Judging Judge Fixed Effects

Brigham R. Frandsen[*] Lars J. Lefgren[*] Emily C. Leslie[*]

January 30, 2019

Abstract

We propose a test for the identifying assumptions invoked in designs based on random assignment to one of many "judges." We show that standard identifying assumptions imply that the conditional expectation of the outcome given judge assignment is a continuous function with bounded slope of the judge propensity to treat. The implication leads to a two-part test that generalizes the Sargan-Hansen overidentification test and assesses whether implied treatment effects across the range of judge propensities are possible given the domain of the outcome. We show the asymptotic validity of the testing procedure, demonstrate its finite-sample performance in simulations, and apply the test in an empirical setting examining the effects of pre-trial release on defendant outcomes in Miami. When the assumptions are not satisfied, we propose a weaker average monotonicity assumption under which IV still converges to a proper weighted average of treatment effects.

1 Introduction

Examining the impact of incarceration length on subsequent labor market earnings, Kling (2006) leveraged plausibly exogenous variation in sentence length arising from the random assignment of offenders to judges. Specifically, he instrumented the offender's realized sentence with the average sentence of all other offenders who faced the same judge. The paper pioneered a now widespread methodology, dubbed the

[*] Department of Economics, Brigham Young University

We thank Jeffrey Denning, Kirill Evdokimov, Giovanni Mellace, and seminar participants at Brigham Young University, MIT, and Dartmouth University for helpful feedback. We thank Crystal Yang for help with gaining access to data from Miami-Dade County. We thank the College of Family, Home, and Social Science of Brigham Young University for generous financial support.

"judged fixed effects" design, in which the exogenous assignment to a judge, administrator, or other decision maker identifies the effects of some treatment on outcomes. Recent examples employing this strategy include the effect of incarceration on economic and family outcomes (Green and Winik, 2010; Loeffler, 2013; Aizer and Doyle, 2015; Mueller-Smith, 2015; Bhuller et al., 2016; Eren and Mocan, 2017; Arteaga, 2018; Norris et al., 2018; Bhuller et al., 2018; Dobbie et al., 2018b), the effect of pretrial detention on a variety of legal and economic outcomes (Gupta et al., 2016; Leslie and Pope, 2017; Dobbie et al., 2018a), the effect of consumer bankruptcy on household financial well-being (Dobbie and Song, 2015; Dobbie et al., 2017), the effect of bankruptcy on firm outcomes (Chang and Schoar, 2013), the effect of foster care on child outcomes (Doyle, 2007, 2008), the effect of disability on labor supply, mortality, and intergenerational welfare use (Maestas et al., 2013; Dahl et al., 2014; Autor et al., 2017; Black et al., 2018), and the effect of patents on innovation (Galasso and Schankerman, 2015; Sampat and Williams, 2015).

The judge fixed effects design is an example of a more general estimation scenario in which one or more instrumental variables are used to estimate the causal effects of binary treatment variable. Examples include experiments making use of a variety of information nudges to influence program take-up or enrollment (Hastings and Weinstein, 2008; Barghava and Manoli, 2015; Bergman et al., 2017) and randomized controlled trials with multiple intervention arms all aimed an increasing participation in a single treatment (Olken, 2007; McKenzie et al., 2008; Thornton, 2008).It is crucial that the treatment variable of interest be binary, but the instrument(s) need not be; a valid continuous instrument can be broken up into a series of binary indicators and remain valid. While we refer to the judge fixed effects design as the leading special case, this article applies to any design in this broader instrumental variables treatment effecs framework.

The intuition and transparency of the judge fixed effects design notwithstanding, identification relies on assumptions that may be controversial in some settings and at the very least deserve scrutiny. The design's popularity and the policy relevance of the research questions it has been employed to answer further underscore the importance of testing the identifying assumptions.

In this paper we develop a test of the identifying assumptions underlying the judge fixed effects design. Most commonly, researchers using this design adopt the local average treatment effects (LATE) framework introduced by Imbens and Angrist (1994), which allows for heterogeneous treatment effects. The key assumptions are that individuals are randomly assigned to judges, an exclusion restriction whereby judges exert no influence on outcomes other than through their decision whether to "treat" an individual, and that judge assignment has a weakly monotonic effect on each individual's treatment status. In the context of Kling's (2006) original study, this latter monotonicity condition implies that if judge A is more likely to incarcerate offenders than judge B, every individual incarcerated by judge B would also have been incarcerated by judge A had judge A handled the case. A variation on the LATE framework relaxes monotonicity slightly, provided the average treatment effect among individuals who violate monotonicity is identical to the average treatment effect among some subset of individuals who satisfy it (de Chaisemartin, 2017).

Our test follows from the implication that outcomes averaged at the judge level will be a continuous function with bounded slope of the judge-level treatment probability (or "propensity"). The slope of the function at any given point corresponds with the marginal treatment effect, so that bounds on the slope arise from bounds on the support of the outcome variable. We develop a two-part test of this implication that takes into account that judge propensity is estimated and demonstrate that our test has substantial statistical power in empirically reasonable settings. One part of the test can be characterized as a generalization of the traditional Sargan-Hansen overidentification test (Sargan, 1958; Hansen, 1982), which probes the stricter condition that average outcomes given the instrument are a *linear* function of the treatment propensity. The other part of the test assesses whether implied treatment effects are possible given the support of the outcome. Our test remains feasible even when additional covariates are not available, a situation that would preclude the use of traditional balance tests of randomization and existing tests examining how judge severity varies across subgroups.

Our procedure jointly tests exclusion and monotonicity, and therefore a rejection suggests one or both may fail. When a priori institutional considerations support the plausibility of the exclusion restriction (implying that a rejection is evidence against the monotonicity condition), we propose a weaker *average monotonicity* condition under which the instrumental variables estimator converges to a proper weighted average of individual treatment effects. If, instead, a priori information suggests the exclusion restriction fails, researchers may consider falling back on the relaxation of the exclusion restriction described in Kolesár et al. (2015) to interpret estimates as consistent estimates of treatment effects.

We apply our test in an examination of the effect of pretrial release on the probability of conviction. We do so using exogenous variation in judge assignment for offenders in Miami-Dade county, a setting used in Dobbie et al. (2018a). We reject the null hypothesis that the exclusion restriction and monotonicity assumption hold. Interpreting the estimates as causal effects of pretrial detention therefore requires alternative assumptions such as the weaker average monotonicity assumption we propose and/or the relaxation of the exclusion restriction described in Kolesár et al. (2015).

Our paper builds on previous work related to tests of instrument validity. As mentioned above, our test can be seen as a generalization of the traditional Sargan (1958) overidentification test to a setting with heterogeneous effects. Angrist and Imbens (1995) show that the monotonicity condition can be tested in a setting with a treatment that takes on more than two values. Our test complements their result by allowing testing with a binary treatment as well, an important case in practice. Kitagawa (2015), Huber and Mellace (2015), Mourifié and Wan (2017), and Norris et al. (2018) also proposed tests for instrument validity in settings with a binary treatment similar to ours. These tests require a priori knowledge of the instruments' order with respect to the probability of treatment, knowledge not commonly available in empirical settings. Applying such tests as if the instrument order were known can lead to substantial overrejection, as we show in simulations below. Most common in the literature are informal tests that examine the correlation of judge severity across observable subgroups (Bhuller et al., 2018; Dobbie et al., 2018a). This approach cannot detect violations of the exclusion restriction, and is only a weak test of monotonicity; strict monotonicity requires not only that subgroup-specific propensities across judges be positively correlated with the overall propensity, but that they are monotonically increasing with the each judge's overall propensity. This approach does, however, test an implication of the weaker average monotonicity assumption we propose for when our test rejects, and thus complements our approach well.

2 Econometric Framework

Consider a binary treatment indicator, D_i for individual i, such as pre-trial release or placement in foster care, whose possible effects on outcome Y_i are of interest. Denote the potential outcome realized by individual i if untreated by $Y_i(0)$, and if treated by $Y_i(1)$. Individual i's treatment effect is therefore $\delta_i := Y_i(1) - Y_i(0)$. A class of parameters of interest consists of weighted averages of the treatment effect:

$$\alpha_w = \frac{E\left[w_i \delta_i\right]}{E\left[w_i\right]}, \tag{1}$$

for nonnegative weights w_i. Treatment status is influenced by the assignment to a judge, denoted by $J_i \in \{0, \ldots, J\}$. An individual's potential treatment status as a function of the judge assignment is given by $D_i(j) \in \{0, 1\}$. Define each judge's propensity to assign treatment as $p(j) := E\left[D_i(j)\right]$. Observed variables include the judge assignment J_i, treatment status $D_i := D_i(J_i)$, and outcome $Y_i := Y_i(D_i)$.

Treatment effect parameters of the form (1) can be consistently estimated under the local average treatment effects (LATE) assumptions (Imbens and Angrist, 1994). This framework makes the following assumptions:

Condition 1 (LATE instrumental variables validity) *For all $j \in \{0, \ldots, J\}$, the following hold jointly:*

a. Random assignment and exclusion: the triple $(Y_i(0), Y_i(1), D_i(j))$ is jointly independent of J_i;

b. Nontrivial instrument: the propensity $p(j)$ is a nontrivial function of j;

c. Monotonicity: For all $j, w \in \{0, \ldots, J\}$, either $D_i(j) \geq D_i(w)$ for all i, or $D_i(j) \leq D_i(w)$ for all i.

Conditions 1a and 1b will be satisfied if judge assignment is random, and judges vary in their influence on treatment status, but do not otherwise affect outcomes. Given part a, the second part can be verified, since under the first part of the condition $p(j) = E\left[D_i | J_i = j\right]$. Part a, however, cannot be directly verified. Part c, monotonicity, means that any individual who is treated when assigned to a particular judge would also be treated if assigned to a judge of equal or greater propensity. It implies that individuals can be partitioned into a group that never receives treatment regardless of judge assignment (never-takers), a group that always receives treatment regardless of judge assignment (always-takers), and groups corresponding to each propensity value p who are treated when assigned to a judge with $p(j) \geq p$ and not otherwise. We refer to members of these latter groups as p-compliers. Imbens and Angrist (1994) show that under Condition (1), weighted average treatment effects of the form (1) are identified, including the average treatment effect among each complier group, and an overall complier average treatment effect that puts positive weights on each of the complier groups.

The LATE Condition 1 implies that expected outcomes conditional on judge assignment lie on a continuous funtion of the judge's propensity, $p(J_i)$, and that the slope of the function at some propensity value p is equal to the average treatment effect among p-compliers. Any bounds on the magnitude of possible treatment effects are also bounds on the slope of the conditional expectation function of Y_i given $p(J_i)$. At a minimum, treatment effects are bounded by the outcome variable's support. The following theorem formalizes this implication:

Theorem 2 *Suppose Condition 1 holds and Y_i has compact support. Then there exists $K < \infty$ such that $E[Y_i|J_i = j] = \phi(p(j))$ where $\phi \in \mathrm{Lip}_K([0,1])$.*

Proof. All proofs are in the Appendix. ■

The result of the theorem means the judge fixed effects identifying assumptions have testable implications. The requirement that the outcome have compact support does not limit its usefulness: one can always replace the outcome by a set of indicator variables of the form $1(Y_i \leq y)$, and doing so leads to tests that exploit the whole outcome distribution. We focus initially on the conditional expectation of Y_i itself, and explore extensions to distributions and quantiles in an appendix.

