlling (perhaps linearly) for non-focal propensities:

$$Y_i = \alpha + \delta_1 D_{1i} + \pi p_2(J_i) + \varepsilon_i.$$

In what follows below, we adapt the identifying assumptions from Bhuller and Sigstad (2022) and Humphries et al. (2023a) to this setting. Throughout, we assume that examiners are assigned randomly, vary sufficiently in their propensities (i.e., they satisfy the instrument rank condition), and only influence outcomes through D_i. Note that some of the conditions below invoke the concept of *partial correlation*. The partial correlation between random variables A and B given C is equal to the usual (Pearson) correlation between the residuals from a linear regression of A on C and the residuals from a linear regression of B on C.

[41] The index notation that we chose is intentional. As discussed below, the Bhuller and Sigstad conditions imply a mapping between examiners and treatments.

B.1 Bhuller and Sigstad (2023) Assumptions

In the following assumptions, consider J_i (individual i's examiner assignment) to be a random variable for each individual i, whose distribution is determined by the mechanism assigning examiners to individuals. Therefore, $D_{1i}(J_i)$, $D_{2i}(J_i)$, $p_1(J_i)$, and $p_2(J_i)$ are also random variables for each individual i. The following conditions govern the relationships among these random variables for each individual i.

- *Average conditional monotonicity (ACM).* ACM is defined for each specific treatment given another. ACM of treatment 1 given $p_2(J_i)$, denoted ACM(1|2), requires that, for every individual i, the partial correlation between $p_1(J_i)$ and $D_{1i}(J_i)$ given $p_2(J_i)$ be nonnegative. In other words, a hypothetical examiner-level linear regression of $D_{1i}(J_i)$ on $p_1(J_i)$ and $p_2(J_i)$ yields a positive coefficient on $p_1(J_i)$ for each individual i. ACM of treatment 2 given $p_1(J_i)$ is defined similarly.

- *No cross effects (NC).* NC is also defined specifically for each treatment. The NC condition for treatment 1 given $p_2(J_i)$, denoted NC(1|2), says that, for every individual i, the partial correlation between $p_1(J_i)$ and $D_{2i}(J_i)$ is zero. The NC condition for treatment 2 is defined similarly.

Intuitively, assumptions ACM(1|2) and NC(1|2) together ensure that increasing $p_1(J_i)$, controlling linearly for $p_2(J_i)$, on average increases $D_{1i}(J_i)$ and on average has zero effect on $D_{2i}(J_i)$. The key consequence is that the 2SLS coefficient on D_{1i} in a model with D_{1i} and D_{2i} as endogenous regressors and $p_1(J_i)$ and $p_2(J_i)$ as excluded instruments identifies a proper weighted average of $\delta_i^{0\to1}$. Similarly, ACM(2|1) and NC(2|1) imply that the coefficient on D_{2i} identifies a proper weighed average of $\delta_i^{0\to2}$.

B.2 Humphries et al. (2023) Assumptions

The conditions in Humphries et al. (2023a) consider variation in treatment assignment holding the examiner propensity for one of the treatments fixed. In the case of three examiners, this means considering how treatment status would change if an individual were switched between two examiners who have the same propensity for one of the treatments. To make the condition below concrete, suppose examiner $J_i = 1$ has higher propensity for treatment 1 than examiner $J_i = 0$, but they have equal propensities for treatment 2 (i.e., $p_1(1) > p_1(0)$ and $p_2(0) = p_2(1)$). Similarly, suppose that examiner $J_i = 2$ has higher propensity for treatment 2 than examiner 0, but they have equal propensities for treatment 1 (i.e., $p_2(2) > p_2(0)$ and $p_1(0) = p_1(2)$). Humphries et al. (2023a) provide the following condition under which the coefficient on D_{1i} in an IV procedure that employs $p_1(J_i)$ as the excluded instrument and conditions (perhaps nonparametrically) on $p_2(J_i)$ will recover a proper weighted average of $\delta_i^{0\to1}$.

- *Unordered partial monotonicity (UPM).* UPM of treatment 1 given treatment 2, denoted UPM$(1|2)$ means the following hold for all i:

  1. $D_{1i}(1) \geq D_{1i}(0)$
  2. $D_{0i}(1) \leq D_{0i}(0)$

3. $D_{2i}(1) = D_{2i}(0)$

UPM$(1|2)$ implies that, if an individual were to switch from examiner 0 to examiner 1 (which increases $p_1(J_i)$ holding $p_2(J_i)$ fixed), that individual might switch into treatment 1, but would never switch out. The second inequality means the individual might switch out of treatment 0, but would never switch in. The equality means no individual's treatment 2 status would change when switching from examiner 0 to examiner 1.[42] In other words, the only change that could happen if an individual were to switch from examiner 0 to examiner 1 is a switch from treatment 0 to treatment 1. Similarly, UPM$(2|1)$ means that the only change that could happen if an individual were to switch from examiner 0 to examiner 2 (which increases $p_2(J_i)$ holding $p_1(J_i)$ constant) is a switch from treatment 0 to treatment 2.

