Journal of Economic Literature 2025, 63(2), 401–439
https://doi.org/10.1257/jel.20241719

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

Eric Chyn, Brigham Frandsen, and Emily Leslie*

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 this research design, outline the assumptions under which the design identifies causal effects, describe empirical tests of the conditions for identification, and discuss trade-offs associated with choices researchers must make for estimation. We demonstrate concepts and best practices in an empirical case study that uses an examiner tendency research design to study the effects of pretrial detention. (*JEL* C21, C26, K14, K41)

1. *Introduction*

In 1932, criminologists in New Jersey documented wide disparities in the sentencing tendencies of trial judges. The most severe judge imprisoned 57.7 percent of the convicted defendants randomly assigned to their courtroom, and the most lenient only 33.6 percent (Gaudet, Harris, and St John 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 nonexperimental settings. In pioneering work, Kling (2006) estimated the impact of incarceration length on post-release labor market earnings by leveraging plausibly exogenous variation

*Chyn: Department of Economics, University of Texas at Austin, and the National Bureau of Economic Research. Frandsen: Department of Economics, Brigham Young University. Leslie: Department of Economics, Brigham Young University. 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.

† Go to https://doi.org/10.1257/jel.20241719 to visit the article page and view author disclosure statement(s).

in sentencing arising from the rules that assign offenders to judges.[1] Often referred to as the "judge fixed effects" or "judge leniency" design, this approach—hereafter called the "examiner tendency" design—has been used in at least 136 studies across a variety of settings involving different types of decision-makers (see Supplemental Appendix Table A1). Recent examples include studies of the effects of pretrial 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 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 controlled trial (RCT).

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 examiner tendency designs. We aim to clarify the conditions under which examiner tendency designs 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 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—*pairwise monotonicity*—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 without strong conditions on how outcomes respond or how examiners decide treatments**. 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,

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

difficult to motivate—conditions on how examiners allocate individuals to treatments (Humphries et al. 2024; Bhuller and Sigstad 2024).

- **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 biases due to 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, Leslie, and McIntyre 2023).
- **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. 2023).
- **While the conditions for identification ultimately rest on institutional and economic foundations, specification tests can empirically examine whether they are plausible**. 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. Recently proposed procedures test whether exclusion and monotonicity conditions hold when effects are heterogeneous, including Kitagawa (2015), Norris, Pecenco, and Weaver (2021), and Frandsen, Lefgren, and Leslie (2023). 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 for discussing 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, 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 identifying 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 pretrial 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 basic 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 for discussing most of the practical issues around identification, estimation, and inference. In Section 5, we consider additional issues that arise when treatment effects are heterogeneous. The estimation and inference approaches that we propose for the simpler constant treatment effects case carry over to the more general heterogeneous treatment effects setting.

2.1 *A Basic Econometric Model*

We seek to estimate the effects of a binary treatment, such as pretrial 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 $Y_i = Y_i(0) + \big(Y_i(1) - Y_i(0)\big)D_i$, 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 $\varepsilon_i = Y_i(0) - E\big[Y_i(0)\big]$, $\alpha = E\big[Y_i(0)\big]$, and $E\big[Y_i(0)\big]$ denotes the expected value of outcomes in the untreated state.

Despite the simplicity of the model, estimating 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 $p(j) = E\big[D_i(j)\big]$. 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—that is, 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 pretrial 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 convicted. This correlation is why an ordinary least squares (OLS) 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 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* $\min\{\Pr(J_i = j), \Pr(J_i = j')\} \geq \mu$.

[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 to incorporate covariates.

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:

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

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 a 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 pretrial 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 over these competing values using the following utility function:

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

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 $c_j \geq \max\{-a_j, -b_j\}$ reflects the intuition that judges prefer correct

[3] The nontrivial 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 nontrivial function of the instrument. That is, $\mathbb{E}[D_i|Z_i = w]$ is a nontrivial function with respect to values w in the support of Z_i.

[4] See Canay, Mogstad, and Mountjoy (2024) for an alternative model of examiner decision-making that is based on a generalized Roy model (Heckman and Vytlacil 2005).

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:

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

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

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

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 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

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

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]

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

[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.

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 *A Two-Stage Least Squares*

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

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

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. The associated standard errors can be estimated using common statistical packages.

