

# Canonical Research Designs VI: Examiner Designs (aka Judge IV, aka leniency IV)

Paul Goldsmith-Pinkham

April 13, 2023

# Roadmap for Today

- Today we're focusing on a particular research design: the examiner design
- Like many of the designs we are studying, they have shown up in papers for a while, but have really started to take off recently with the rise of high-quality administrative microdata.
- Key issues we'll touch on:
  - Identification: what are we getting at?
  - Estimation: What's the best estimation method for using the examiner design?
  - Inference: How should we do inference?

## High-level description of examiner design

- In many applications, there is an administrator, judge, or monitor who plays an important role in deciding an outcome
- These outcomes include:
  - bail
  - bankruptcy
  - getting a loan
  - parole
  - disability insurance
  - patent granting
  - cancer screening
- In many cases, this examiner is effectively randomly assigned, *and* there is wide-range of differences (and discretion) in how likely they are to decide the outcome

## High-level description of examiner design

- In many applications, there is an administrator, judge, or monitor who plays an important role in deciding an outcome
- These outcomes include:
  - bail
  - bankruptcy
  - getting a loan
  - parole
  - disability insurance
  - patent granting
  - cancer screening
- In many cases, this examiner is effectively randomly assigned, *and* there is wide-range of differences (and discretion) in how likely they are to decide the outcome
- Key ingredients:
  1. random assignment of examiner
  2. discretion over a (typically binary) outcome
  3. heterogeneity in behavior

# Consider two examples

- Example 1: Bail setting in Philadelphia (Stevenson (2018))
- After you're arrested, you can be:
  - held in jail
  - released on bail (you pay \$)
  - released on recognizance (no bail \$)
- This decision is made by one of six magistrates at preliminary hearings
  - The particular magistrate that a defendant faces depends on a rotating schedule
  - Hence, the randomness occurs from *the time of day*

## Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes

Megan T. Stevenson\*  
George Mason University

This article uses a natural experiment to analyze whether incarceration during the pretrial period affects case outcomes. In Philadelphia, defendants randomly receive bail magistrates who differ widely in their propensity to set bail at affordable levels. Using magistrate leniency as an instrument, I find that pretrial detention leads to a 13% increase in the likelihood of being convicted, an effect largely explained by an increase in guilty pleas among defendants who otherwise would have been acquitted or had their charges dropped. I find also that pretrial detention leads to a 42% increase in the length of the incarceration sentence and a 41% increase in the amount of nonbail court fees owed. This latter finding contributes to a growing literature on fines-and-fees in criminal justice, and suggests that the use of money bail contributes to a “poverty-trap”: those who are unable to pay bail wind up accruing more court debt. (JEL K14)

I have had the “you can wait it out or take the deal and get out” conversation with way too many clients.  
—a public defender, Philadelphia

## Consider two examples

- Example 1: Bail setting in Philadelphia (Stevenson (2018))
- After you're arrested, you can be:
  - held in jail
  - released on bail (you pay \$)
  - released on recognizance (no bail \$)
- This decision is made by one of six magistrates at preliminary hearings
  - The particular magistrate that a defendant faces depends on a rotating schedule
  - Hence, the randomness occurs from *the time of day*

Philadelphia employs six Arraignment Court Magistrates at a time, and one of the six will be on duty 24 hours a day, 7 days a week, including holidays. Each day is composed of three work shifts: graveyard (11:30 p.m.–7:30 a.m.), morning (7:30 a.m.–3:30 p.m.) and evening (3:30 p.m.–11:30 p.m.). Each magistrate will work for five days on a particular shift, take five days off, then do five days on the next shift, five days off, and so forth. For example, a magistrate may work the graveyard shift from January 1st to January 5th, have January 6th–10th off, then work the morning shift from January 11th–15th, have the 16th–20th off, do the evening shift from January 21st–25th, take the next five days off, and then start the cycle all over again.