In the current special case where examiners 0 and 1 have identical propensities for treatment 2 and examiners 0 and 2 have identical propensities for treatment 1, the assumption UPM(1|2) implies that the 2SLS coefficient on D_{1i}, with $p_1(J_i)$ as the excluded instrument and conditioning on $p_2(J_i)$, identifies a proper weighted average of $\delta_i^{0\to1}$.

Here "conditioning on $p_2(J_i)$" is equivalent to including it as a linear control because $p_2(J_i)$ takes on only two values. Beyond this special case, however, conditioning on $p_2(J_i)$ would either require nonparametrically controlling for $p_2(J_i)$, or assuming additionally that $E[p_{1i}(J_i)|p_2(J_i)]$ is linear in $p_2(J_i)$.

Identifying proper weighted averages of $\delta_i^{0\to2}$ requires the analogous assumption UPM$(2|1)$. If both UPM$(1|2)$ and UPM$(2|1)$ hold, then 2SLS estimation with both D_{1i} and D_{2i} as endogenous regressors and $p_1(J_i)$ and $p_2(J_i)$ as excluded instruments identifies effects of both treatments.

B.3 Equivalence of Results

In this just identified example, the Bhuller and Sigstad (2022) conditions are equivalent to the Humphries et al. (2023a) conditions. That is, ACM$(1|2)$, ACM$(2|1)$, NC$(1|2)$, and NC$(2|1)$ imply UPM$(1|2)$ and UPM$(2|1)$ and vice versa. To see this, note that these conditions restrict only how individuals' treatment status responds to examiner assignment. The two sets of conditions are equivalent if they allow the same responses of individual treatment status to examiner assignment.

There are 27 possible ways that the three examiners can allocate a defendant to one of three treatments. Appendix Table A.4 below lists all the possible treatment permutations. We'll refer to each treatment permutation as a "response type." Each response type is defined by its potential treatment states as a function of examiner assignment: $(D_i(0), D_i(1), D_i(2)) \in \{0,1,2\}^3$. For example, the response type "always 0" is allocated to treatment 0 by all three examiners, and so has potential treatment states $(0,0,0)$.

The second column of the table shows that the Bhuller and Sigstad (2022) assumptions, ACM$(1|2)$, ACM$(2|1)$, NC$(1|2)$, and NC$(2|1)$, allow only six response types:

- $(0,0,0)$
- $(1,1,1)$

[42] In Humphries et al. (2023a), the equality is expressed as a weak inequality, but in our three-examiner, three-treatment scenario here where examiners 0 and 1 share the same propensity for treatment 2, the weak inequality must be satisfied with equality.

- $(2, 2, 2)$
- $(0, 1, 0)$
- $(0, 0, 2)$
- $(0, 1, 2)$.

The third column shows that the Humphries et al. (2023a) assumptions, UPM$(1|2)$ and UPM$(2|1)$, allow the same six response types. It also shows how each of the prohibited response types violates those conditions.

The two sets of assumptions make identical restrictions on how individual treatment status responds to examiner assignment, and therefore are equivalent in this special case.

The argument above establishes via brute force that the two assumptions are equivalent. Further intuition is provided in Figure A.1 which illustrates the pattern of treatment assignment that must occur in this setting. The rows represent each of the three judges while the columns represent the six allowed response types. The pattern in each cell of the figure indicates whether defendants of a given response type would be assigned to treatment 0 (dots), 1 (crosshatch dots) or 2 (crosshatch) when they are assigned to a specific examiner.

Figure A.1 below illustrates three response types that we might describe as "always-takers" of one of the three treatments. Defendants whose potential treatment status is defined by the vector (0,0,0) always receive treatment 0. Similarly, response types (1,1,1), and (2,2,2) receive the same treatment regardless of examiner assignment. For the remaining three response types, treatment status varies with examiner assignment. All of them will be assigned to treatment 0 if assigned to examiner 0. Examiner 1 moves some of them into treatment 1, and examiner 2 moves some of them into treatment 2.

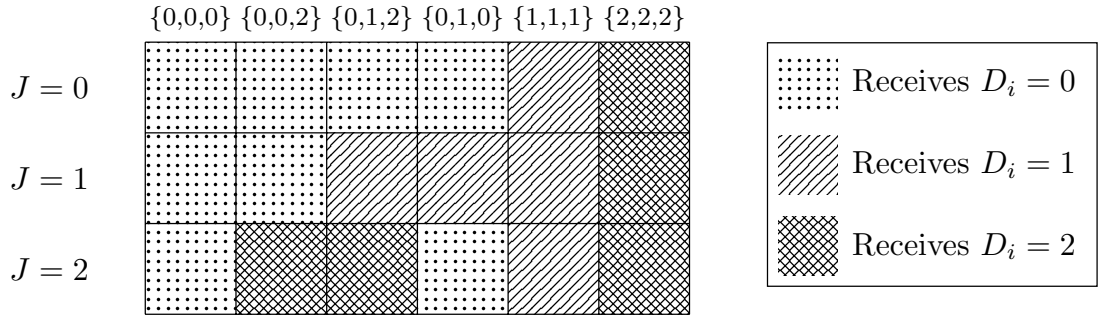
The figure demonstrates that not only do the allowable response types imply the existence of a reference treatment, they also imply the existence of a reference examiner. Identification of the average effects $\bar{\delta}_1$ and $\bar{\delta}_2$ effects is possible because a comparison between those assigned to examiners 1 and 0 isolates the impact of receiving treatment 1 relative to treatment 0. Similarly, the comparison between those assigned to examiners 2 and 0 isolates the impact of receiving treatment 2 relative to treatment 0. For a researcher to argue that only the six allowed response types will exist in their setting, they must be willing to argue that there is a labeling of examiners such that one is the reference examiner, while the other examiners only move people across exactly one treatment margin relative to the reference examiner.