Common alternatives to the LATE assumptions also share this implication. The traditional instrumental variables assumptions do not assume monotonicity, but do require constant treatment effects. In this framework the result of Theorem 2 is even stronger: $E\left[Y_i|J_i = j\right]$ is not only continuous, but a *linear* function of propensities. de Chaisemartin (2017) presents a weaker monotonicity assumption under which a proper weighted average of treatment effects is still identified. His "compliers-defiers assumption" is that, for each pair of judges, there is a subset of compliers that (1) is the same size as the set of defiers for that judge pair, and (2) has the same average treatment effect as the set of defiers (those whose treatment status response violates monotonicity) for that judge pair. Replacing traditional monotonicity with this compliers-defiers assumption leaves the result of Theorem 2 unchanged. Thus, the test described below applies equally well when these alternative assumptions are invoked.

3 Testing Procedure

Our proposed test is based on two observations that follow from Theorem 2: (1) average outcomes conditional on judge assignment should fit a continuous function of judge propensities; and (2) the slope of that continuous function should be bounded in magnitude by the width of the outcome variable's support. The tests consists of examining whether observed outcomes averaged by judge are consistent with such a function.

Figure 1 illustrates graphically the intution behind the test. The top panel depicts a situation in which the assumptions are satisfied, so that average outcomes by judge lie on a continuous function of judge propensity, and that slope of that function is within the required bounds. The bottom panel illustrates two ways that violations of the assumptions may appear. In the first (labeled “A” on the figure), two judges have identical propensities, but different average outcomes; thus no continuous function can pass through both points. In the second (labeled “B”), two adjacent judges do not have identical propensities, but their average outcomes are sufficiently different that the slope of the curve connecting them exceeds the possible treatment effect values.

This suggests a conceptually straightforward procedure for testing the judge fixed effects design’s assumptions:

1. Regress the outcome Y_i on a flexible function of the judge propensity, $\phi\left(p\left(J_i\right)\right)$
2. Jointly test fit and slope by
   (a) Regressing the residuals from step 1, $u_i = Y_i - \phi\left(p\left(J_i\right)\right)$, on judge indicators and testing whether the coefficients are jointly zero;
   (b) Testing whether the slopes of the function are within the bounds dictated by the support of Y_i.

This procedure presents two complications. The first is specifying the propensity regression in step 1. The step 1 regression of the outcome on the judge propensities should be as flexible as the researcher’s assumptions regarding treatment effect heterogeneity. A linear regression imposes constant treatment effects and makes the test procedure above equivalent to the usual Sargan-Hansen overidentification test (Sargan, 1958; Hansen, 1982). To impose minimal assumptions on treatment effect heterogeneity, Theorem 2 suggests one should choose a flexible specification that approximates Lipschitz functions well, such as polynomials or splines (Chen, 2007). Our simulations and application use b-splines (see Racine, 2018), but other bases could be used as well. Let the number of terms in the chosen series be $m+1$, and let the function class in which the chosen specification lies be denoted $\mathcal{S}_m$; for example, degree-m polynomials or degree-r splines with $m-r$ knots. In the context of the judges design, the number of terms in the approximating series is limited by the number of judges; settings with a large number of judges, such as our application, allow the specification to be quite flexible.

The second complication is accounting for the estimation of the judge propensities and the step 1 residuals when performing the tests in step 2. The simplest estimator for $p\left(J_i\right)$ is simply the fitted value from a regression of treatment status D_i on a vector of judge indicators $W_i = \left(1, 1\left(J_i = 1\right), \ldots, 1\left(J_i = J\right)\right)'$, which amounts to the fraction treated among individuals assigned to judge j, although it may be generalized by adding controls to the first stage regression. Denote the estimated fitted values $\hat{P}_i$. The first-step residuals also depend on a linear regression coefficient: collecting the terms of the spline (or whichever basis is chosen) in the estimated propensity into the vector $\hat{S}_i$, the estimated residual for the i-th observation is

$$\hat{u}_i = Y_i - \hat{S}_i' \left(\sum_{i=1}^{n} \hat{S}_i \hat{S}_i'\right)^{-1} \sum_{i=1}^{n} \hat{S}_i Y_i.$$

The fit component of our test is based on the second-step coefficients obtained by regressing $\hat{u}_i$ on W_i:

$$\hat{\gamma} = \left(\sum_{i=1}^{n} W_i W_i'\right)^{-1} \sum_{i=1}^{n} W_i \hat{u}_i.$$

Under the conditions of Theorem 2, $\hat{\gamma}$ converges in probability to zero. Our procedure tests this via the following Wald statistic:

$$\hat{T} = n \hat{\gamma}' \hat{\Omega}^{-1} \hat{\gamma}, \tag{2}$$

where $\hat{\Omega}$ is a consistent estimator of the limiting covariance of $\sqrt{n}\hat{\gamma}$, accounting for the first-step estimates $\left(\sum_{i=1}^{n}\hat{S}_i\hat{S}_i'\right)^{-1}\sum_{i=1}^{n}\hat{S}_iY_i$ and $\hat{P}_i$. Given an iid sample, a suitable estimator is:

$$\hat{\Omega} = \left(n^{-1}\sum_{i=1}^{n} W_iW_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\left(W_i\hat{u}_i - \hat{R}_i\right)\left(W_i\hat{u}_i - \hat{R}_i\right)'\right)\left(n^{-1}\sum_{i=1}^{n} W_iW_i'\right)^{-1},$$

where $\hat{R}_i$ is and adjustment term defined in the appendix.

Given the assumptions so far, the test statistic (2) converges in distribution to a chi-squared random variable with degrees of freedom equal to the difference between the number of judges and the number of terms in the specification for ϕ, as the following theorem formalizes:

Theorem 3 *Suppose Condition 1 holds and $\phi \in \mathcal{S}_m$, where $m < J$. Suppose further that $\{Y_i, D_i, J_i\}_{i=1}^{n}$ comprise an iid sample and $E\left[|Y_i|^3\right] < \infty$. Then*

$$\hat{T} \underset{d}{\rightarrow} \chi^2\left(J - m\right).$$

Performing the fit component of the test means computing the test statistic and obtaining the associated p-value from the appropriate chi-squared distribution.

The slope component of the test examines whether the slopes of the function relating outcomes to judge propensities lie between $-K$ and K, recalling that K is the width of the outcome variable's support. The function relating average outcomes given judge assignment to judge propensities is specified as

$$\phi\left(p\right) = \delta_0 S_0\left(p\right) + \cdots + \delta_m S_m\left(p\right),$$

where $S_0, \ldots, S_m$ are elements of a polynomial series, spline series, or whichever basis is chosen for ϕ. When ϕ is specified as a quadratic b-spline, the maximum slope occurs at one of the knots, $\{t_0 = 0, t_1, \ldots, t_{m-2}, t_{m-1} = 1\}$. The slope at the l-th knot is given by

$$\phi'(t_l) = \frac{2}{t_{l+1} - t_{l-1}} (\delta_{l+1} - \delta_l), \quad l = 0, \ldots, m-1,$$

where we define $t_{-1} = t_0 = 0$ and $t_{m-1} = t_m = 1$. The restriction on the slope of ϕ corresponds to the following set of inequality constraints:

$$\left\{ -K \leq \frac{2}{t_{l+1} - t_{l-1}} (\delta_{l+1} - \delta_l) \leq K \right\}_{l=0}^{m-1}.$$

Given estimates $\hat{\delta} = \left(\sum_{i=1}^n \hat{S}_i \hat{S}_i' \right)^{-1} \sum_{i=1}^n \hat{S}_i Y_i$ and corresponding variance matrix that accounts for the estimation of $\hat{P}_i$, we implement the moment inequality testing procedure proposed by Andrews and Soares (2010). This procedure first performs generalized moment selection to eliminate inequalities that are far from binding, and then constructs a modified method of moments (MMM) test statistic to test the remaining inequalities. The appendix describes the details of the implementation.

Finally, we combine the fit component and slope component of the test via a weighted Bonferroni procedure to produce a single joint test. If we denote the p-value from the fit component of the test as p_f and the p-value from the slope-component of the test as p_s, then a joint level-α test rejects if either $p_f < \omega\alpha$ or $p_s < (1-\omega)\alpha$, for some weight $\omega \in [0,1]$. Equivalently, one can define a joint p-value as $\min\{p_f/\omega, p_s/(1-\omega)\}$ and reject if the joint p-value is less than α. The choice of ω governs the direction of power between the fit component and the slope component, with values near one directing more power to the fit component of the test. In the "just identified" case when there are only two judges, the fit component of the test will have no power, and therefore choosing $\omega = 0$ is appropriate. As the number of judges grows, the specification of ϕ becomes more flexible and the number of inequalities being tested in the slope component grows, causing the slope component of the test to lose power; choosing ω to be near one will be most appropriate in cases with many judges, as in our application. The procedure described above has asympotic size of at most α; the simulations below show the test also performs well in finite samples.

The test has power against alternatives in which the conditional expectation of Y_i given the assigned judge differs from a continuous function of the judge propensity, or in which the function has slopes that exceed the maximum possible treatment effect size. This includes violations of random assignment or exclusion (Condition 1a) or violations of monotonicity (Condition 1c). The test has power against violations of random assignment when, for example, a given judge, $\tilde{j}$, is more likely than other judges to be assigned to a certain group of defendants, and that group differs in its potential outcomes from other defendants. Similarly, the test has power against violations of the exclusion restriction that arise when judges affect outcomes through channels besides the treatment, so that defendants assigned to one judge $\tilde{j}$ experience different average outcomes than defendants assigned to another judge with a similar propensity.

Finally, the test has power against violations of monotonicity when, for example, a given judge $\tilde{j}$ with similar propensity to another judge nevertheless treats a quite different set of defendants, and treatment affects those defendants differently from the average. Under the preceding types of violations, average outcomes conditional on $J_i = \tilde{j}$ will differ discretely from average outcomes conditional on other judges no matter how close their propensity. As a result, the conditional expectation function will not be well approximated by a continuous function with bounded slope, the coefficients from regressing residuals on judge indicators will have nonzero probability limits, or the slopes will exceed the maximum treatment effects size and the test will have asymptotic power.

The test does not have power against all violations of Condition 1, however, as is also the case for other specification tests in the literature (Kitagawa, 2015; Huber and Mellace, 2015; Mourifié and Wan, 2017). For example, the test will not detect violations of the monotonicity condition where two judges with similar propensity nevertheless treat different types of defendants, but those types have identical average treatment effects. (Monotonicity violations are unimportant in this case anyway, because they do not induce bias unless treatment effects are heterogeneous.) Similarly, the test will not detect violations of the exclusion restriction where two judges with similar propensity values also have exclusion violations in similar magnitude and direction. The test will not have power against knife-edge alternatives like this, but will nevertheless have power against a wide class of alternatives as described in the previous paragraph.