# Consider two examples

- Example 2: Bankruptcy judges in Chapter 13 (Dobbie et al.(2017))
- When you file for Chapter 13 bankruptcy, you file a repayment plan to be approved by a judge
- This decision is made by different judges staffed at a court on a given day
  - The assigned judge is done randomly by the court
  - Hence, the randomness occurs within *location-time* of filing

Consumer Bankruptcy and Financial Health  
Will Dobbie, Paul Goldsmith-Pinkham, and Crystal Yang  
NBER Working Paper No. 21032  
March 2015  
JEL No. D14,K35

## **ABSTRACT**

This paper estimates the effect of Chapter 13 bankruptcy protection on post-filing financial outcomes using a new dataset linking bankruptcy filings to credit bureau records. Our empirical strategy uses the leniency of randomly-assigned judges as an instrument for Chapter 13 protection. Over the first five post-filing years, we find that Chapter 13 protection decreases an index measuring adverse financial events such as civil judgments and repossessions by 0.316 standard deviations, increases the probability of being a homeowner by 13.2 percentage points, and increases credit scores by 14.9 points. Chapter 13 protection has little impact on open unsecured debt, but decreases the amount of debt in collections by \$1,315.

Will Dobbie  
Industrial Relations Section  
Princeton University  
Firestone Library  
Princeton, NJ 08544-2098  
and NBER  
wdobbie@princeton.edu

Crystal Yang  
Harvard Law School  
1585 Massachusetts Avenue  
Griswold 301  
Cambridge, MA 02138  
cyang@law.harvard.edu

Paul Goldsmith-Pinkham  
Graduate School of Business Administration  
Harvard University  
Soldiers Field  
Boston, MA 02163  
paulgp@gmail.com

## Consider two examples

- Example 2: Bankruptcy judges in Chapter 13 (Dobbie et al.(2017))
- When you file for Chapter 13 bankruptcy, you file a repayment plan to be approved by a judge
- This decision is made by different judges staffed at a court on a given day
  - The assigned judge is done randomly by the court
  - Hence, the randomness occurs within *location-time* of filing

Bankruptcy judges are federal judges appointed to 14-year terms by the Court of Appeals in their judicial district. There are a total of 94 federal bankruptcy courts in the United States, including at least one bankruptcy court in each state, the District of Columbia, and Puerto Rico. Each bankruptcy court hears all cases originating from counties in its jurisdiction, and are often further divided into offices that hear all cases originating from a subset of counties in the court's jurisdiction. Bankruptcy judges often hear cases across multiple offices within their court, but only hear cases filed in their bankruptcy court. These cases are typically assigned to judges using a random number generator or a blind rotation system within each office.<sup>8</sup>



## Notation

- Before we get to intuition, let's start with some notation
- We consider  $n$  individuals indexed by  $i$ , with two outcomes:  $D_i$  and  $Y_i$
- Each individual is assigned to one of  $K$  examiners:  $Q_i \in \{0, \dots, K - 1\}$
- Hence, very easy to consider the potential outcomes  $D_i(q)$  for each of the potential examiners, where we observe only one!
  - If  $K = 2$ , this is just our simple binary case.
  - With  $K > 2$ , it becomes more complicated
  - Note that there's no meaningful ordering to the  $K$  examiners
- We can *also* consider the potential outcomes  $Y_i(q)$ . What are we ignoring?

## Notation

- Before we get to intuition, let's start with some notation
- We consider  $n$  individuals indexed by  $i$ , with two outcomes:  $D_i$  and  $Y_i$
- Each individual is assigned to one of  $K$  examiners:  $Q_i \in \{0, \dots, K - 1\}$
- Hence, very easy to consider the potential outcomes  $D_i(q)$  for each of the potential examiners, where we observe only one!
  - If  $K = 2$ , this is just our simple binary case.
  - With  $K > 2$ , it becomes more complicated
  - Note that there's no meaningful ordering to the  $K$  examiners
- We can *also* consider the potential outcomes  $Y_i(q)$ . What are we ignoring?
  - Potential impact of  $D_i$  on  $Y_i$ . When we do IV, we'll need to consider  $Y_i(D_i(Q_i), Q_i)$ , and then shut down the direct effect of  $Q_i$  in order to do IV

# Intuition

- First consider the variation in  $D_i$  across the  $Q_i$ .
  - Conditional on some covariates  $W_i$ , we assume that strong ignorability holds
- If  $W_i$  is just a constant, then we can consider  $\tau_{q,q'} = E(D_i|Q_i = q) - E(D_i|Q_i = q')$ 
  - This is the relative effect of judge  $q$  vs.  $q'$  on bail decisions or bankruptcy discharge
- We have an RCT, but with judges!
- Useful to also define  $\mu_D(q) = E(D_i|Q_i = q)$ 
  - $\hat{\mu}_D(q)$  is the corresponding empirical estimate