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 toward the OLS estimand (i.e., $Cov(Y_i, D_i)/Var(D_i)$). The bias of 2SLS in this case is an example 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

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

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 $k/n \rightarrow \kappa$ 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 JIVE (Angrist, Imbens, and Krueger 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, Kostøl, and Mogstad (2014) estimate the leniency of the disability

[7] Over 90 percent of the studies we survey in Supplemental Appendix Table A1 used a jackknife or leave-out procedure for calculating the examiner propensity measure.

insurance examiner assigned to each case by calculating the examiner's tendency among all their other cases. 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. In this case, 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 toward 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. Frandsen, Leslie, and McIntyre (2023) provide detailed discussion of this estimator. Denoting the set of observations in individual i's cluster as $\mathcal{C}_i$, the CJIVE fitted value is defined as: $\hat{p}_i^{CJIVE} = Z_i' \hat{\pi}_{-\mathcal{C}_i}$, 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, since 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 we explore 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. Including 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 (see Section 6). Factors that may be affected by treatment or judge assignment should not be included in X_i because their inclusion 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 redefine $\varepsilon_i = Y_i(0) - \mathbb{E}\left[Y_i(0) \mid X_i\right]$.

The presence of covariates complicates the leave-out or jackknife remedy for many-instruments bias discussed above. Two recent estimators adapt JIVE to the case with covariates: the unbiased 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 consistent. 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 that 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 be misleading when

[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.

there are few cases per examiner. Intuitively, the reason 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 sufficiently many cases. For example, the case study in Section 7 restricts the sample 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 of 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. (2023), 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, 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.[12] In the design-based approach to inference recommended by Abadie et al. (2020) and Abadie et al. (2023), 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 to treatment. In some IV settings, the conventional asymptotic approximations break down under weak identification and the usual standard errors can yield misleading

[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 error 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 of clustering at the examiner level include Dobbie, Goldin, and Yang (2018) and Bald et al. (2022).

inference (Andrews, Stock, and Sun 2019b; Mikusheva and Sun 2021). This section discusses when weak identification is likely to cause practical problems and how to address weak identification in 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—is very high.[14] 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 screening on the sign of the estimated first stage. In our case, this means proceeding with the analysis only if the covariance between treatment status and the JIVE instrument is positive, that is, $\widehat{Cov}\left(D_i, \hat{p}_i^{JIVE}\right) > 0$. 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]

[13] Although the underlying examiner dummies are many, the JIVE fitted value (i.e., $\hat{p}_i^{JIVE}$) can be treated like a single instrument under certain conditions. See Bhuller et al. (2020) for additional discussion.

[14] With either heteroskedasticity or dependence due to clustering, note that the degree of endogeneity is not simply the correlation between ν_i and ε_i.

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

[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.

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-statistic 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 determine 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 use simulations to illustrate 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 $D_i = 1\left(\Phi(v_i) \geq \tau_j\right)$, where Φ is the standard normal cumulative distribution function (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 and the function q is determined by Φ. The simulation assumes that the 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 $\delta = 0.3$.

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: (i) 2SLS using judge dummies; (ii) JIVE; (iii) 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); (iv) JIVE, screening on having a positive first-stage coefficient, an approach recommended for IV from Angrist and Kolesár (2024).

Figure 1 illustrates 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 for 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

[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.