4 Implications if the test rejects

A test rejection constitutes evidence that either the exclusion restriction or the monotonicity assumption (or both) fail to hold. Intuitively, it implies either that judges influence outcomes beyond their propensity to assign treatment, or judges disagree on their implicit ordering of which defendants should be treated. Strictly speaking, the test does not distinguish which assumption is violated, although a priori information may be informative about the nature of the violation. But regardless of which assumption is violated, the consequence is that instrumental variables estimators using the full set of judge indicators cannot be guaranteed to consistently estimate causal parameters such as (1) except under special circumstances, described below.

4.1 Exclusion restriction violations

If exclusion restriction violations cannot be ruled out, researchers may consider assumptions that relax the exclusion restriction but still allow for identification of causal parameters. One such assumption is the "many invalid instruments" condition proposed by Kolesár et al. (2015), who allow the exclusion restriction to be violated, and instead assume that the direct effects of instruments on the outcome are uncorrelated with their effects on treatment.[1] In the current setting, this means that a judge's direct effect on defendant outcomes is uncorrelated with his or her propensity to assign treatment. Whether this assumption is plausible will depend on the specific setting.

Mueller-Smith (2015) proposes a more traditional strategy for dealing with exclusion restriction violations when the channels of the violations are observed, namely, to treat the channels through which judges affect outcomes besides treatment as additional endogenous variables. Under the traditional linear simultaneous equations framework, which includes constant treatment effects and that the number of judges is greater than the number of channels, the treatment effect of interest is identified.

4.2 Monotonicity violations

If a priori considerations rule out exclusion restriction violations, then a rejection provides evidence against the strict monotonicity assumption, Condition 1c. A natural next step by researchers would be to relax monotonicity, perhaps by adopting the weaker compliers-defiers condition (de Chaisemartin, 2017). Unfortunately, the implication our procedure tests is also implied by the compliers-defiers condition, meaning a rejection constitutes evidence against that assumption as well.

[1] The proof that causal effects can be identified under this assumption uses a model with constant treatment effects. A similar result may exist for a heterogeneous effects setting, but has not been demonstrated.

Strict monotonicity can be relaxed in another way, however, that still preserves the interpretation of the IV estimand as a proper weighted average of individual treatment effects of form (1), and may hold even when strict monotonicity's testable implication is not satisfied. Define $\bar{D}_i = \sum_{j=1}^{J} \lambda_j D_i(j)$ as individual i's average treatment status across judges, where λ_j is the probability of being assigned $J_i = j$, and consider the following *average monotonicity* condition:

Condition 4 (Average monotonicity) $\omega_i := \sum_{j=1}^{J} \lambda_j (p(j) - p) \left(D_i(j) - \bar{D}_i\right) \geq 0$ *almost surely.*

Condition 4 means that for each individual, the covariance between the individual's judge-specific treatment status and judge overall treatment propensities is weakly positive. This means that individuals may violate monotonicity with specific judges, as long as they comply with monotonicity for enough other judges so that the overall covariance stays nonnegative.

Under this weaker notion of average monotonicity (and the exclusion restriction) the IV two-stage least squares estimand can be interpreted as a proper weighted average, as the following theorem shows.

Theorem 5 (IV weighted average) *Suppose Condition 1a and 1b hold and define*

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

Then

$$\beta_{2SLS} = \frac{E[\omega_i \delta_i]}{E[\omega_i]},$$

where ω_i is defined in Condition 4.

Thus, falling back on Condition 4 when the test we propose rejects provides a way to nevertheless interpret estimates causally.

But can the weaker average monotonicity condition be verified? Although the individual weights ω_i are not identified, as they are a function of potential treatment status, their conditional expectation is, which provides a testable implication of Condition 4. Given random assignment and the exclusion restriction, the conditional expectation of ω_i given some covariate X_i conditional upon which J_i is independent of $\left(Y_i\left(0\right), Y_i\left(1\right), \left\{D_i\left(j\right)\right\}\right)$ is given by the covariance between judges' x-specific treatment propensity and judges' overall propensity:

$$E\left[\omega_i | X_i = x\right] = \sum_{j=1}^{J} \lambda_j \left(p\left(j\right) - p\right)\left(p_x\left(j\right) - p_x\right),$$

where $p_x = E\left[D_i | X_i = x\right]$ and $p_x\left(j\right) = E\left[D_i | J_i = j, X_i = x\right]$. The assumption that $\omega_i \geq 0$ implies $E\left[\omega_i | X_i = x\right] \geq 0$ for all x in the support of X_i, which may be tested by examining the observed covariance between judges' group-specific treatment propensities and overall propensity. This gives a formal motivation to the informal tests that overall propensity is positively correlated with group-specific propensities in the applied literature (Bhuller et al., 2018; Dobbie et al., 2018a).

If the many-invalid-instruments assumption and our average monotonicity assumption seem reasonable in a given application, then using judge assignment for identification is still appropriate even if our test rejects the conventional exclusion and monotonicity conditions. However, these weaker assumptions do constrain the scope of inference. In particular, they do not allow for the identification of marginal effects along the entire distribution of judge propensities, as can be achieved in the conventional framework (Mogstad et al., 2017). The weaker assumptions rely on averaging across the entire set of judges, while identification of marginal effects throughout the distribution requires assumptions to hold judge by judge.

5 Simulations

The section illustrates how the proposed test's performance in terms of finite-sample size and power depends on features of the underlying data. The simulations' data generating process mimics a setting with J judges, to whom individuals are assigned with uniform probablity:

$$J_i \sim U\left\{1, \ldots, J\right\}.$$

A judge's propensity to assign treatment is given by:

$$p\left(J_i\right) = \theta J_i / \left(J\right).$$

The outcome is generated as

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i,$$

and treatment is determined by

$$D_i = 1\left(\Phi\left(-\nu_i\right) \leq p\left(J_i\right)\right),$$

where

$$v_i = \rho\varepsilon_i + \sqrt{1-\rho^2}\eta_i$$

and

$$\left(\varepsilon_i, \eta_i\right) \sim N\left(0, I_2\right).$$

In this setup the parameter θ governs the strength of the instruments and ρ determines the degree of treatment endogeneity. Note that this setup satisfies Condition 1. The main simulation results set $\omega = 1$, which directs power to the fit component of the test. Further simulation results below show how the test performs under different choices for ω.

The first set of simulations examines how the test's size depends on the number of observations. We set the simulation parameters as $J = 10, \theta = 1, \beta_0 = \beta_1 = 1$, and $\rho = .5$. We consider sample sizes of $n \in \{500; 1,000; 2,000; 5,000; 10,000\}$, and for each sample size draw 999 samples from the data generating process described above and perform the test with nominal size $\alpha = .05$, recording the rejection rate for each sample size. The simulations show that the test has very close to nominal size even for modest sample sizes. Figure 2 plots the rejection rate as a function of the sample size. The horizontal line is at .05. The simulated rejection rate is very near the nominal level throughout the range of sample sizes.

The next set of simulations explores the test's power to detect a violation of the exclusion restriction, Condition 1a, which can arise if judges have direct effects on outcomes other than through treatment. The data generating process is as described above, except judges now have direct effects on the outcomes:

$$Y_i = \beta_0 + \beta_1 D_i + \sum_{j=1}^{J} \gamma_j 1\left(J_i = j\right) + \varepsilon_i,$$

where the individual judge effects γ_j are drawn from a normal distribution with mean zero and standard deviation that varies from zero (corresponding to no violation) to 1 (severe violation). We set $n = 1000$ for this set of simulations. The simulations show that the proposed test's power increases rapidly with the severity of the violation. Figure 3 plots the rejection rate as a function of the standard deviation of the direct judge effects. At the far left the rejection rate is very near .05, reproducing the result that the test has correct size when the assumptions are satisfied. As the standard deviation of the direct judge effects grows, the rejection rate increases rapidly. Power exceeds 90 percent when the standard deviation is 0.2 and is essentially 100 percent for standard deviations above 0.3.

The next set of simulations illustrates the test's power to detect violations of the monotonicity assumption, Condition 1c, which can occur if judges do not implicitly agree on the order in which defendants should be treated. To allow for monotonicity violations in the simulations, we introduce heterogeneity in defendants and judges. We introduce an additional set of J judges (indexed $J+1,\ldots,J$) who order most defendants identically to the first J judges, but order a fraction $\phi < .5$ of defendants, whom we call defiers, in the opposite order. Since violations of monotonicity only lead to bias when treatment effects vary, we set defiers' treatment effect to $-\beta_1$. Let the binary variable F_i with $\Pr\left(F_i = 1\right) = \phi$ indicate whether a defendant is a defier. Treatment assignment and outcomes are then determined as follows:

$$\begin{aligned} Y_i &= \begin{cases} \beta_0 - \beta_1 D_i + \varepsilon_i & , \quad F_i = 1 \\ \beta_0 + \beta_1 D_i + \varepsilon_i & , \quad \text{otherwise} \end{cases}, \\ D_i &= \begin{cases} 1\left(\Phi\left(-\nu_i\right) \leq 1 - p\left(J_i\right)\right) & , \quad F_i = 1 \text{ and } J_i \geq J+1 \\ 1\left(\Phi\left(-\nu_i\right) \leq p\left(J_i\right)\right) & , \quad \text{otherwise} \end{cases}. \end{aligned}$$

The simulations show the proposed test has good power to detect violations of monotonicity of this sort. Figure 4 plots the test's rejection rate as a function of the fraction of defiers ϕ. At the far left ($\phi = 0$, corresponding to no violation) the test rejects at a rate near $\alpha = .05$ as expected. As the fraction ϕ increases and the violation of monotonicity becomes more severe, the rejection rate increases rapidly. Power exceeds 80 percent when the fraction of defiers is greater than .3.

We also run simulations that allow us to compare how our test performs relative to the the test described in Kitagawa (2015). The Kitagawa test assumes a priori knowledge of the instruments' order with respect to the probability of treatment. In the context of judge assignment, this assumption is problematic because the judge propensities to treat are estimated rather than directly observed. To assess the size of the Kitagawa test, we run simulations with four judges (the test quickly becomes computationally burdensome as the number of judges increases) with population propensities $.25, .495, .5$, and $.505$ using three different samples sizes ($n \in \{5,000; 10,000; 100,000\}$). The monotonicity assumption in this scenario means individuals fall into one of five compliance categories: always-takers, never-takers, and one of three complier groups. We set the treatment effect equal to zero, and set always-takers' outcomes to $Y_i = 0$, judge 1 compliers' outcomes to $Y_i = 1$, judge 2 compliers' outcomes to $Y_i = 2$, judge 3 compliers' outcomes to $Y_i = 3$, and never-takers' outcomes to $Y_i = 4$. With a nominal test size of 5%, rejection rates for the Kitagawa test are 10.3% for a sample size of 5,000, and grow to 22.2% for $n = 10,000$ and 27.9% with a sample size of 100,000. In comparison, rejection rates for our test in the same simulations are 5.2%, 5.1% and 4.7%. The results show that the Kitagawa test substantially overrejects and that the distortion does not decrease with the sample size over the range considered. For a large enough sample size, of course, and given a dgp, estimation error in the propensities will become negligible and the Kitagawa test will have correct size. But for any given sample size, there is a dgp for which the Kitagawa test will fail to control size; that is, the Kitagawa test is not uniformly asymptotically valid when propensities are estimated, as illustrated in our simulations.