## What is this measure?

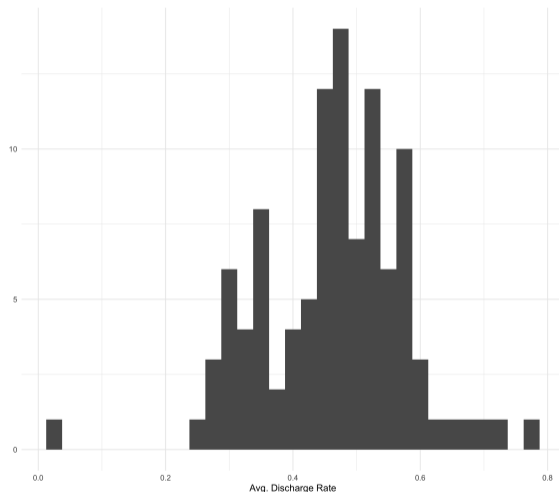
- In the context of bankruptcy or bail judges,  $\mu_D(q)$  measures judge  $q$ 's *leniency* – e.g. their average propensity
- We can literally estimate this with a simple linear regression of dummies! E.g. let  $Q_i$  be the set of dummies for  $Q_i$ :

$$D_i = Q_i \mu_D + u_i$$

- Note that given the simplicity of this measure, the  $\mu_D$  are equivalent to the predicted values for  $D_i$ 
  - E.g.  $\hat{D}$  from the regression on  $Q_i$
- This predicted measure dovetails exactly into the measure that papers have historically used!

## What's the instrument that people have used?

- In many papers, the measure that has been used is a “leniency” metric for judges: e.g. the  $\hat{D}_i$
- This is the overall average leniency across judges in DGPY [note, not observation weighted – hence outliers]
- What's the problem with just the overall, in the DGPY context?
  - Judges are not random overall – location (and time) specific
- Importantly, the average judge characteristic may capture location specific features
  - Control for  $W_i$



## How to incorporate $W_i$ ?

- Now consider the simple linear regression with  $W_i$ :

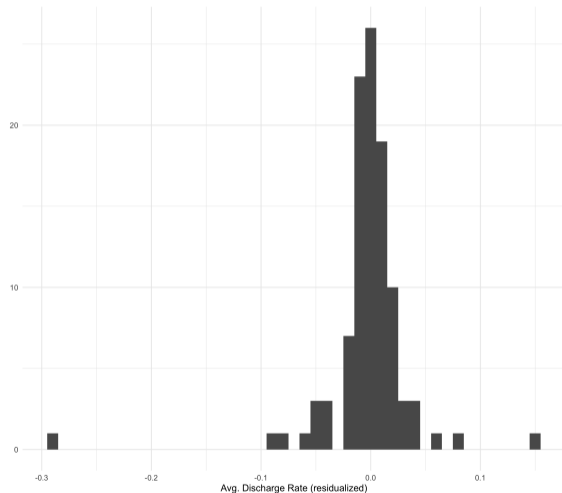
$$D_i = Q_i \mu_D + W_i \gamma + u_i$$

- Now, the  $\mu_D$  are equivalent to the predicted values for  $D_i$  after residualizing out for  $W_i$  (which usually includes a constant)
  - E.g.  $\hat{D}^\perp$  from the regression on  $Q_i^\perp$
  - Call this measure  $Z_i = \hat{D}^\perp$  our leniency measure
- This captures the average variation in leniency within the average of a location
  - In the simple case where judges are nested within courts, this just captures variation across  $Q$  due to judge specific random variation
  - Mechanically, this would literally capture

$$\hat{D}^\perp = n^{-1} \left( \underbrace{\sum_i 1(Q_i = q) D_i}_{\text{judge mean}} - \underbrace{\sum_i 1(W_i = w) D_i}_{\text{location mean}} \right)$$

## What's the instrument that people have used?

- Once you do this exercise with judge leniency, you get a much more recentered object
- This captures across location differences, and give true variation
- Point worth noting – due to the location effect, we can't estimate the “true” judge leniency – just relative leniency within an office