Appendix Table A.4: Response types allowed under ACM, NC, and UPM Assumptions

(1) Response type: $(\mathbf{D}(\mathbf{0}), \mathbf{D}(\mathbf{1}), \mathbf{D}(\mathbf{2}))$	(2) **Satisfies ACM and NC conditions?**	(3) **Satisfies UPM inequalities?**
$(\mathbf{0}, \mathbf{0}, \mathbf{0})$	**Yes**	**Yes**
$(0, 0, 1)$	No	Violates UPM$(2\|1)$ #3
$(\mathbf{0}, \mathbf{0}, \mathbf{2})$	**Yes**	**Yes**
$(\mathbf{0}, \mathbf{1}, \mathbf{0})$	**Yes**	**Yes**
$(0, 1, 1)$	No	Violates UPM$(2\|1)$ #3
$(\mathbf{0}, \mathbf{1}, \mathbf{2})$	**Yes**	**Yes**
$(0, 2, 0)$	No	Violates UPM$(1\|2)$ #3
$(0, 2, 1)$	No	Violates UPM$(1\|2)$ #3 and UPM$(2\|1)$ #3
$(0, 2, 2)$	No	Violates UPM$(1\|2)$ #3
$(1, 0, 0)$	No	Violates UPM$(1\|2)$ #1 & #2 and UPM$(2\|1)$ #2 & #3
$(1, 0, 1)$	No	Violates UPM$(1\|2)$ #1 & #2
$(1, 0, 2)$	No	Violates UPM$(1\|2)$ #1 & #2 and UPM$(2\|1)$ #3
$(1, 1, 0)$	No	Violates UPM$(2\|1)$ #2 & #3
$(\mathbf{1}, \mathbf{1}, \mathbf{1})$	**Yes**	**Yes**
$(1, 1, 2)$	No	Violates UPM$(2\|1)$ #3
$(1, 2, 0)$	No	Violates UPM$(1\|2)$ #1 & #3 and UPM$(2\|1)$ #2 & #3
$(1, 2, 1)$	No	Violates UPM$(1\|2)$ #1 & #3
$(1, 2, 2)$	No	Violates UPM$(1\|2)$ #1 & #3 and UPM$(2\|1)$ #3
$(2, 0, 0)$	No	Violates UPM$(1\|2)$ #2 & #3 and UPM$(2\|1)$ #1 & #2
$(2, 0, 1)$	No	Violates UPM$(1\|2)$ #2 & #3 and UPM$(2\|1)$ #1 & #3
$(2, 0, 2)$	No	Violates UPM$(1\|2)$ #2 & #3
$(2, 1, 0)$	No	Violates UPM$(1\|2)$ #3 and UPM$(2\|1)$ #1 & #2
$(2, 1, 1)$	No	Violates UPM$(1\|2)$ #3 and UPM$(2\|1)$ #1 & #3
$(2, 1, 2)$	No	Violates UPM$(1\|2)$ #3
$(2, 2, 0)$	No	Violates UPM$(2\|1)$ #1 & #2
$(2, 2, 1)$	No	Violates UPM$(2\|1)$ #1 & #3
$(\mathbf{2}, \mathbf{2}, \mathbf{2})$	**Yes**	**Yes**

Notes: This table demonstrates the equivalence of the assumptions proposed in Bhuller and Sigstad (2022) and Humphries et al. (2023a) in a setting with three distinct treatments and three judges. Each row is one of the 27 possible treatment permutations for the three judges. We refer to each row as a "response type" which is defined by potential treatment states as a function of examiner assignment. For example, the first row is the response type for "always 0" which is the type of defendant who is allocated to treatment 0 by all examiners. In this setting, there are a total of four ACM and NC conditions from Bhuller and Sigstad (2022). Column 2 shows that only six response types are possible when these four conditions hold. In the framework from Humphries et al. (2023a), there are two UPM conditions that have associated inequality conditions. Column 3 shows that the associated UPM inequalities hold for the six response types that are possible when the ACM and NC conditions hold.