Finally, we show how the test performs under different choices for ω. To demonstrate this, we run two sets of simulations: one with a small number of judges ($J = 2$), in which we will see that small values for ω are best, and another with a large number of judges ($J = 20$) in which larger values for ω are best. The outcome variable is binary, as in our application below, with expected value conditional on judge assignment given by

$$\Pr\left(Y_i = 1 | J_i\right) = \beta_0 + \beta_1 J_i/k + \sum_{j=0}^{J} \gamma_j 1\left(J_i = j\right),$$

and treatment propensity given by

$$\Pr\left(D_i = 1 | J_i\right) = \alpha_0 + \alpha_1 J_i/k.$$

As above, the γ_j terms represent violations of the exclusion restriction when they are nonzero; in this set of simulations they are drawn from normal distribution with standard deviation .2. This simulation setup also allows the assumptions to be violated when $\beta_1 > \alpha_1$, as this would imply an average treatment effect greater than one, which is impossible for a binary outcome. This corresponds to a violation of the slope condition. In this simulation setup we set $\beta_1 = .3$ and $\alpha_1 = .2$. Using a simulated sample size of $n = 1,000$, we perform our test for several choices of ω between zero and one, and examine how the test's power depends on ω in the few-judge case ($J = 2$) and the many-judge case ($J = 20$). The simulation results show that in the few-judge case, power is greatest when $\omega = 0$, since the fit component of the test has no power in this case. The upper panel of Figure 5 shows that the test's power is over 80 percent when $\omega = 0$, and drops to zero when $\omega = 1$. The situation is reversed in the many-judge case. The lower panel of Figure 5 shows that power is very poor (around 10 percent) when $\omega = 0$, and increases for higher values of ω. Typical instances of the judge fixed effects design, including our application below, involve relatively many judges. These simulation results suggest choosing ω to be high in these cases. In the application we set $\omega = 1$.

6 Empirical Application: Pretrial Detention and Case Outcomes

To illustrate how to implement and interpret our test in practice, we replicate results from the literature on the impact of pretrial detention on case outcomes. When an individual is charged with a crime, he may be released or held in jail while the case is being adjudicated. Several recent papers on pretrial detention (Gupta et al., 2016; Leslie and Pope, 2017; Dobbie et al., 2018a) find that holding defendants pretrial increases the probability that they will be convicted. In most settings, the only purpose of the first hearing for the majority of defendants is to determine their pretrial status. The judge at this hearing decides whether and how high to set bail. Causal identification of the impact of detention on conviction comes from variation across arraignment judges in the rates at which they detain people.

6.1 Background

We analyze a dataset from criminal cases in Miami-Dade that is identical to the data from this location used in Dobbie et al. (2018a). Following arrest in Miami-Dade, defendants are brought to a police station where they can secure their release by posting bail according to a schedule based on seriousness of offense. For the 70% of defendants who do not post bail immediately, there is a bail hearing within 24 hours of arrest. At the hearing, the bail judge on duty may change the bail amount or impose nonmonetary conditions, like monitoring.

The weekday bail hearing shifts are presided over by a single judge, while approximately 60 judges preside at weekend bail hearings based on a rotating schedule. Defendants are automatically assigned to the bail judge on duty, leaving little scope for manipulating judge assignment given the short window between arrest and the hearing. The process of trial judge assignment is not connected to the bail hearing, so the leniency of the bail judge does not have a systematic relationship with the tendencies of judges involved in later stages of the case.

6.2 Data

Our data include all criminal cases on record in Miami-Dade between 2006 and 2014. The court data include arrest charge, filing charge, and disposition charge; the outcome for each charge; and the punishment for each guilty charge. We observe the bail judge, bail amount and type, and if/when bail was posted. We know each defendant's name, gender, race, date of birth, and address. The individually identifying information allows us to construct criminal history and future criminal activity during the observation period. We restrict attention to cases in which there was a weekend bail hearing; these are the cases assigned a bail judge based on a rotating schedule. Table 1 shows summary statistics for our sample, which is identical to the Miami-Dade sample used in Dobbie et al. (2018a). The sample is predominantly male, and split fairly evenly between whites and blacks. Defendants who are released pretrial are less likely to have a prior offense from the past year and more likely to be white than those detained pretrial. They are also less likely to be convicted, be sentenced to incarceration, or to recidivate after case disposition. These differences are consistent with the hypothesis that pretrial detention influences case outcomes. They also highlight the nonrandom assignment of pretrial status, and the need to go beyond simple comparisons of means to identify a causal effect.

6.3 Research Design

We are interested in the relationship between pretrial release and conviction represented in the following equation:

$$convicted_{ict} = \beta_0 + \beta_1 released_{ic} + \beta_2 X_{ict} + \epsilon_{ict}$$

where $convicted_{ic}$ is an indicator for whether individual i in case c was convicted, $released_{ic}$ is an indicator for whether the individual was released within three days of the bail hearing, and X_{ic} is a vector of defendant and case characteristics. OLS will yield biased estimates if there are unobserved factors that are correlated with both pretrial release and case outcomes. For example, defendants with above average lawyers may have better outcomes at the bail hearing stage and a lower probability of conviction.

To estimate the causal effect of pretrial status on case outcomes, we begin by instrumenting for whether an individual was released pretrial with a measure of the leniency of his bail hearing judge. Following Dobbie et al., we use a leave-out mean of residualized pretrial outcomes as an instrumental variable. We first regress actual pretrial release status on a vector of bail year by bail day of week and court by bail month by bail day of week fixed effects, to account for the possibility of bail judges who are more likely to work certain days or months. The variation we use for identification is therefore differences in leniency between judges working the same day of the week within a month and year. Let the residual from this regression be $Released^*_{ict}$. The instrumental variable for each defendant is the average of his bail judge's residuals for that year, excluding his own:

$$Z_{ict} = \left(\frac{1}{n_{tj} - n_{itj}}\right) \left(\Sigma_{k=0}^{n_{tj}}(Released^*_{ikt} - \Sigma_{c=0}^{n_{itj}} Released^*_{ict}\right)$$

where n_{tj} is the number of cases seen by judge j in year t and n_{itj} is the number of cases of defendant i seen by judge j in year t.

6.4 Identification and Results

For our instrumental variable strategy to identify the causal effect of pretrial release, the exclusion restriction and a monotonicity condition must hold. The exclusion restriction in this setting requires both random assignment of judges and that judges impact the outcome of interest only through pretrial detention. We provide evidence that judge assignment is conditionally random in Table 2. Both the coefficients and standard errors, and the p-values on the tests of joint significance indicate that defendant characteristics are powerful predictors of pretrial status, but not of judge leniency.

Even if defendants are randomly assigned to judges, the exclusion restriction could be violated if bail judges influence case outcomes through channels besides pretrial release. For example, a bail judge who orders pretrial drug testing or treatment influences outcomes other than through pre-trial detention and violates the exclusion restriction.

Monotonicity could be violated in this context if, for example, some judge were harsh on average, but lenient toward female defendants. One approach employed in the existing literature to assess the monotonicity assumption is to check whether judge leniency within one subgroup is positively correlated with judge leniency within another subgroup. Another test in the same spirit is to examine whether the first stage estimates of the relationship between the residualized measure of judge leniency Z_{ict} and individual pretrial status are positive for all subgroups, meaning that judges who are more lenient overall are more likely to release members of any observable subgroup. As discussed above, this approach tests the weaker average monotonicity condition, but is likely not a powerful test of strict monotonicity.

Our joint test of the exclusion restriction and monotonicity assumption probes whether there is a continuous relationship between judge leniency and case outcomes (averaged at the judge level), taking into account that judge leniency is estimated. This can be assessed graphically. Figure 6 plots average case outcomes by release rates for each judge. Intuitively, our test assesses whether the data are consistent with all points lying on a single continuous curve, to within sampling variation. Visually, this appears to be unlikely.

The formal test confirms the visual evidence. Table 3 shows results when we apply our test to the data. We implement our test choosing $\omega = 1$, since with a large number of judges, the slope component of the test has little power. The conclusions are unchanged for a wide range of choices for ω, however. The top panel shows that we reject the null hypothesis on the full sample for various numbers of knots in the spline function.[2]

As discussed above, one possibility is to rely instead on the weaker average monotonicity assumption, provided that the rejection is not due to exclusion restriction violations. To assess the plausibility of our average monotonicity assumption, we check that the first-stage coefficient is positive within subsamples defined by case and defendant characteristics (see Tables 4 and 5), an approach used in several papers in the literature (Gupta et al., 2016; Leslie and Pope, 2017; Dobbie et al., 2018a). The results confirm that the first stage coefficient is positive and statistically significant within several important observable subsamples. Based on these estimates, the

[2] If we suspected that monotonicity violations along certain observable characteristics were to blame, then we could test jointly across these dimensions. Assuming independence of the subsamples, we could add up the chi-squared test statistics and degrees of freedom after running the test on all subsamples defined by the relevant observables to get the joint test statistic and its chi-squared degrees of freedom.

average monotonicity assumption may be justified, and IV estimates have a causal interpretation despite the violation of strict monotonicity, again, provided the rejection was not due to exclusion restriction violations.

To the extent that there are also exclusion restriction violations, we must invoke the many invalid instruments assumption to interpret our IV estimates causally. If we observe any potential non-focal treatment dimensions in the data, we can use them to probe the many invalid instruments assumption. In our setting, we observe whether or not the defendant was represented by a public defender. The bail judge has no control over the specific public defender (and prosecutor) involved in the case, but the court does have the final say on eligibility for representation by a public defender. We construct the propensity for defendants appearing before each judge to receive a public defender using the same approach as we did for the construction of the focal propensity. The first stage for this non-focal propensity is just as strong as the focal first stage, suggesting that this is, in fact, a channel through which judges exercise real influence. Finally, we check whether there is a statistically significant correlation between the focal and non-focal propensities, and find a correlation coefficient of -0.18 with a p-value smaller than .0001. The strength of this relationship makes the many invalid instruments assumption less palatable for this context.

Table 6 reports the estimated effect of pretrial release on conviction, and F-statistics for the first stage. Pretrial release is estimated to reduce the probability of conviction by 16 percentage points. The F-statistic from using the single judge propensity to release instrument overstates the strength of the first stage. Using jackknife instrumental variables estimation with the judge dummies reduces the size of the F-statistic substantially..

7 Conclusion

Judge fixed effects designs, or, more generally, treatment effects estimation involving several dummy instrumental variables, are increasingly popular. Traditional overidentification tests rely on often implausible constant treatment effects assumptions for their validity. More recently developed tests that allow for heterogeneous treatment effects, on the other hand, require that the order of the instruments by treatment propensity be known, something that is rarely the case in judge fixed effects designs. We have proposed a test of the identifying assumptions in judge fixed effects designs that allows for heterogeneous treatment effects and accounts for the estimation of judge propensities, established its asymptotic properties, and demonstrated its finite sample performance in simulations and a real-world application to the effects of pretrial detention. Finally, we provided guidance on steps researchers can take when the test reveals evidence against the identifying assumptions.