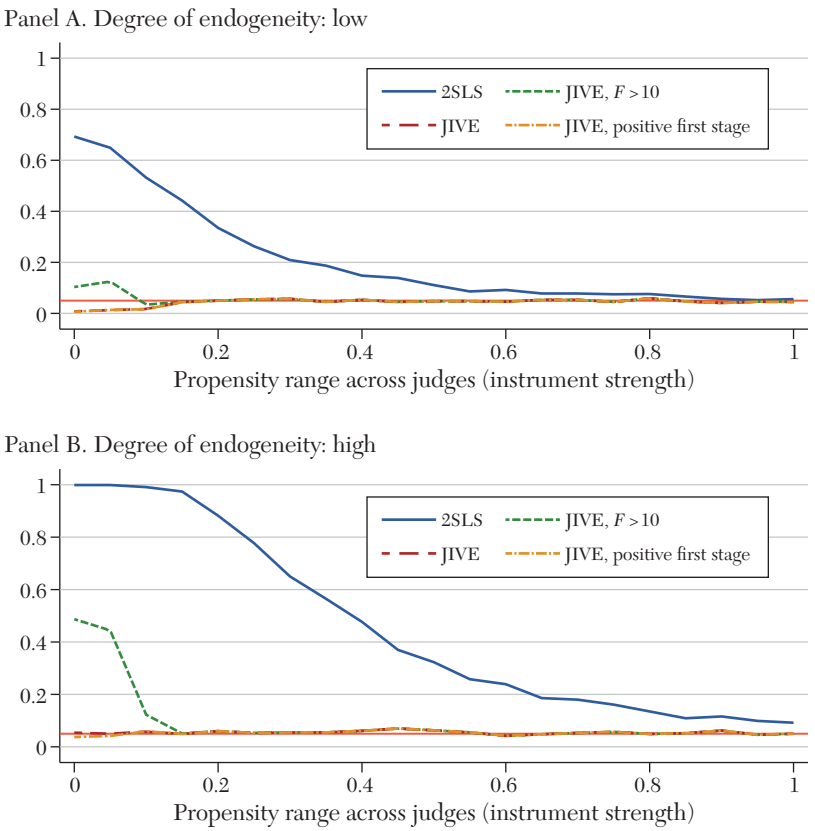

Figure 1. Weak Instrument Simulation Exercise: IV Rejection Rate, Nominal 5 Percent Test

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 $\rho = 0.6$.

examiner tendency design simulation results provided in Bhuller et al. (2020, supplemental 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 rat- when identification is weak, an illustration of the well-known inference distortion with weak instruments (Andrews, Stock, and Sun 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]

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

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 approach: 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 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 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 this 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 scenario. 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 nonnegative weights w_i.

Heterogeneous treatment effects have important implications for interpreting 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:

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

As discussed in Frandsen, Lefgren, and Leslie (2023), a setting that 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:

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

[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.

where λ_j is the probability of being assigned to examiner j, $p(j)$ is the examiner propensity to treat, $\bar{p}$ is the average propensity across all examiners ($\bar{p} = \sum_{j=1}^{k} \lambda_j p_j$), 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: $\sum_{j=1}^{k} \lambda_j \big(p(j) - \bar{p}\big)\big(D_i(j) - \bar{D}_i\big)$, which is proportional to the correlation across examiners between 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: those whom some examiners would assign to treatment and others would not. An examiner with an above average treatment propensity ($p(j) > \bar{p}$) 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 ($p(j) < \bar{p}$) 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., $\lambda_1 = \lambda_2 = 0.5$) and the treatment propensities for examiners 1 and 2 are $p(1) = 0.75$ and $p(2) = 0.25$, respectively. This implies $p(1) - \bar{p} = 0.25$ and $p(2) - \bar{p} = -0.25$. 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 $\lambda_j \big(p(j) - \bar{p}\big)\big(D_i(j) - \bar{D}_i\big) < 0$ 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* $j, \ell \in \{0, \ldots, k\}$, *either* $D_i(j) \geq D_i(\ell)$ *or* $D_i(j) \leq D_i(\ell)$ *for each individual* i.

[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 whom all examiners would always incarcerate).

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 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 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 conditions hold.

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

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

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 treatment (i.e., v_{ij} can vary across examiners), as long as these disagreements are not extensive enough to make anyone's treatment status negatively correlated with

[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, Bertrand, and Mullainathan 2012).

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

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 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 supplemental 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, Santos, and Torgovitsky 2018). When average monotonicity holds alone, marginal treatment effects are no longer identified.

Chan, Gentzkow, and Yu (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 that 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 than the three alternatives to pairwise monotonicity described above.[23]

[23] Sigstad (2023) studies judicial panels in several settings and provides empirical evidence that suggests average monotonicity is a more realistic condition even in settings where pairwise monotonicity is frequently violated. As we note in the

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, Gentzkow, and Yu (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. Identifying MTEs requires a monotonicity condition that holds for every pair of examiners (e.g., conventional pairwise monotonicity or the compliers–defiers condition). As we will show, MTEs are not identified when pairwise monotonicity fails.

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, Tobias, and Vytlacil (2001), Heckman, Urzua, and Vytlacil (2006), and Heckman and Vytlacil (2007) is defined as the average treatment effect 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 pretrial detention example, the MTEs give the effects of pretrial 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 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 the ATE will implicitly extrapolate beyond the support of observed propensities.[24]

conclusion, further research assessing the plausibility of monotonicity conditions and the magnitude of bias due to violations remains an ongoing topic for future research.

[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.