Appendix Figure A.1: Six Potential Treatment Response Types and Treatment Assignment



Notes: This figure illustrates the pattern of treatment assignment that must hold to satisfy the conditions in Bhuller and Sigstad (2022) and Humphries et al. (2023a) in a three examiner and three treatment setting. The rows represent examiners while the columns represent the six response types (i.e., the potential treatment status for a group of defendants) that are permitted. The pattern in each cell indicates whether a defendant of a given response type would be assigned to treatment 0 (dots), 1 (crosshatch dots) or 2 (crosshatch) when they are assigned to a specific examiner.

References

Jason Abaluck, Leila Agha, and David Chan. Personalized medicine and patient selection: Discretion vs. guidelines. 2016.

David Abrams. Building criminal capital vs. specific deterrence: The effect of incarceration length on recidivism. In *5th annual conference on empirical legal studies paper*, 2010.

Amanda Y. Agan, Jennifer L. Doleac, and Anna Harvey. Misdemeanor prosecution. Working Paper 28600, National Bureau of Economic Research, March 2021. URL `http://www.nber.org/papers/w28600`.

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

Sergey Alexeev and Don Weatherburn. Fines for illicit drug use do not prevent future crime: evidence from randomly assigned judges. *Journal of Economic Behavior & Organization*, 200: 555–575, 2022.

Sofia Amaral, Gordon B Dahl, Victoria Endl-Geyer, Timo Hener, and Helmut Rainer. Deterrence or backlash? arrests and the dynamics of domestic violence. Technical report, National Bureau of Economic Research, 2023.

Lars H Andersen and Christopher Wildeman. Measuring the effect of probation and parole officers on labor market outcomes and recidivism. *Journal of Quantitative Criminology*, 31:629–652, 2015.

Abhay P Aneja and Carlos F Avenancio-León. No credit for time served? incarceration and credit-driven crime cycles. *Unpublished manuscript*, 2020.

Samuel Antill. Do the right firms survive bankruptcy? *Journal of Financial Economics*, 144(2): 523–546, 2022.

Aloisio Araujo, Rafael Ferreira, Spyridon Lagaras, Flavio Moraes, Jacopo Ponticelli, and Margarita Tsoutsoura. The labor effects of judicial bias in bankruptcy. *Journal of Financial Economics*, 150(2):103720, 2023.

William Arbour. Can Recidivism Be Prevented From Behind Bars? Evidence From a Behavioral Program. Technical report, 2021. URL `https://www.economics.utoronto.ca/public/workingPapers/tecipa-683.pdf`.

William Arbour and Steeve Marchand. Can parole reduce both time served and crime. *Available at SSRN 4114251*, 2024.

Patrick Arni and Amelie Schiprowski. Job search requirements, effort provision and labor market outcomes. *Journal of Public Economics*, 169:65–88, 2019.

David Arnold, Will Dobbie, and Crystal S Yang. Racial bias in bail decisions*. *The Quarterly Journal of Economics*, page qjy012, 2018. doi: 10.1093/qje/qjy012. URL http://dx.doi.org/10.1093/qje/qjy012.

Carolina Arteaga. The cost of bad parents: Evidence from incarceration on children's education. Technical report, UCLA, 2018.

T Ash and E Nix. How asylum seekers in the united states respond to their judges: Evidence and implications. Technical report, Working Paper, 2023.

Elsa Augustine, Johanna Lacoe, Steven Raphael, and Alissa Skog. The impact of felony diversion in san francisco. *Journal of Policy Analysis and Management*, 41(3):683–709, 2022.

David Autor, Andreas Ravndal Kostol, Magne Mogstad, and Bradley Setzler. Disability benefits, consumption insurance, and household labor supply. Working Paper 23466, National Bureau of Economic Research, June 2017. URL http://www.nber.org/papers/w23466.

Sebastian Axbard. Convicting corrupt officials: Evidence from randomly assigned cases. *Review of Economic Studies*, page rdae069, 2024.

Pieter Bakx, Bram Wouterse, Eddy van Doorslaer, and Albert Wong. Better off at home? effects of nursing home eligibility on costs, hospitalizations and survival. *Journal of Health Economics*, 73:102354, 2020. ISSN 0167-6296. doi: https://doi.org/10.1016/j.jhealeco.2020.102354. URL https://www.sciencedirect.com/science/article/pii/S0167629619307635.

Anthony Bald, Eric Chyn, Justine Hastings, and Margarita Machelett. The causal impact of removing children from abusive and neglectful homes. *Journal of Political Economy*, 130(7):1919–1962, 2022.

Omar Bamieh, Decio Coviello, Andrea Ichino, and Nicola Persico. Effect of firing frictions on turnover. *WorkINPS Papers*, 45, 2021.

E. Jason Baron and Max Gross. Temporary Stays and Persistent Gains: The Causal Effects of Foster Care. Technical report, 2020.

E Jason Baron and Max Gross. Is there a foster care-to-prison pipeline? evidence from quasi-randomly assigned investigators. Technical report, National Bureau of Economic Research, 2022.