References

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

Donald W. K. Andrews and Gustavo Soares. Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157, 2010. doi: 10.3982/ECTA7502.

Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90:430–442, 1995.

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

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

Saurabh Barghava and Dayanand Manoli. Psychological frictions and the incomplete take-up of social benefits: Evidence from an irs field experiment. *American Economic Review*, 105(11):3489–3529, 2015.

Peter Bergman, Jeffrey Denning, and Day Manoli. Broken tax breaks? evidence from a tax credit information experiment with 1,000,000 students. *Working Paper*, 2017.

Manudeep Bhuller, Gordon B Dahl, Katrine V Løken, and Magne Mogstad. Incarceration, recidivism and employment. Working Paper 22648, National Bureau of Economic Research, September 2016. URL `http://www.nber.org/papers/w22648`.

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

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

Tom Chang and Antoinette Schoar. Judge specific differences in chapter 11 and firm outcomes. Technical report, University of Southern California, 2013.

Xiaohong Chen. *Large Sample Sieve Estimation of Semi-Nonparametric Models*, chapter 76, pages 5549–5632. 2007.

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

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

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

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

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

Will Dobbie, Hans Gr onqvist, Susan Niknami, Mårten Palme, and Mikael Priks. The intergenerational effects of parental incarceration. Working Paper 24186, National Bureau of Economic Research, January 2018b. URL http://www.nber.org/papers/w24186.

Joseph J. Doyle, Jr. Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, 97(5):1583–1610, December 2007. doi: 10.

1257/aer.97.5.1583. URL http://www.aeaweb.org/articles?id=10.1257/aer.97.5.1583.

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

Ozkan Eren and Naci Mocan. Juvenile punishment, high school graduation and adult crime: Evidence from idiosyncratic judge harshness. Working Paper 23573, National Bureau of Economic Research, July 2017. URL http://www.nber.org/papers/w23573.

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

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

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

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, July 1982.

Justine Hastings and Jeffrey Weinstein. Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics*, 123(4):1373–1414, 2008.

Martin Huber and Giovanni Mellace. Testing instrument validity for late identification based on inequality moment constraints. *The Review of Economics and Statistics*, 97(2):398–411, 2015. doi: 10.1162/REST_a_00450. URL https://doi.org/10.1162/REST_a_00450.

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

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

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

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

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

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

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

David McKenzie, Suresh de Mel, and Christopher Woodruff. Returns to capital: Results from a randomized experiment. *The Quarterly Journal of Economics*, 123 (4):1329–1372, 2008.

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment effects. Working Paper 23568, National Bureau of Economic Research, July 2017. URL `http://www.nber.org/papers/w23568`.

Ismael Mourifié and Yuanyuan Wan. Testing local average treatment effect assumptions. *The Review of Economics and Statistics*, 99(2):305–313, 2017. URL `https://EconPapers.repec.org/RePEc:tpr:restat:v:99:y:2017:i:2:p:305-313`.

Michael Mueller-Smith. The criminal and labor market impacts of incarceration. 2015.

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

Benjamin Olken. Monitoring corruption: Evidence from a field experiment in indonesia. *Journal of Political Economy*, 115(2):200–249, 2007.

Jeffrey S. Racine. A primer on regression splines. unpublished manuscript, May 2018.

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

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

Rebecca Thornton. The demand for, and impact of, learning hiv status. *American Economic Review*, 98(5):1829–1863, 2008.

Appendix

Proofs

Proof of Theorem 2. Condition 1 satisfies the conditions of Theorem 1 in Imbens and Angrist (1994), which implies

$$E\left[Y_i|J_i=j\right]=\left(p\left(j\right)-p\left(1\right)\right)E\left[Y_i\left(1\right)-Y_i\left(0\right)|D_i\left(j\right)>D_i\left(0\right)\right]+E\left[Y_i|J_i=1\right]. \quad (3)$$

By Condition 1c, monotonicity, for each individual i one can define a marginal propensity, $\bar{p}_i := \inf\left\{p : p=p\left(j\right), D_i\left(j\right)=1\right\}$, such that when assigned a judge with $p\left(j\right)\geq\bar{p}_i$, the individual is treated, and otherwise is untreated. For never-takers, we define $\bar{p}_i=\infty$. For always-takers, $\bar{p}_i=p\left(1\right)$. Note that $\bar{p}_i$ depends only on $D_i\left(j\right)$, and by Condition 1a is therefore independent of J_i, and that potential treatment status can be written $D_i\left(J_i\right)=\bar{p}_i\leq p\left(J_i\right)$. The right hand side of equation (3) then can be written

$$\phi\left(p\left(j\right)\right)=\left(p\left(j\right)-p\left(1\right)\right)E\left[Y_{i}\left(1\right)-Y_{i}\left(0\right)|p\left(1\right)<\bar{p}_{i}\leq p\left(j\right)\right]+E\left[Y_{i}|J_{i}=1\right],$$

which depends on j only through $p\left(j\right)$. By monotonicity the average slope of ϕ through two points p and p' (where $p'\geq p$) can be written:

$$\phi\left(p'\right)-\phi\left(p\right)=\left(p'-p\right)E\left[Y_{i}\left(1\right)-Y_{i}\left(0\right)|p\leq\bar{p}_{i}\leq p'\right].$$

Let $\mathcal{Y}$ be the compact support of Y_i. Noting that $K:=\sup\mathcal{Y}-\inf\mathcal{Y}$ is finite and that $\left|E\left[Y_{i}\left(1\right)-Y_{i}\left(0\right)|p\leq\bar{p}_{i}\leq p'\right]\right|\leq K$ yields the result. ■

Proof of Theorem 3. Define the estimated judge propensity to treat as

$$\begin{aligned}\hat{p}\left(Z_{i}\right) &= W_{i}'\hat{\alpha},\\ \hat{\alpha} &= \left(\sum_{i=1}^{n}W_{i}W_{i}'\right)^{-1}\sum_{i=1}^{n}W_{i}D_{i}.\end{aligned}$$

Define $v_i:=D_i-W_i'\alpha$ and $u_i:=Y_i-S_i'\delta$, where S_i is a vector powers of $p\left(J_i\right):=W_i'\alpha$, judge J_i's (population) propensity to treat, and δ is the vector of coefficients from the population regression of Y_i on S_i. Write $S_i'\delta:=f\left(\lambda,W_i\right)$, where $\lambda=\left(\alpha',\delta'\right)'$. Letting

$$\hat{\delta}=\left(n^{-1}\sum_{i=1}^{n}\hat{S}_{i}\hat{S}_{i}'\right)^{-1}n^{-1}\sum_{i=1}^{n}\hat{S}_{i}Y_{i},$$

we can write $\hat{u}_i=Y_i-f\left(\hat{\lambda},W_i\right)$ where $\hat{\lambda}=\left(\hat{\alpha}',\hat{\delta}'\right)'$, which has limiting behavior as follows:

$$\sqrt{n}\left(\hat{\lambda}-\lambda\right)=n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix}Q_{W}^{-1}W_{i}v_{i}\\ Q_{S}^{-1}S_{i}u_{i}\end{pmatrix}+o_{p}\left(1\right),$$

where for some random vector A_i we adopt the notation $Q_A := E\left[A_i A_i'\right]$. By a mean value expansion we can write $\hat{u}_i = u_i - \nabla\left(\tilde{\lambda}, W_i\right)'\left(\hat{\lambda} - \lambda\right)$, where $\nabla\left(\lambda, W_i\right)$ is the Jacobian of $f\left(\lambda, W_i\right)$ with respect to λ,

$$\nabla\left(\lambda, W_i\right) = \left(\begin{array}{c} W_i \Delta_i' \delta \\ S_i \end{array}\right)$$

and

$$\hat{\Delta}_i = \left(\frac{dS_0\left(\hat{p}_i\right)}{dp}, \ldots, \frac{dS_m\left(\hat{p}_i\right)}{dp}\right)'.$$

The estimator on which the test statistic is based can therefore be expanded as:

$$\begin{aligned} \sqrt{n}\hat{\gamma} &= \sqrt{n}\left(n^{-1}\sum_{i=1}^{n} W_i W_i'\right)^{-1} n^{-1}\sum_{i=1}^{n} W_i \hat{u}_i \\ &= Q_W^{-1}\left(n^{-1/2}\sum_{i=1}^{n} W_i u_i - r_i\right) + o_p\left(1\right), \end{aligned}$$

where

$$r_i = E\left[W_i\left(\begin{array}{c} W_i \Delta_i' \delta \\ S_i \end{array}\right)'\right]\left(\begin{array}{c} Q_W^{-1} W_i\left(D_i - W_i'\alpha\right) \\ Q_S^{-1} S_i u_i \end{array}\right),$$

a consistent estimator for which is

$$\hat{R}_i = \left(n^{-1}\sum_{j=1}^{n} W_j\left(\begin{array}{c} W_j \Delta_j' \hat{\delta} \\ \hat{S}_j \end{array}\right)'\right)\left(\begin{array}{c} \hat{Q}_W^{-1} W_i\left(D_i - \hat{p}_i\right) \\ \hat{Q}_S^{-1} \hat{S}_i \hat{u}_i \end{array}\right). \qquad (4)$$

By the central limit theorem we therefore have

$$\sqrt{n}\hat{\gamma} \underset{d}{\rightarrow} N\left(0, \Omega\right),$$

where

$$\Omega = Q_W^{-1} Var\left(W_i u_i - r_i\right) Q_W^{-1}$$

is consistently estimated by $\hat{\Omega}$ in the text. The quadratic form

$$n\hat{\gamma}'\hat{\Omega}^{-1}\hat{\gamma}$$

is therefore asymptotically a chi-squared random variable with degrees of freedom equal to the rank of $\hat{\Omega}^{-1}$, in this case $k-m$. ■

Proof of Theorem 5. Define and note the following:

$$\begin{aligned}
Y_{ij} &= Y_i\left(1\right) D_i\left(j\right) + Y_i\left(0\right)\left(1 - D_i\left(j\right)\right) \\
\bar{D}_i &= \sum_{j=1}^{J} \lambda_j D_i\left(j\right) \\
p &= \sum_{j=1}^{J} \lambda_j p(j) \\
\bar{Y}_i &: = \left(\bar{D}_i Y_i\left(1\right) + \left(1 - \bar{D}_i\right) Y_i\left(0\right)\right) \\
&= \sum_{j=1}^{J} \lambda_j Y_{ij} \\
E\left[\bar{Y}_i\right] &= E\left[\sum_{j=1}^{J} \lambda_j Y_{ij}\right] \\
&= \left[\sum_{j=1}^{J} \Pr\left(J_i = j\right) E\left[Y_i | J_i = j\right]\right] \\
&= E\left[Y_i\right]
\end{aligned}$$

The IV estimand is the covariance between assigned judge propensity and individual outcome divided by the variance of the judge propensity:

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

Iterating expectations in the numerator and denominator, the right hand side becomes:

$$\begin{aligned} & \frac{\sum_{j=1}^{J} \lambda_j \left(E\left[\left(p(j) - p\right)\left(Y_i - E\left[Y_i\right]\right)|J_i = j\right]\right)}{\sum_{j=1}^{J} \lambda_j \left(p(j) - p\right)^2} \\ = \quad & \frac{\sum_{j=1}^{J} \lambda_j \left(E\left[\left(p(j) - p\right)\left(Y_{ij} - \bar{Y}_i\right)|J_i = j\right]\right)}{\sum_{j=1}^{J} \lambda_j \left(p(j) - p\right) E\left[\left(D_i\left(j\right) - \bar{D}_i\right)\right]}, \end{aligned}$$

where the second line follows from random assignment which implies $E\left[\bar{Y}_i|J_i = j\right] = E\left[\bar{Y}_i\right] = E\left[Y_i\right]$. Noting that $\lambda_j\left(p(j) - p\right)$ is deterministic and that random assignment implies $E\left[Y_{ij}|J_i = j\right] = E\left[Y_{ij}\right]$, the IV estimand can be written:

$$\begin{aligned} & \frac{\sum_{j=1}^{J} \lambda_j \left(p(j) - p\right)\left(E\left[Y_{ij}\right] - E\left[\bar{Y}_i\right]\right)}{E\left[\sum_{j=1}^{J} \lambda_j \left(p(j) - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)\right]} \\ = \quad & \frac{E\left[\sum_{j=1}^{J} \lambda_j \left(p(j) - p\right)\left(Y_{ij} - \bar{Y}_i\right)\right]}{E\left[\sum_{j=1}^{J} \lambda_j \left(p(j) - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)\right]} \\ = \quad & \frac{E\left[\sum_{j=1}^{J} \lambda_j \left(\left(p(j) - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)\right)\left(Y_i\left(1\right) - Y_i\left(0\right)\right)\right]}{E\left[\sum_{j=1}^{J} \lambda_j \left(p(j) - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)\right]}, \end{aligned}$$

where the first equality follows from the interchangeability of integration and summation, and the final equality from the definitions of Y_{ij} and $\bar{Y}_i$.

Hence, given random assignment and the exclusion restriction (and notably without imposing monotonicity), the IV estimand can be written as a weighted average of individual-level treatment effects:

$$\beta_{IV} = \frac{E\left[\omega_i \left(Y_i\left(1\right) - Y_i\left(0\right)\right)\right]}{E\left[\omega_i\right]},$$

where the weights are given by

$$\omega_i := \sum_{j=1}^{J} \lambda_j \left(p(j) - p\right)\left(D_i\left(j\right) - \bar{D}_i\right).$$

■

Extensions

The proposed test has power against alternatives that shift the mean of Y_i, but will not have power against alternatives where other features of Y_i are changed but not the mean. The test naturally extends to have power against shifts in other features of the distribution, as well. Instead of regressing Y_i on the instrument indicators and the propensity power series, one can simultaneously regress a set of indicator variables of the form $1\left(Y_i \leq y_j\right)$ for a grid of $\{y_j\}$ values and jointly test whether the coefficients on the instrument indicators are zero across all equations.

Alternatively, one can replace the mean regression with a set of quantile regressions of Y_i with a grid of quantile values $\tau_j \in (0,1)$. The test then consists of jointly testing the hypothesis that the coefficients on the instrument dummies are zero across all quantile regressions. In this case and the dummy dependent variable alternative above, the test is carried out using a variance matrix accounting for the estimation of $\hat{p}\left(J_i\right)$, analogous to the procedure described in the main text. These extensions allow the test to have power against a wider array of alternatives, although at the expense of more computational burden and perhaps a lack of specific power against alternatives where only the mean is shifted.

Generalized Moment Selection Implementation

The slope component of the test implements the moment inequality testing procedure proposed by Andrews and Soares (2010). This procedure is based on the following modified method of moments (MMM) test statistic:

$$\hat{M} = \sum_{l=0}^{m-1} \left(\left[\frac{K - \hat{\phi}'(t_l)}{s.e.\left(\hat{\phi}'(t_l)\right)} \right]_-^2 + \left[\frac{K + \hat{\phi}'(t_l)}{s.e.\left(\hat{\phi}'(t_l)\right)} \right]_-^2 \right),$$

where $[x]_- = x1(x < 0)$,

$$\hat{\phi}'(t_l) = \frac{2}{t_{l+1} - t_{l-1}} \left(\hat{\delta}_{l+1} - \hat{\delta}_l \right),$$

$$s.e.\left(\hat{\phi}'(t_l)\right) = n^{-1/2} \frac{2}{t_{l+1} - t_{l-1}} \left(\hat{\Sigma}_{l+1,l+1} + \hat{\Sigma}_{l,l} - 2\hat{\Sigma}_{l+1,l} \right)^{1/2},$$

and $\hat{\Sigma}$ is a consistent estimator of the variance matrix of $\hat{\delta} = \left(\sum_{i=1}^n \hat{S}_i \hat{S}_i' \right)^{-1} \sum_{i=1}^n \hat{S}_i Y_i$ that takes into account estimation of $\hat{P}_i$:

$$\hat{\Sigma} = \hat{Q}_S^{-1} \left(\sum_{i=1}^n \left(\hat{S}_i \hat{u}_i - \hat{\Delta}_i W_i' \hat{Q}_W^{-1} W_i \hat{v} \right) \left(\hat{S}_i \hat{u}_i - \hat{\Delta}_i W_i' \hat{Q}_W^{-1} W_i \hat{v} \right)' \right) \hat{Q}_S^{-1}.$$

Under the regularity conditions described in Andrews and Soares (2010), the distribution of the MMM test statistic can be approximated by the distribution of

$$\hat{M}^* = \sum_{l \in \mathcal{L}^-} [Z_l^*]_-^2 + \sum_{l \in \mathcal{L}^+} [-Z_l^*]_-^2,$$

where Z^* is an m-element multivariate normal random variable with unit variances and correlation matrix corresponding to the asymptotic variance of

$$\left([0_{m\times 1} : I_m] - [I_m : 0_{m\times 1}]\right)\hat{\delta},$$

and the moments selected by the generalized moment selection are given by:

$$\mathcal{L}^{-} = \left\{ l : \frac{K - \hat{\phi}'(t_l)}{s.e.\left(\hat{\phi}'(t_l)\right)} \leq \sqrt{\ln n} \right\},$$

and

$$\mathcal{L}^{+} = \left\{ l : \frac{K + \hat{\phi}'(t_l)}{s.e.\left(\hat{\phi}'(t_l)\right)} \leq \sqrt{\ln n} \right\}.$$

The p-value from the slope component of the test can be found to arbitrary precision by simulating many multivariate draws, constructing $\hat{M}^*$ for each draw, and computing the fraction of draws for which $\hat{M}^* \geq \hat{M}$.

Tables

Table 1: Summary Statistics

	Detained mean	Released mean	Full Sample mean
Panel A: Bail Information			
Release on Recognizance	0.05	0.11	0.07
Non-Monetary Bail	0.11	0.38	0.20
Monetary Bail	0.83	0.51	0.73
Bail Amount (in thousands)	59.94	24.80	48.38
Released in 14 Days	0.06	1.00	0.37
Released Before Trial	0.33	1.00	0.55
Panel B: Defendant Characteristics			
Male	0.87	0.79	0.84
White	0.46	0.50	0.48
Black	0.54	0.50	0.52
Age at Bail Decision	36.52	34.00	35.69
Prior Offense in Past Year	0.40	0.22	0.34
Panel C: Charge Characteristics			
Number of Offenses	1.66	1.59	1.64
Felony Offense	0.51	0.56	0.53
Misdemeanor Only	0.49	0.44	0.47
Any Drug Offense	0.27	0.30	0.28
Any DUI Offense	0.00	0.00	0.00
Any Violent Offense	0.14	0.30	0.19
Any Property Offense	0.41	0.23	0.35
Panel E: Outcomes			
Any Guilty Offense	0.67	0.41	0.58
Guilty Plea	0.58	0.31	0.49
Any Incarceration	0.25	0.17	0.22
Rearrest in 0-2 Years	0.53	0.37	0.48
Rearrest Prior to Disposition 0-2 Years	0.14	0.16	0.15
Rearrest After Disposition	0.40	0.25	0.35
Observations	62644	30714	93358

Table 2: Test of Random Judge Assignment

	(1) Released in 3 Days	(2) Judge Leniency
Male	-0.11***	0.00
	(0.00)	(0.00)
Black	-0.03***	0.00
	(0.00)	(0.00)
Age at Bail Decision	-0.03***	-0.00
	(0.00)	(0.00)
Prior Offense in Past Year	-0.16***	0.00
	(0.00)	(0.00)
Number of Offenses	-0.02***	0.00
	(0.00)	(0.00)
Felony Offense	0.34***	0.02
	(0.07)	(0.01)
Any Drug Offense	0.04***	0.00
	(0.01)	(0.00)
Any Violent Offense	0.17***	-0.00
	(0.01)	(0.00)
Any Property Offense	-0.12***	-0.00
	(0.00)	(0.00)
Missing Race	-0.05	-0.00
	(0.03)	(0.00)
Joint Test p-value	0.00	0.37
N	93358.00	93358.00

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Test Results

	5 knots	10 knots	15 knots	20 knots
Full sample	935	839	819	749
	(504)	(499)	(494)	(489)
	[0.000]	[0.000]	[0.000]	[0.000]

Note: This table displays the test statistics, degrees of freedom, and associated p-values from the proposed test. Each column shows results using a different number of knots in the spline function.

Table 4: First Stage Results by Case Characteristics

	Crime Severity		Crime Type		
	Misd.	Felony	Drug	Property	Violent
Judge Leniency	0.699***	0.281***	0.571***	0.479***	0.060
	(0.057)	(0.048)	(0.061)	(0.053)	(0.065)
Observations	44130	49228	24987	30855	17015

Note: This table shows the results from regressing pretrial release status on the judge leniency instrument for subgroups defined by case characteristics. All specifications include controls for crime type and severity; whether the defendant is black; whether the defendant is male; age categories; whether the defendant has a prior offense within the past year; the number of counts; and whether the charges include any drug crimes, and violent crimes, and any property crimes. Standard errors are twoway clustered at the individual and judge levels.

Table 5: First Stage Results by Defendant Characteristics

	Priors		Race	
	No prior	Prior	Black	White
Judge Leniency	0.491*** (0.042)	0.463*** (0.054)	0.527*** (0.041)	0.439*** (0.046)
Observations	61697	31661	48900	44313

Note: This table shows the results from regressing pretrial release status on the judge leniency instrument for subgroups defined by defendant characteristics. All specifications include controls for crime type and severity; whether the defendant is black; whether the defendant is male; age categories; whether the defendant has a prior offense within the past year; the number of counts; and whether the charges include any drug crimes, and violent crimes, and any property crimes. Standard errors are twoway clustered at the individual and judge levels.