Heckman and Vytlacil (2005) show that, under pairwise monotonicity and strict exclusion, these marginal treatment effects are identified provided there is 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. In other words, this parameter is 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:

$$\delta^{LATE}(p',p) = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{P}(D_i = 1|Z_i = 1) - \mathbb{P}(D_i = 1|Z_i = 0)}$$

$$= \frac{\mathbb{E}[Y_i|p(J_i) = p'] - \mathbb{E}[Y_i|p(J_i) = p]}{p' - p}.$$

With this expression for the LATE in mind, the MTE is intuitively identified by comparing outcomes for individuals assigned to examiners whose propensities to administer treatment are 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, Lefgren, and Leslie (2023), 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

[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.

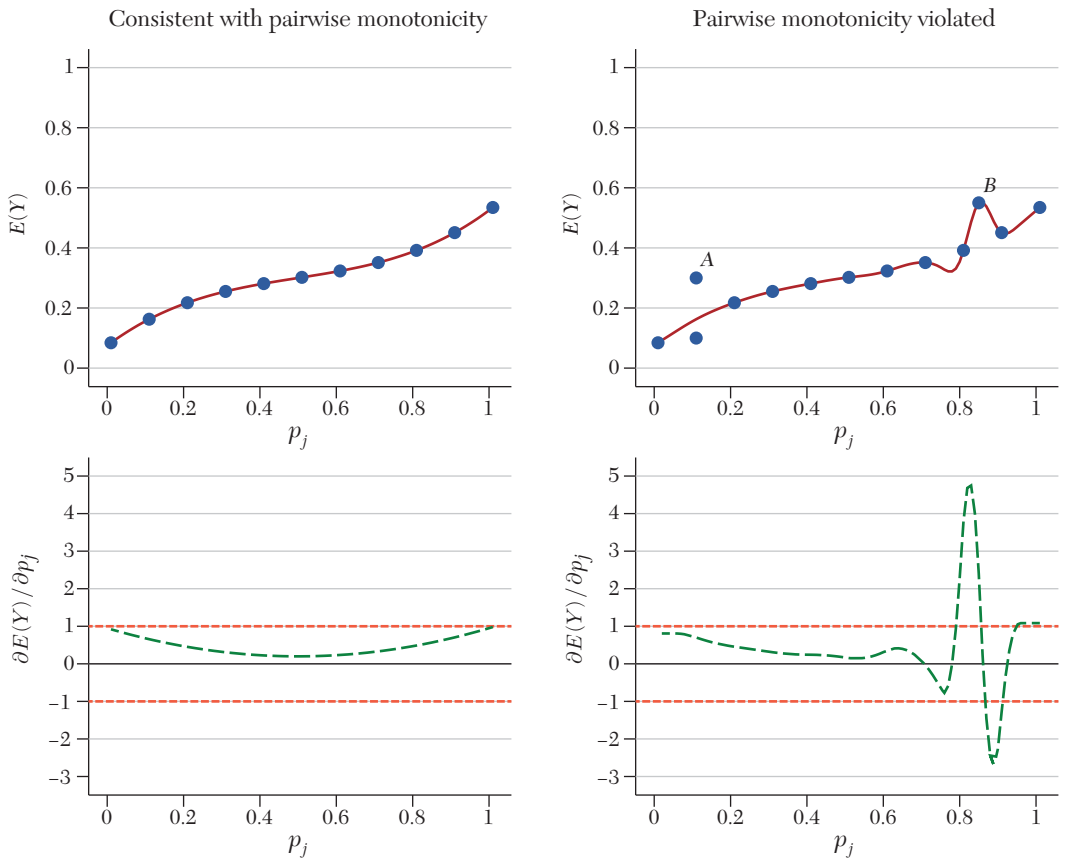

Figure 2. Marginal Treatment Effects Illustration

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).

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 would take on impossible values, that is, 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, 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 can no longer be interpreted as 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 be 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 proceed much like that for the effects of a binary treatment. Jackknife IV with examiner dummies as excluded instruments 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.

[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.

[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 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 pretrial 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 even though there are four distinct categories of defendants based on whether or not individuals are detained pretrial 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 pretrial 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, Lefgren, and Leslie (2023)—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 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]

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

[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 simultaneously instrumented for all treatments using examiner indicators.