Paul Beaudry and Tim Willems. On the macroeconomic consequences of over-optimism. *American Economic Journal: Macroeconomics*, 14(1):38–59, 2022.

Andrea Berggren. Antibiotic consumption and health care utilization in children. *Available at SSRN 4241676*, 2022.

Shai Bernstein, Emanuele Colonnelli, Xavier Giroud, and Benjamin Iverson. Bankruptcy spillovers. *Journal of Financial Economics*, 133(3):608–633, 2019a.

Shai Bernstein, Emanuele Colonnelli, and Benjamin Iverson. Asset allocation in bankruptcy. *The Journal of Finance*, 74(1):5–53, 2019b.

Manudeep Bhuller and Henrik Sigstad. Feedback and learning: The causal effects of reversals on judicial decision-making. *Available at SSRN 4000424*, 2021.

Manudeep Bhuller and Henrik Sigstad. 2sls with multiple treatments, 2022. URL `https://arxiv.org/abs/2205.07836`.

Manudeep Bhuller, Gordon B Dahl, Katrine V Løken, and Magne Mogstad. Incarceration spillovers in criminal and family networks. Working Paper 24878, National Bureau of Economic Research, August 2018.

Manudeep Bhuller, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad. Incarceration, recidivism, and employment. *Journal of Political Economy*, 128(4):1269–1324, 2020. doi: 10.1086/705330. URL `https://doi.org/10.1086/705330`.

Manudeep Bhuller, Laura Khoury, and Katrine Vellesen Løken. Prison, mental health and family spillovers. *NHH Dept. of Economics Discussion Paper*, (19), 2021.

Bernard Black, Eric French, Jeremy McCauley, and Jae Song. The effect of disability insurance receipt on mortality. Technical report, Northwestern University, 2018.

Emely Ek Blæhr and Rikke Søgaard. Instrumental variable-based assessment of the effect of psychotherapy on suicide attempts, health, and economic outcomes in schizophrenia. *Health Economics*, 30:903–914, 2021.

Diana Bonfim and Gil Nogueira. Corporate reorganization as labor insurance in bankruptcy. *SSRN Electronic Journal (January)*, 2021.

Diana Bonfim and Gil Nogueira. Corporate reorganization and the reallocation of labor in bankruptcy. *Available at SSRN 3761746*, 2024.

Marieke Bos, Andrew Hertzberg, Andres Liberman, et al. *Are We Overdiagnosing Mental Illnesses?: Evidence from Randomly Assigned Doctors*. Swedish House of Finance, 2021.

Marieke Bos, Andrew Hertzberg, and Andres Liberman. The effects of diagnosing a young adult with a mental illness: Evidence from randomly assigned doctors. *Available at SSRN 4573307*, 2023.

Tolga Caskurlu. Effects of patent rights on acquisitions and small firm r&d. *Available at SSRN 2427191*, 2019.

Jacelly C Cespedes, Carlos R Parra, and Clemens Sialm. The effect of principal reduction on household distress: Evidence from mortgage cramdown. Technical report, National Bureau of Economic Research, 2021.

David C. Chan, David Card, and Lowell Taylor. Is there a va advantage? evidence from dually eligible veterans. *American Economic Review*, 113(11):3003–3043, 2023.

David C. Chan, Jr., Matthew Gentzkow, and Chuan Yu. Selection with variation in diagnostic skill: Evidence from radiologists. Working Paper 26467, National Bureau of Economic Research, November 2019. URL http://www.nber.org/papers/w26467.

Tom Chang and Antoinette Schoar. Judge specific differences in chapter 11 and firm outcomes. In *AFA 2007 Chicago Meetings Paper*, 2006.

Daniel L Chen. Insiders and outsiders: Does forbidding sexual harassment exacerbate gender inequality? 2010.

Ing-Haw Cheng, Felipe Severino, and Richard R Townsend. How do consumers fare when dealing with debt collectors? evidence from out-of-court settlements. *The Review of Financial Studies*, 34(4):1617–1660, 2021.

Elior Cohen. The effect of housing first programs on future homelessness and socioeconomic outcomes. *Federal Reserve Bank of Kansas City Working Paper*, (22-03), 2022.

Robert Collinson, John Eric Humphries, Nicholas S Mader, Davin K Reed, Daniel I Tannenbaum, and Winnie van Dijk. Eviction and poverty in american cities. Technical report, National Bureau of Economic Research, 2022.

Zack Cooper, Joseph J Doyle Jr, John A Graves, and Jonathan Gruber. Do higher-priced hospitals deliver higher-quality care? Technical report, National Bureau of Economic Research, 2022.

Tomás Cortés, Nicolás Grau Veloso, and Jorge Rivera Cayupi. Juvenile incarceration and adult recidivism. 2019.

Gordon B. Dahl, Andreas Ravndal Kostøl, and Magne Mogstad. Family welfare cultures *. *The Quarterly Journal of Economics*, 129(4):1711–1752, 2014. doi: 10.1093/qje/qju019. URL http://dx.doi.org/10.1093/qje/qju019.