Table 6: First Stage and IV Results

	Judge IV
Effect of pretrial release	
Convicted	-0.165* (0.079)
First-stage F for Z	226.291 (0.000)
First-stage F for dummies	5.972 (0.000)

Note: The top panel displays the estimated effect of pretrial release on conviction, instrumenting for whether bail was met using judge leniency in column 1 and tercile leninecy in column 2. The bottom panel shows first-stage F statistics, and their corresponding p-values, using first the leniency instrument and then judge or tercile dummies. The first stage using dummies uses jackknife instrumental variables estimation.

Figures

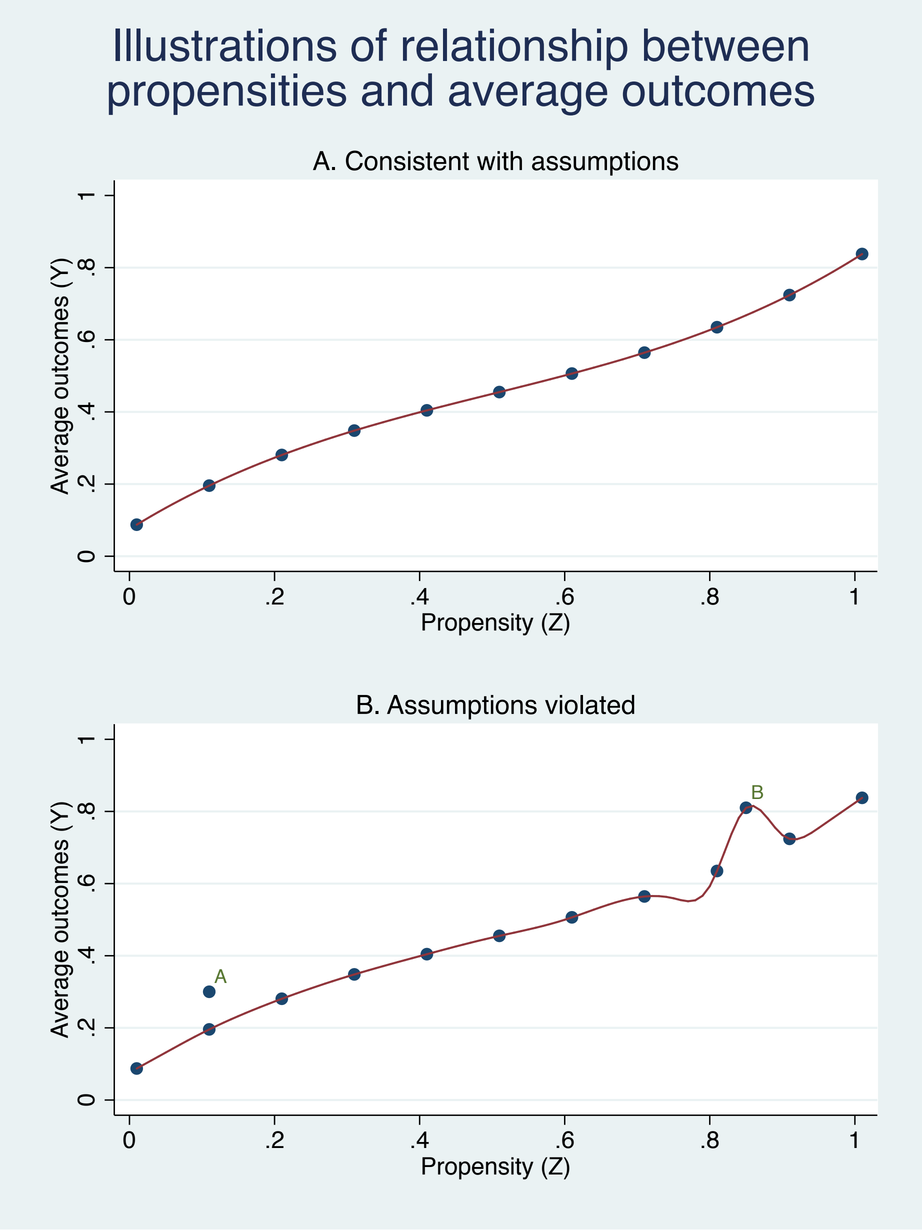

Figure 1: Illustrations of hypothetical relationships between true judge propensities to assign treatment and expected outcomes. Each dot represents a single judge. The pattern in Panel A is consistent with the exclusion restriction and monotonicity, because all the dots lie on a continuous function whose slope is nowhere larger in magnitude that the largest possible treatment effects, given a binary outcome. The pattern in Panel B could only arise if one or more of the assumptions were violated. The judge labeled "A" has exactly the same propensity as another judge, but different expected outcomes. The judge labeled "B" lies on a segment of the curve whose slope is larger than one, implying an impossibly large treatment effect.

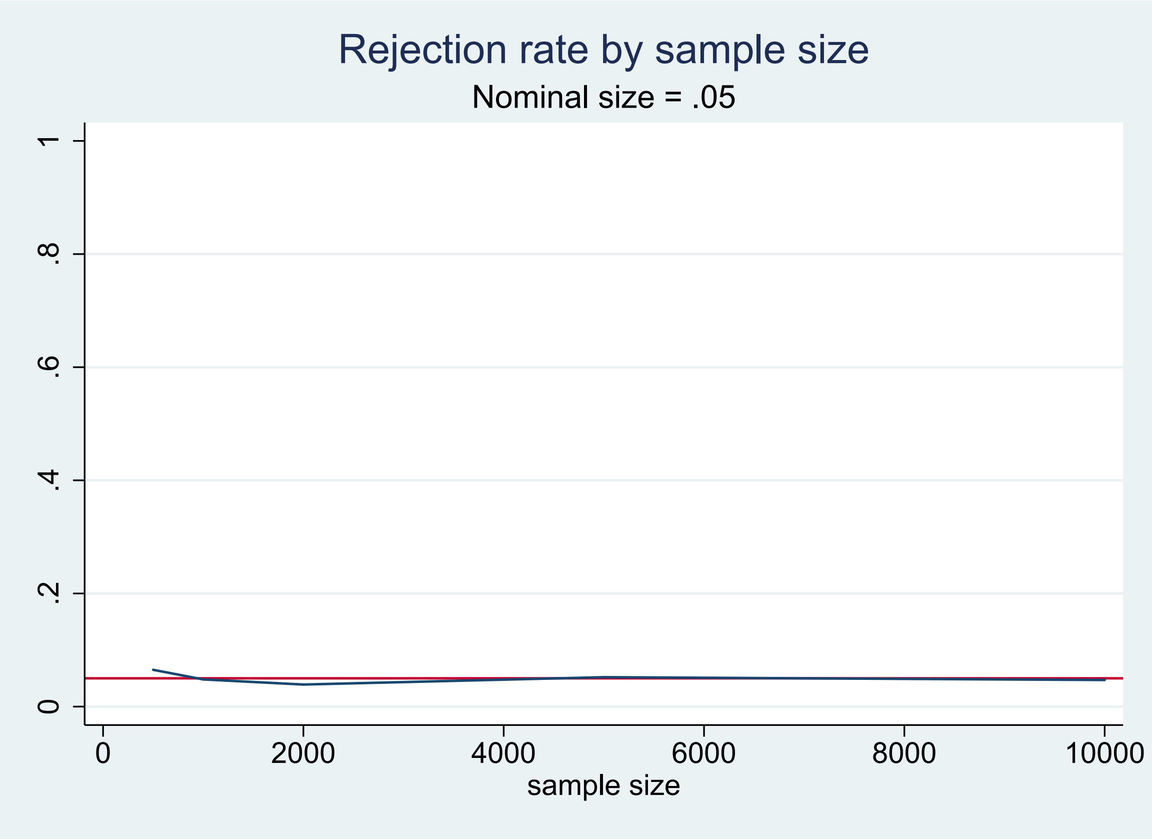

Figure 2: Monte Carlo simulation rejection rates from the test for instrument validity as a function of the sample size (x-axis). The nominal size of the tests is .05. Based on 999 iterations.

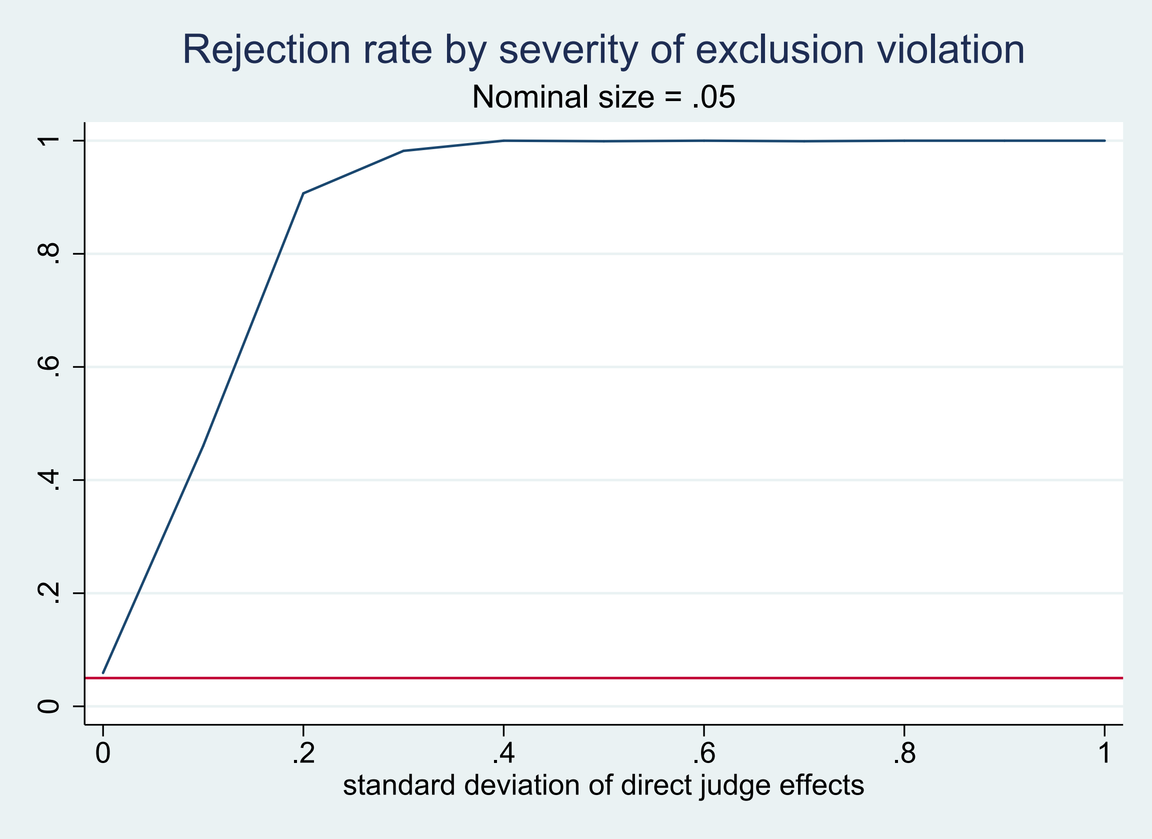

Figure 3: Monte Carlo simulation rejection rates from the test for instrument validity as a function of the severity of the exclusion restriction violation, as measured by the standard deviation of the direct judge effects (x-axis). The nominal size of the tests is .05. Based on 999 iterations.