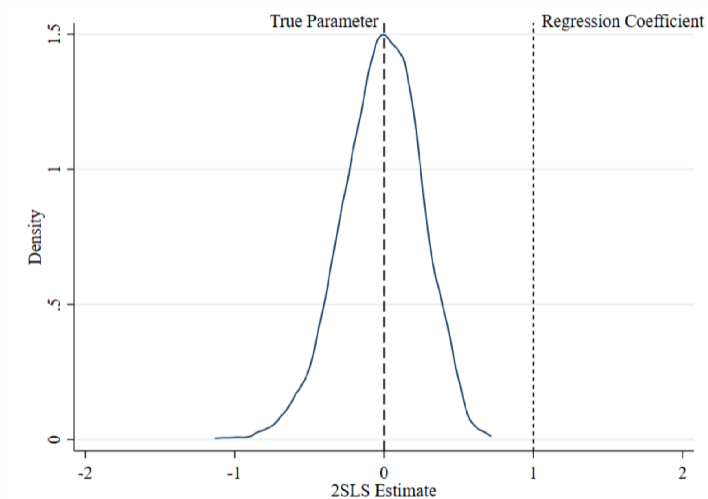
## Important note: Leave-out

- In practice, individuals use the “leave-one-out” mean, rather than the actual average
- Why? Because if you include your own observation, that will be endogeneously correlated
- The “leave-one-out” leniency is the mechanical solution that deals with the many instrument problem



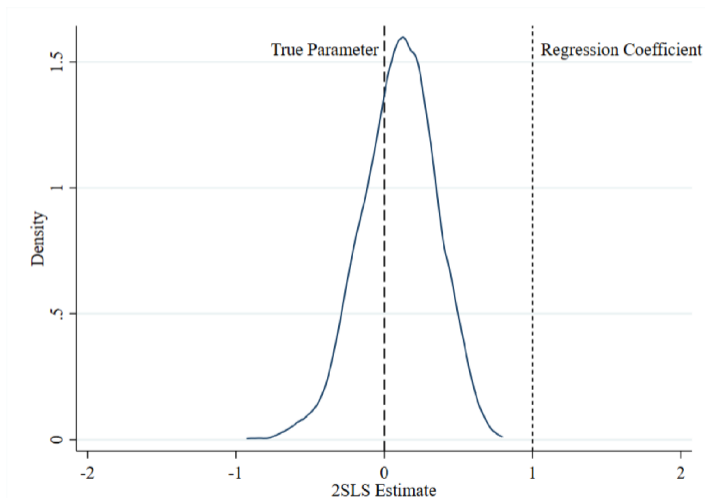
# Recall Many IV bias is pernicious

Monte Carlo:  $Y_i = \varepsilon_i$ ,  $D_i = \Pi Z_{i1} + \eta_i$ : IV with one  $Z_{i1}$



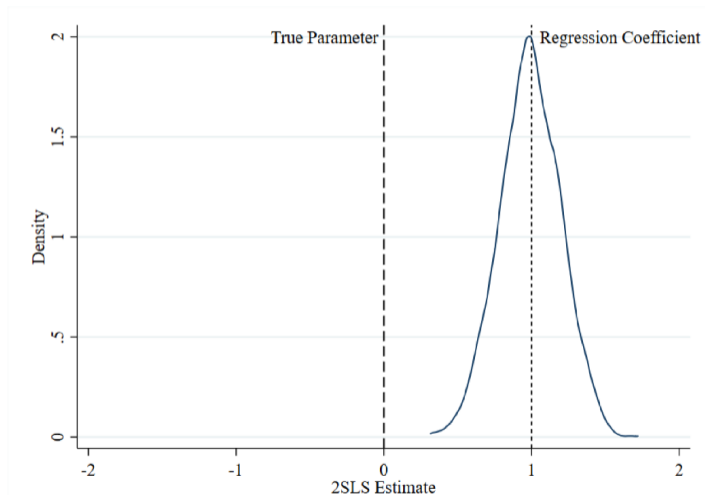
# Recall Many IV bias is pernicious

Monte Carlo:  $Y_i = \varepsilon_i$ ,  $D_i = \Pi Z_{i1} + \eta_i$ : IV with ten  $Z_{ij}$



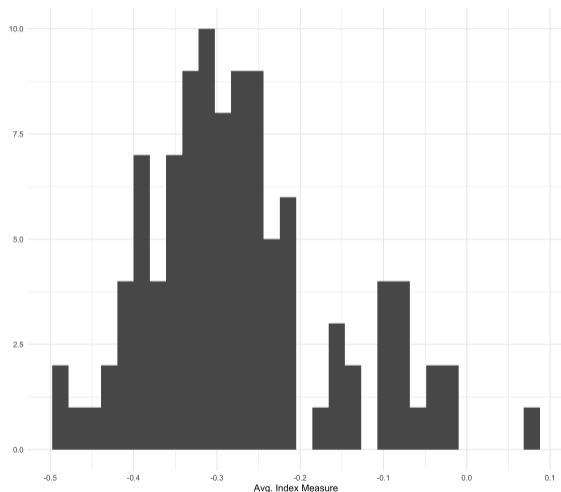
# Recall Many IV bias is pernicious

Monte Carlo:  $Y_i = \varepsilon_i$ ,  $D_i = \Pi Z_{i1} + \eta_i$ : IV with 100  $Z_{ij}$