Gustavo De Souza. Employment and welfare effects of the quota for disabled workers in brazil. 2023.

Rafael Di Tella and Ernesto Schargrodsky. Criminal recidivism after prison and electronic monitoring. *Journal of political Economy*, 121(1):28–73, 2013.

Rebecca Diamond, Adam Guren, and Rose Tan. The effect of foreclosures on homeowners, tenants, and landlords. Technical report, National Bureau of Economic Research, 2020.

Stephanie Holmes Didwania. The immediate consequences of federal pretrial detention. *American Law and Economics Review*, 22(1):24–74, 2020.

Christian Dippel and Dustin Frye. The effect of land allotment on native american households during the assimilation era. *Working Paper*, 2020.

Will Dobbie and Jae Song. Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. *American Economic Review*, 105(3):1272–1311, March 2015. doi: 10.1257/aer.20130612. URL http://www.aeaweb.org/articles?id=10.1257/aer.20130612.

Will Dobbie, Paul Goldsmith-Pinkham, and Crystal S Yang. Consumer bankruptcy and financial health. *Review of Economics and Statistics*, 99(5):853–869, 2017.

Will Dobbie, Jacob Goldin, and Crystal S. Yang. The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–40, February 2018a. doi: 10.1257/aer.20161503. URL http://www.aeaweb.org/articles?id=10.1257/aer.20161503.

Will Dobbie, Hans Grönqvist, Susan Niknami, Mårten Palme, and Mikael Priks. The intergenerational effects of parental incarceration. Technical report, National Bureau of Economic Research, 2018b.

Joseph Doyle, John Graves, and Jonathan Gruber. Evaluating measures of hospital quality: Evidence from ambulance referral patterns. *Review of Economics and Statistics*, 101(5):841–852, 2019.

Joseph J. Doyle, Jr. Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, 97(5):1583–1610, December 2007. doi: 10.1257/aer.97.5.1583. URL http://www.aeaweb.org/articles?id=10.1257/aer.97.5.1583.

Joseph J. Doyle, Jr. Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *Journal of Political Economy*, 116(4):746–770, 2008. doi: 10.1086/590216.

Joseph J Doyle Jr. Causal effects of foster care: An instrumental-variables approach. *Children and Youth Services Review*, 35(7):1143–1151, 2013.

Arnaud Dyevre. Public r&d spillovers and productivity growth. Technical report, Working Paper, 22 januari. Te vinden op www. arnauddyevre. com, 2023.

Sarah Eichmeyer and Jonathan Zhang. Pathways into opioid dependence: Evidence from practice variation in emergency departments. *American Economic Journal: Applied Economics*, 14(4): 271–300, 2022.

Sarah Eichmeyer and Jonathan Zhang. Primary care providers' influence on opioid use and its adverse consequences. *Journal of Public Economics*, 217:104784, 2023.

Ozkan Eren and Naci Mocan. Juvenile punishment, high school graduation, and adult crime: Evidence from idiosyncratic judge harshness. *Review of Economics and Statistics*, 103(1):34–47, 2021.

Melissa Famulari. Identifying the causal effects of remedial writing on collegiate performance: Using placement exam grader toughness. Technical report, working paper., University of California, 2012. http://www. econ. ucsd. edu . . . , 2012.

Joan Farre-Mensa, Deepak Hegde, and Alexander Ljungqvist. The bright side of patents. Technical report, National Bureau of Economic Research, 2016.

Henri Fraisse. Households debt restructuring: The re-default effects of a debt suspension. *The Journal of Law, Economics, and Organization*, 33(4):686–717, 2017.

Eric French and Jae Song. The effect of disability insurance receipt on labor supply. *American economic Journal: economic policy*, 6(2):291–337, 2014.

Alberto Galasso and Mark Schankerman. Patents and cumulative innovation: Causal evidence from the courts. *The Quarterly Journal of Economics*, 130(1):317–369, 2015. doi: 10.1093/qje/qju029. URL http://dx.doi.org/10.1093/qje/qju029.

Alberto Galasso and Mark Schankerman. Patent rights, innovation, and firm exit. *The RAND Journal of Economics*, 49(1):64–86, 2018.

Sergio Galletta, Elliott Ash, and Daniel L Chen. Do judicial sentiments affect social attitudes? *Center for Law & Economics Working Paper Series*, 8, 2019.

Patrick Gaule. Patents and the success of venture-capital backed startups: Using examiner assignment to estimate causal effects. *The Journal of Industrial Economics*, 66(2):350–376, 2018.

Jessica W Gillooly. "lights and sirens": Variation in 911 call-taker risk appraisal and its effects on police officer perceptions at the scene. *Journal of Policy Analysis and Management*, 41(3): 762–786, 2022.

Felipe M Gonçalves and Steven Mello. Police discretion and public safety. Technical report, National Bureau of Economic Research, 2023.

Juanita González-Uribe and Santiago Reyes. Identifying and boosting "gazelles": Evidence from business accelerators. *Journal of Financial Economics*, 139(1):260–287, 2021.