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 (2024) give the conditions under which linear 2SLS with several endogenous treatments recovers proper weighted averages of treatment effects, and Humphries et al. (2024) 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. Supplemental appendix B describes the conditions required in the Bhuller and Sigstad (2024) and Humphries et al. (2024) frameworks. In general, the two frameworks are distinct, but we illustrate the stringency of the conditions in the case where they are equivalent: when there are 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 (2024) and Humphries et al. (2024) 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 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 (2024) and Humphries et al. (2024) are 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,

[31] IV methods beyond linear 2SLS can identify treatment effects in the discrete choice models discussed by Heckman, Urzua, and Vytlacil (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. (2024) 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 Supplemental Appendix B.

the requirements can be limiting in terms of examiner decision-making patterns. For example, threshold-crossing models with a single unobservable to determine treatment can be sufficient for identification (see Bhuller and Sigstad 2024 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 among 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*

Estimates from examiner tendency designs only have a causal interpretation when several identifying conditions hold. 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. 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 pretreatment 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, Hausman, and Newey 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 (e.g., 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 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, Lefgren, and Leslie (2023). 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, Lefgren, and Leslie (2023), 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 preexisting 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, this test is equivalent to examining 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 2024; Humphries et al. 2024) and established best practices have not yet emerged.

7. *Case Study: Effects of Pretrial Detention*

In this section, we provide a concrete guide to implementing our suggested best practices using an empirical example that analyzes the effects of pretrial 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, Goldin, and Yang (2018), we use a sample of court records from misdemeanor and felony cases in Miami–Dade County, Florida, over the period 2006–2014 (Chyn, Frandsen, and Leslie 2025). Following arrest, defendants in Miami–Dade were brought to a police station where they could secure pretrial release by posting bail that varied based on the seriousness of their offense. The 70 percent of defendants who did 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, Goldin, and Yang (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, our main outcome of interest.

For our analysis, we follow Dobbie, Goldin, and Yang (2018) and restrict our attention to cases assigned to a weekend bail hearing because these are cases where bail judges are assigned on a rotating schedule. In the main analysis, we restrict the sample to cases that 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 cases, presided over by 146 unique judges.

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

TABLE 1
Pretrial Detention Case Study: Defendant-Level Summary Statistics

	Unweighted			Estimate
	Full sample (1)	Released (2)	Detained (3)	Complier weighted (4)
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	0.55
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
Observations	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 in column 4 presents complier-weighted averages for any guilty offense (i.e., conviction), which is a posttreatment outcome variable. For this measure, we report the estimated mean in the untreated state (i.e., the share of compliers who are convicted when they are not released pretrial).

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.

[34] Defendants released before trial are also more likely to have a felony or violent offense. This finding that release is 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 criterion for pretrial release. Given this objective, released defendants may not necessarily be a group of defendants who are associated with more severe charges.

We are interested in studying the causal effects of pretrial release, 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 and court-by-month-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. In words, based on our definition of X_i, our identifying assumption is that judges are randomly assigned within court, year, month, and day-of-week groups. Other case characteristics are omitted from X_i 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 posttreatment 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, 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

[35] It is straightforward to use a 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.

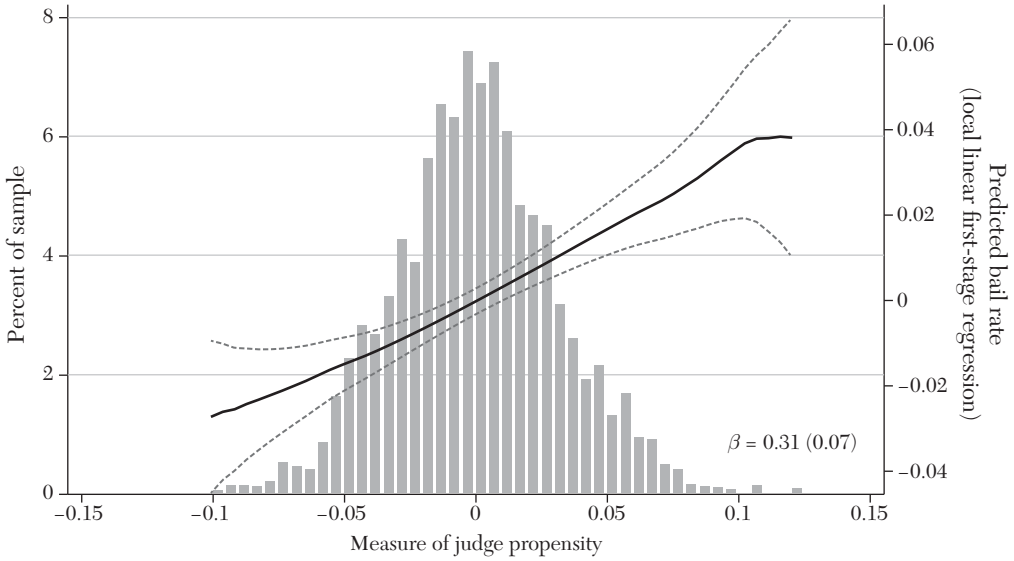

Figure 3. Distribution of Judge Leniency Measure

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.