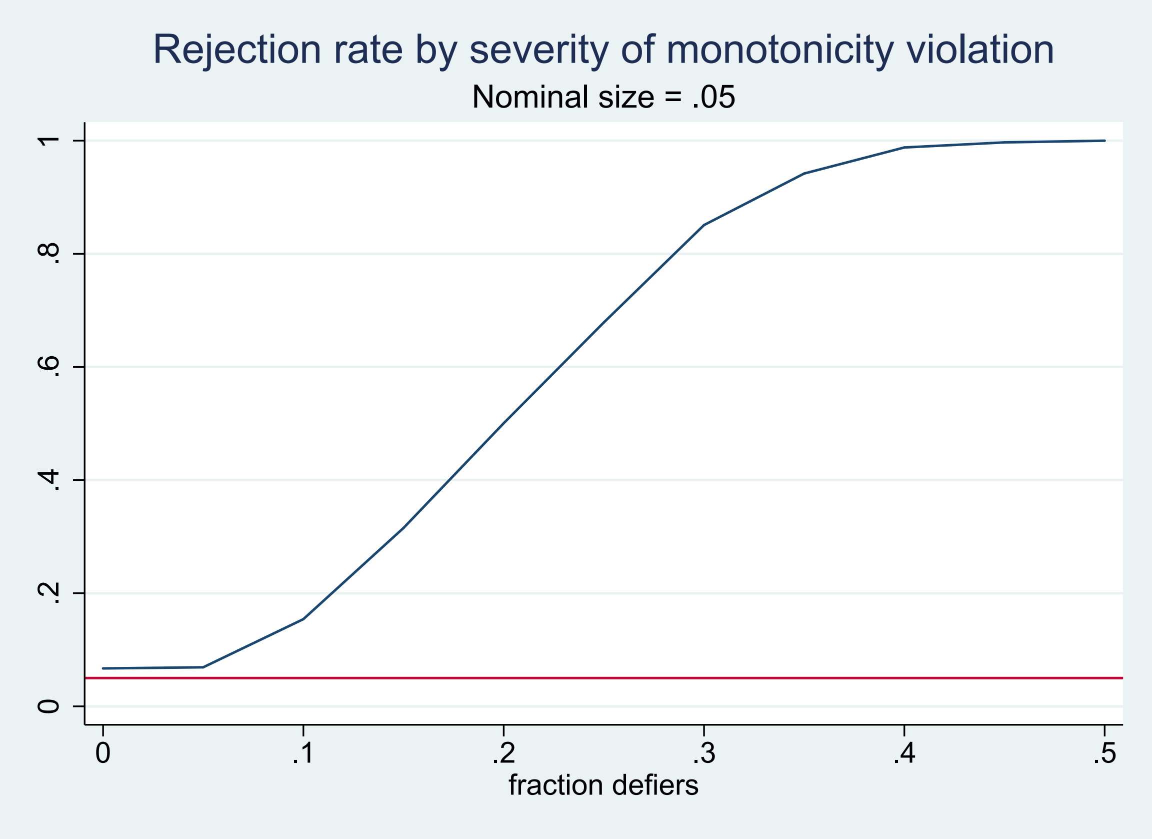

Figure 4: Monte Carlo simulation rejection rates from the test for instrument validity as a function of the severity of the monotonicity violation, as measured by the fraction of defendants for whom judges disagree on the ordering. The nominal size of the tests is .05. Based on 999 iterations.

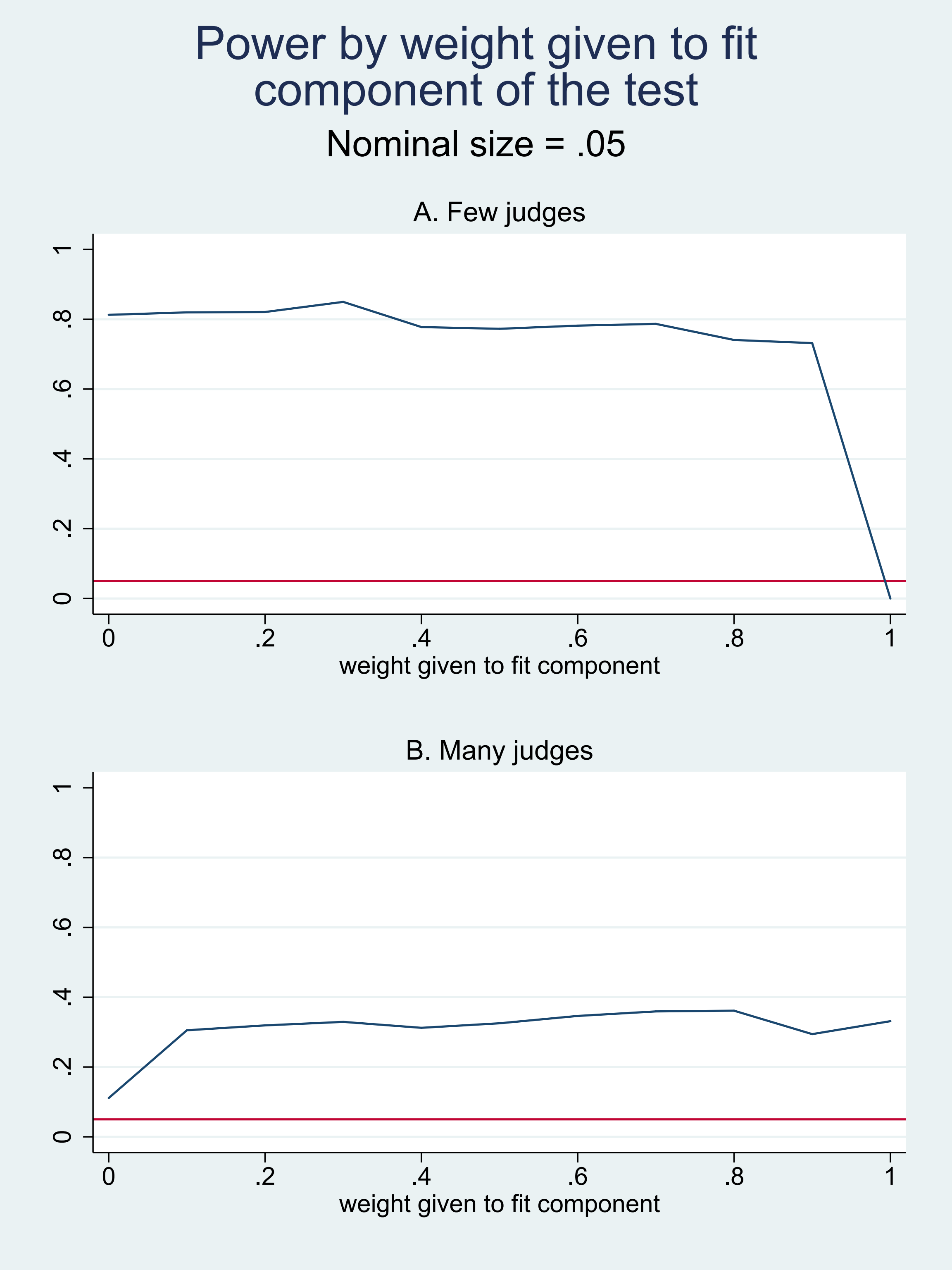

Figure 5: Monte Carlo simulation rejection rates from the test for instrument validity as a function of the weight given to the fit component of the test. The upper panel sets $J = 2$. The lower panel sets $J = 20$. The nominal size of the tests is .05. Based on 999 iterations.

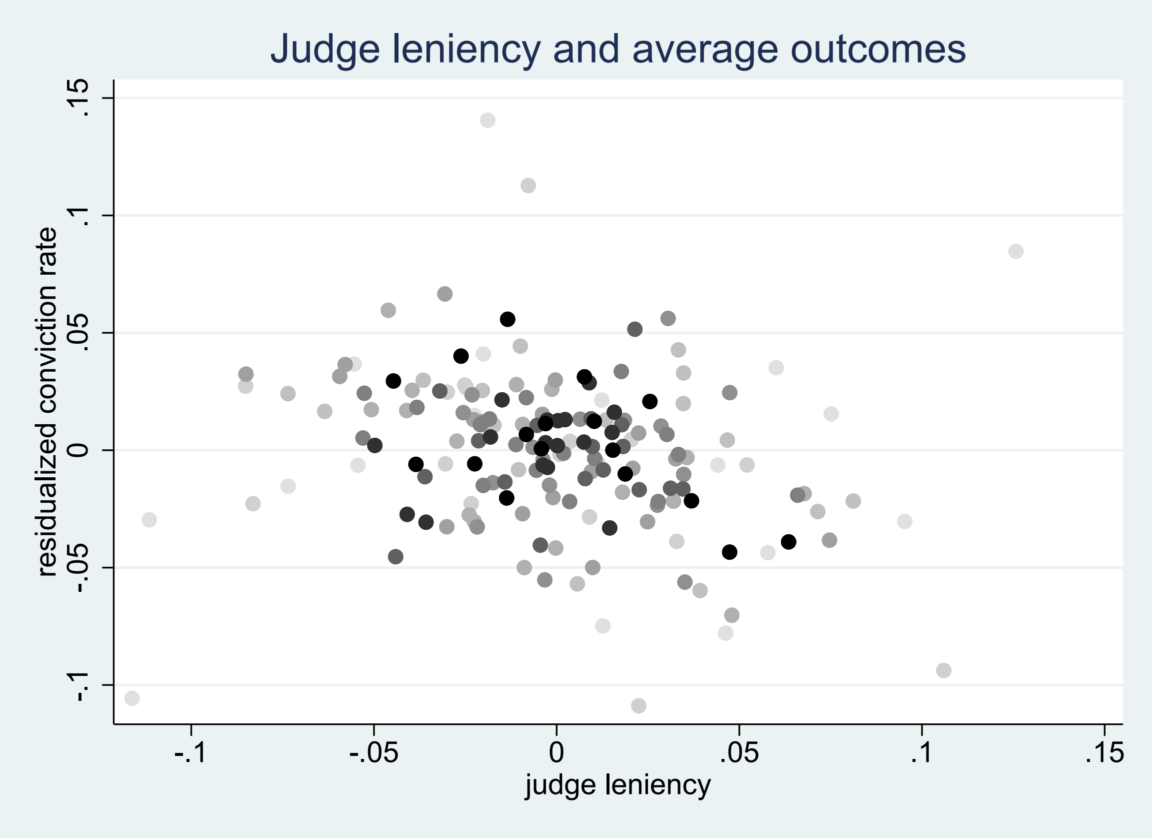

Figure 6: Each dot corresponds to an individual bail judge. The x-axis measures leniency, and is the average of the leave-out mean instrument for each bail judge. The y-axis measures conviction rates for defendants who appeared before each bail judge. Specifically, it is the average of the residual from regressing a dummy variable for conviction on time/place fixed effects, and a vector of defendant and case characteristics. Dot darkness reflects the number of bail hearings presided over by the judge, with darker dots for higher caseloads. The median judge presided over 506 cases during our sample period.