Nicolas Grau, Gonzalo Marivil, and Jorge Rivera. The effect of pretrial detention on labor market outcomes. Working Papers wp518, University of Chile, Department of Economics, May 2021. URL https://ideas.repec.org/p/udc/wpaper/wp518.html.

Donald P. Green and Daniel Winik. Using random judge assignment to estimate the effects of incarceration and probation on recidivism among drug offenders. *Criminology*, 48(2):357–387, 2010.

Marie-Pascale Grimon. *Effects of the child protection system on parents*. 2023.

Morten Grindaker, Andreas Kostøl, and Kasper Roszbach. Executive labor market frictions, corporate bankruptcy and ceo careers. *Corporate Bankruptcy and CEO Careers (December 16, 2021)*, 2021.

Arpit Gupta, Christopher Hansman, and Ethan Frenchman. The heavy costs of high bail: Evidence from judge randomization. *The Journal of Legal Studies*, 45(2):471–505, 2016.

Charu Gupta. One product, many patents: Imperfect intellectual property rights in the pharmaceutical industry. *Many Patents: Imperfect Intellectual Property Rights in the Pharmaceutical Industry (June 22, 2023)*, 2023.

David J Harding, Jeffrey D Morenoff, Anh P Nguyen, and Shawn D Bushway. Imprisonment and labor market outcomes: Evidence from a natural experiment. *American Journal of Sociology*, 124(1):49–110, 2018.

David J Harding, Jeffrey D Morenoff, Anh P Nguyen, Shawn D Bushway, and Ingrid A Binswanger. A natural experiment study of the effects of imprisonment on violence in the community. *Nature human behaviour*, 3(7):671–677, 2019.

Ronja Helénsdotter. Surviving childhood: Health effects of removing a child from home. 2023.

Eric Helland and Minjae Yun. More talk, less conflict: Evidence from requiring informal discovery conferences. *American Law and Economics Review*, 25(1):129–189, 2023.

Daniel Herbst. The impact of income-driven repayment on student borrower outcomes. *American Economic Journal: Applied Economics*, 15(1):1–25, 2023.

Randi Hjalmarsson and Matthew J Lindquist. The causal effect of military conscription on crime. *The Economic Journal*, 129(622):2522–2562, 2019.

Colleen Honigsberg and Matthew Jacob. Deleting misconduct: The expungement of brokercheck records. *Journal of Financial Economics*, 139(3):800–831, 2021.

John Eric Humphries, Nicholas S. Mader, Daniel I. Tannenbaum, and Winne L. van Dijk. Does eviction cause poverty? quasi-experimental evidence from cook county. Working Paper 26139, National Bureau of Economic Research, August 2019. URL `https://www.nber.org/papers/w26139`.

John Eric Humphries, Aurelie Ouss, Kamelia Stavreva, Megan Stevenson, and Winnie van Dijk. Conviction, incarceration, and recidivism: Understanding the revolving door, 2023a. URL `https://ssrn.com/abstract=4507597`.

John Eric Humphries, Aurelie Ouss, Kamelia Stavreva, Megan T Stevenson, and Winnie van Dijk. Conviction, incarceration, and recidivism: understanding the revolving door. *Available at SSRN 4507597*, 2023b.

Kristiina Huttunen, Martti Kaila, David Macdonald, and Emily Nix. Financial crimes and the role of prison to reduce them. *Loyola Marymont University Working Paper*, 2022.

Johnny Huynh and David Chan. Do cash transfers narrow health disparities? evidence from veterans with disabilities. *Working Paper*, 2024.

Benjamin G Hyman. Can displaced labor be retrained? evidence from quasi-random assignment to trade adjustment assistance. In *Proceedings. annual conference on taxation and minutes of the annual meeting of the national tax association*, volume 111, pages 1–70. JSTOR, 2018.

Ivan T Ivanov and James Z Wang. Bank supervision and corporate credit supply. *Management Science*, 70(5):3338–3361, 2024.

Zuchuat Jeremy. How vacancy referrals affect job search: Evidence from job applications data. *Available at SSRN 4366729*, 2023.

Andrew Jordan and Taeho Kim. Strengthening police oversight: the impacts of misconduct investigators on police officer behavior. *Available at SSRN 4099052*, 2022.

Andrew Jordan, Ezra Karger, and Derek Neal. Heterogeneous impacts of sentencing decisions. 2023.

Felipe Diaz Klaassen. Crime and (monetary) punishment, 2021.

Jeffrey R. Kling. Incarceration length, employment, and earnings. *American Economic Review*, 96(3):863–876, June 2006. doi: 10.1257/aer.96.3.863. URL http://www.aeaweb.org/articles?id=10.1257/aer.96.3.863.

Michael LaForest. Parole supervision at the margins. In *2022 APPAM Fall Research Conference*. APPAM, 2022.

Logan M Lee. Halfway home? residential housing and reincarceration. *American Economic Journal: Applied Economics*, 15(3):117–149, 2023.

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

Charles E. Loeffler. Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology*, 51(1):137–166, 2013.