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 pretrial release when there is treatment effect heterogeneity and the conditions of instrument exogeneity, exclusion, and monotonicity hold. This parameter represents causal impacts among the subset of complier defendants who would be released by lenient judges but not by strict judges. We next undertake a series of exercises to examine whether the usual identifying conditions necessary for judge research designs are plausible.

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 pretrial 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 pretrial release still have observable differences in baseline characteristics even after controlling for court-by-time fixed

[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 value 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.

TABLE 2
ASSESSING BALANCE

	Treatment (1)	Leniency measure (2)
Male	−10.097	−0.019
	(0.404)	(0.033)
Black	−2.251	−0.027
	(0.296)	(0.025)
Age	−0.310	−0.002
	(0.012)	(0.001)
Prior offender	−17.371	−0.005
	(0.308)	(0.024)
Number of counts	−2.137	0.015
	(0.131)	(0.011)
Felony charge	24.596	−0.586
	(9.817)	(0.643)
Drug charge	2.201	0.055
	(0.426)	(0.038)
Violent charge	14.565	0.012
	(0.420)	(0.033)
Property charge	−11.428	0.052
	(0.374)	(0.029)
Joint *F*-stat	1,160.491	1.349
p-value	0.000	0.207
Observations	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 pretrial release. Column 1 reports results from a linear probability model with pretrial 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.

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.

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

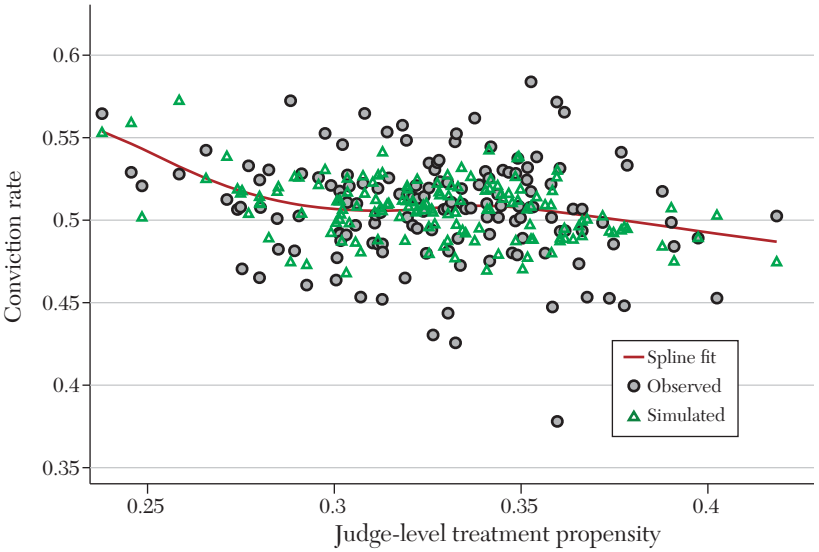

Figure 4. Illustration of Test of Pairwise Monotonicity and Strict Exclusion

Notes: This figure provides an illustration of the joint test of strict exclusion and pairwise monotonicity recommended in the Frandsen, Lefgren, and Leslie (2023). The *y*-axis reports conviction rates while the *x*-axis reports judge-level treatment propensities. The dots (in gray) correspond to the observed conviction and treatment propensity in our sample after controlling for court-by-time fixed effects. The test proposed in Frandsen, Lefgren, and Leslie (2023) is based on fitting a flexible spline function to the observed data on conviction rates and treatment propensities. The solid line (in red) shows the predicted values of the spline function fit to the observed data. Intuitively, the test examines two conditions: (i) whether the fitted function meets slope restrictions implied by the range of possible treatment effect sizes; (ii) 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 1 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, Lefgren, and Leslie (2023).

more lenient judges would detain, perhaps because more lenient judges may be stricter for particular groups of defendants. The joint test proposed by Frandsen, Lefgren, and Leslie (2023) 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: (i) whether the slope of the fitted function is impossibly large because it exceeds the range of possible treatment effects sizes, and (ii) 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 red 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 for 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 are 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, Lefgren, and Leslie (2023), 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 pretrial 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. Since we find no evidence violating 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.

We report our main results on the effects of pretrial release 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 larger in magnitude than the OLS results. The final two columns report our preferred results that use the CJIVE estimator to leave out each defendant's cluster (shift) in

TABLE 3
First Stage Analysis for Pretrial Release

	Full sample (1)	Male (2)	Black (3)	Prior offender (4)	Any drug (5)	Any violent (6)	Any property (7)	Felony case (8)
$\hat{p}_i^{CJIVE}$	0.314	0.248	0.326	0.288	0.302	0.069	0.269	0.161
	(0.067)	(0.069)	(0.083)	(0.078)	(0.120)	(0.108)	(0.090)	(0.088)
Observations	91,282	77,087	47,861	51,394	26,189	19,312	33,047	47,927

Notes: This table is an analysis of the first-stage impact of judge leniency on pretrial 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 Pretrial Release

	OLS (1)	Judge dummies (2)	IJIVE (3)	UJIVE (4)	CJIVE (5)	CJIVE (6)
Released	−0.232	−0.272	−0.294	−0.293	−0.432	−0.500
	(0.007)	(0.066)	(0.107)	(0.109)	(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 pretrial 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.

the calculation of the judge leniency measure. The point estimate in column 5 indicates that pretrial release reduces conviction rates by 43.2 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 −0.500 with a more narrow confidence interval.[37]

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 toward OLS when there are many judges. In addition, the pattern of

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

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 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 nonrandom 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 versus 1.63), have a much lower likelihood of being charged with a felony offense (0.28 versus 0.52), and are much less likely to be charged with a violent crime (0.05 versus 0.21).[39] The last row of Table 1 reports the estimated share of compliers who would be convicted if they had not been released, revealing that 96 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 relative to staying behind bars for an indeterminate length while they await trial.[40]

As a final exercise, we conduct sensitivity analysis in our pretrial 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 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.

[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, Goldin, and Yang (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 pretrial 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 versus 49 days), consistent with the possibility that detaining people faced with low-level charges pretrial induces them to accept plea deals relatively quickly.

TABLE 5
Robustness to Varying Sample Restriction

	50 cases/judge (1)	100 cases/judge (2)	200 cases/judge (3)	300 cases/judge (4)
Released	−0.428	−0.486	−0.432	−0.426
	(0.198)	(0.207)	(0.203)	(0.203)
Observations	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 pretrial 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.

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, interpreting the 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 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's (2023) recent study makes progress in this direction. 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 2SLS estimates are not severely biased due to violations of the traditional monotonicity condition, 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. (2024), Rivera (2023), and Kamat, Norris, and Pecenco (2024) 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.

References

Abadie, Alberto. 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics* 113 (2): 231–63.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1): 265–96.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" Quarterly *Journal of Economics* 138 (1): 1–35.

Abrams, David S., Marianne Bertrand, and Sendhil Mullainathan. 2012. "Do Judges Vary in Their Treatment of Race?" *Journal of Legal Studies* 41 (2): 347–83.

Ackerberg, Daniel A., and Paul J. Devereux. 2009. "Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity." *Review of Economics and Statistics* 91 (2): 351–62.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in IV Regression: Theory and Practice." *Annual Review of Economics* 11: 727–53.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90 (430): 431–42.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometrics* 14 (1): 57–67.

Angrist, Joshua, and Michal Kolesár. 2024. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *Journal of Econometrics* 240 (2): 105398.

Bald, Anthony, Eric Chyn, Justine Hastings, and Margarita Machelett. 2022. The Causal Impact of Removing Children from Abusive and Neglectful Homes." *Journal of Political Economy* 130 (7): 1919–62.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62 (3): 657–81.

Bhuller, Manudeep, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad. 2020. "Incarceration, Recidivism, and Employment." *Journal of Political Economy* 128 (4): 1269–1324.

Bhuller, Manudeep, and Henrik Sigstad. 2024. "2SLS with Multiple Treatments." *Journal of Econometrics* 242 (1): 105785.