Kristian Lum, Erwin Ma, and Mike Baiocchi. The causal impact of bail on case outcomes for indigent defendants in new york city. *Observational Studies*, 3(1):38–64, 2017.

Nicole Maestas, Kathleen J. Mullen, and Alexander Strand. Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *American Economic Review*, 103(5):1797–1829, August 2013. doi: 10.1257/aer.103.5.1797. URL http://www.aeaweb.org/articles?id=10.1257/aer.103.5.1797.

Paulo Manoel and Vinicius Augusto Brunassi Silva. Creditor voting and time constraints in bankruptcy reorganizations. *Creditor Voting and Time Constraints in Bankruptcy Reorganizations (August 26, 2021)*, 2021.

Armando Meier, Jonathan Levav, and Stephan Meier. Early release and recidivism. 2020.

Jacob Moscona. The management of aid and conflict in africa. *Available at SSRN 4575081*, 2023.

Michael Mueller-Smith. The criminal and labor market impacts of incarceration. *Working Paper*, 2015.

Sendhil Mullainathan and Ziad Obermeyer. Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727, 2022.

David Munroe and Laurence H Wilse-Samson. Foreclosure contagion: Measurement and mechanisms. *Available at SSRN 4349905*, 2013.

Rachel Nesbit. The role of mandated mental health treatment in the criminal justice system. Technical report, 2023.

Samuel Norris, Matthew Pecenco, and Jeffrey Weaver. The effects of parental and sibling incarceration: Evidence from ohio. *American Economic Review*, 111(9):2926–2963, 2021.

Yuval Ofek-Shanny, Avner Strulov-Shlain, and Dan Zeltzer. Impacts of home-care subsidies: Evidence from quasi-random assignment. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2023-139), 2023.

Hiroko Okudaira. The economic costs of court decisions concerning dismissals in japan: Identification by judge transfers. *International Review of Law and Economics*, 53:60–75, 2018.

Aurelien Quignon. Crowd-based feedback and early-stage entrepreneurial performance: Evidence from a digital platform. *Research Policy*, 52(7):104792, 2023.

George Botros Zarif Rateb. *The Case for Clemency: Differential Impacts of Pretrial Detention on Case and Crime Outcomes*. PhD thesis, Duke University, 2022.

Beatriz Ribeiro and Claudio Ferraz. Pretrial detention and rearrest: Evidence from brazil. *Working Paper*, 2019.

Alexander Rieber and Steffen Schechinger. Herding Behavior between Rating Agencies. VfS Annual Conference 2019 (Leipzig): 30 Years after the Fall of the Berlin Wall - Democracy and Market Economy 203580, Verein für Socialpolitik / German Economic Association, 2019. URL `https://ideas.repec.org/p/zbw/vfsc19/203580.html`.

Michael A Roach and Max M Schanzenbach. The effect of prison sentence length on recidivism: Evidence from random judicial assignment. *Northwestern Law & Econ Research Paper*, (16-08), 2015.

Bhaven Sampat and Heidi L Williams. How do patents affect follow-on innovation? evidence from the human genome. Working Paper 21666, National Bureau of Economic Research, October 2015. URL `http://www.nber.org/papers/w21666`.

David Silver and Jonathan Zhang. Impacts of basic income on health and economic well-being: Evidence from the va's disability compensation program. 2022.

David White Silver and Jonathan Zhang. *Invisible wounds: Health and well-being impacts of mental disorder disability compensation on veterans*. National Bureau of Economic Research, 2023.

Megan T Stevenson. Distortion of justice: How the inability to pay bail affects case outcomes. *The Journal of Law, Economics, and Organization*, 34(4):511–542, 2018.

Megan T Stevenson, Aurelie Ouss, Winnie van Dijk, John Eric Humphries, and Kamelia Stavreva. Conviction, incarceration, and recidivism: Understanding the revolving door. *Available at SSRN 4507597*, 2023.

Bas van der Klaauw and Heike Vethaak. Empirical evaluation of broader job search requirements for unemployed workers. Technical report, Tinbergen Institute Discussion Paper, 2022.

Hilde T Wermink, AAJ Blokland, J Been, PM Schuyt, N Tollenaar, and R Apel. Estimating effects of short-term imprisonment on crime using random judge assignments. *Justice quarterly*, 41(3): 317–346, 2024.

Ariel White. Misdemeanor disenfranchisement? the demobilizing effects of brief jail spells on potential voters. *American Political Science Review*, 113(2):311–324, 2019.

Jenny Williams and Don Weatherburn. Can electronic monitoring reduce reoffending? *Review of Economics and Statistics*, 104(2):232–245, 2022.

Michelle Yin, Garima Siwach, and Dajun Lin. Vocational rehabilitation services and labor market outcomes for transition-age youth with disabilities in maine. *Journal of Policy Analysis and Management*, 42(1):166–197, 2023.

Mariyana Zapryanova. The effects of time in prison and time on parole on recidivism. *The Journal of Law and Economics*, 63(4):699–727, 2020.