Canay, Ivan A., Magne Mogstad, and Jack Mountjoy. 2024. "On the Use of Outcome Tests for Detecting Bias in Decision Making." *Review of Economic Studies* 91 (4): 2135–67.

Chyn, Eric, Brigham Frandsen, and Emily C. Leslie. 2025. *Data and Code for: "Examiner and Judge Designs in Economics: A Practitioner's Guide."* American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E209883V1.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *Quarterly Journal of Economics* 137 (2): 729–83.

Dahl, Gordon B., Andreas Ravndal Kostøl, and Magne Mogstad. 2014. "Family Welfare Cultures." *Quarterly Journal of Economics* 129 (4): 1711–52.

de Chaisemartin, Clément. 2017. Tolerating Defiance? Local Average Treatment Effects without Monotonicity." *Quantitative Economics* 8 (2): 367–96.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review* 108 (2): 201–40.

Evdokimov, Kirill S., and Michal Kolesár. 2019. Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects." Unpublished.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review* 113 (1): 253–77.

Frandsen, Brigham, Emily Leslie, and Samuel McIntyre. 2023. "Cluster Jackknife Instrumental Variables Estimation." Unpublished.
Gaudet, Frederick J., George S. Harris, and Charles W. St John. 1932. "Individual Differences in the Sentencing Tendencies of Judges." *American Institute of Criminal Law and Criminology* 23: 811–18.
Hansen, Christian, Jerry Hausman, and Whitney Newey. 2008. "Estimation with Many Instrumental Variables." *Journal of Business and Economic Statistics* 26 (4): 398–422.
Heckman, James J., and Rodrigo Pinto. 2018. "Unordered Monotonicity." *Econometrica* 86 (1): 1–35.
Heckman, James, Justin L. Tobias, and Edward Vytlacil. 2001. "Four Parameters of Interest in the Evaluation of Social Programs." *Southern Economic Journal* 68 (2): 211–223.
Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics* 88 (3): 389–432.
Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73 (3): 669–738.
Heckman, James J., and Edward J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics*, Vol. 6B, edited by James J. Heckman and Edward E. Leamer, 4779–4874. Elsevier.
Humphries, John Eric, Aurelie Ouss, Kamelia Stavreva, Megan T. Stevenson, and Winnie van Dijk. 2024. "Conviction, Incarceration, and Recidivism: Understanding the Revolving Door." NBER Working Paper 32894.
Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.
Imbens, Guido W., and Donald B. Rubin. 1997. "Estimating Outcome Distributions for Compliers in Instrumental Variables Models." *Review of Economic Studies* 64 (4): 555–74.
Kamat, Vishal, Samuel Norris, and Matthew Pecenco. 2024. "Conviction, Incarceration, and Policy Effects in the Criminal Justice System." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4777635.
Kitagawa, Toru. 2015. "A Test for Instrument Validity." *Econometrica* 83 (5): 2043–63.
Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96 (3): 863–76.
Kolesár, Michal. 2013. "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity." Unpublished.
Kolesár, Michel, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens. 2015. "Identification and Inference with Many Invalid Instruments." *Journal of Business and Economic Statistics* 33 (4): 474–84.
Lee, David S., Justin McCrary, Marcelo J. Moreira, Jack R. Porter, and Luther Yap. 2023. "What to Do When You Can't Use '1.96' Confidence Intervals for IV." NBER Working Paper 31893.
Lee, Sokbae, and Bernard Salanié. 2018. "Identifying Effects of Multivalued Treatments." *Econometrica* 86 (6): 1939–63.
Mikusheva, Anna, and Liyang Sun. 2021. "Inference with Many Weak Instruments." *Review of Economic Studies* 89 (5): 2663–86.
Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters." *Econometrica* 86 (5): 1589–1619.
Norris, Samuel, Matthew Pecenco, and Jeffrey Weaver. 2021. "The Effects of Parental and Sibling Incarceration: Evidence from Ohio." *American Economic Review* 111 (9): 2926–63.
Rivera, Roman. 2023. "Release, Detain, or Surveil?" Unpublished.
Sargan, J. D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica* 26 (3): 393–415.
Sigstad, Henrik. 2023. "Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs." http://dx.doi.org/10.2139/ssrn.4534809.
Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–86.
Waldfogel, Joel. 1995. "The Selection Hypothesis and the Relationship between Trial and Plaintiff Victory." *Journal of Political Economy* 103 (2): 229–60.
Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press.