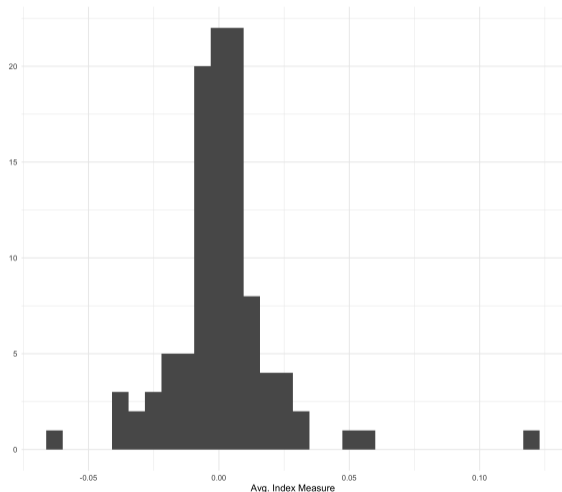
# Nothing stops us from doing the same with outcomes!

- We can do the same exercise with our outcome measures!
- Nothing changes in this setting – we have both location and judge effects, and we can see differences if we don't account for them
- Location variation absorbs a large chunk, but we still see variation in our outcome caused by variation in judges



## Nothing stops us from doing the same with outcomes!

- We can do the same exercise with our outcome measures!
- Nothing changes in this setting – we have both location and judge effects, and we can see differences if we don't account for them
- Location variation absorbs a large chunk, but we still see variation in our outcome caused by variation in judges



# Thinking about instrumental variables

- The reason that this approach is used, of course, is to use the variation in  $\mu_D(q)$  to identify the effect of  $D$  on  $Y$
- So what do we need? Recall
  1. relevance – e.g. we need that our instrument is predictive of  $D$
  2. exclusion – e.g. we need that the examiner *only* has an effect on  $Y$  e.g. potential outcomes  
 $Y_i(D_i(Q_i), Q_i) = Y_i(D_i(Q_i))$
  3. monotonicity – e.g. we need an ordering in the effects
- The last two are the challenging part, and not inherently testable. Let's discuss the issues, but first, how is this done in practice?

## How is this done in practice, usually?

- In many papers (including my own, historically), the *leniency* measure  $Z_i$  has been used, rather than the dummies for judges
- Why? A number of reasons:
  - Faster – one time calculation vs. over identified 2sls
  - Visualization – plotting the reduced form and first stage against leniency is very intuitive
  - Worry about first stage power is easier in the just identified case
- But note the following equivalency from our GMM/2SLS estimator:

$$\hat{\beta}_{2SLS} = \frac{D' Q (Q' Q)^{-1} Q' Y}{\underbrace{D' Q (Q' Q)^{-1} Q' D}_{P_Q}} = \frac{\hat{D}' Y}{\hat{D}' D} \quad (1)$$

- Using many instruments and using the predicted first stage as your instrument are the same thing (adding controls just adds residualization)

## So which is it?

- Our variation is really the random assignment of  $K$  judges. Collapsing to a predicted first stage doesn't change this, and if anything masks the experiment
- The estimator is exactly the same (or close, once we deal with some estimation issues)
  - If this were the right approach, we could do this in every overidentified setting!
- The issue is with inference – e.g. how much uncertainty is there in the underlying projections?
  - Consider the real line with the  $\mu_D(q)$  and  $\hat{\mu}_D(q)$ . Focusing on the estimated leniency measure ignores potential important underlying variation in the first stage estimates
- In some cases, using the overidentified approach and the leniency just-identified approach give very similar standard errors – this is due to relative precision in estimates in the  $\mu$ , and is not guaranteed!



## An important reason to like leniency measures

- The biggest reason why researchers used the leniency measure is it gave a natural way to deal with the “own-observation” problem
- Note that  $\hat{\mu}_D(q)$  includes individual  $i$ 's observation in the estimation procedure, scaled by  $n^{-1}$ 
  - This term is endogeneous!
- Leniency measures that researchers construct use a “leave-one-out” mean to account for this, and instead measure an individual's leniency exposure as a judge's leniency *excluding* own observation
- This variable was easily plugged into 2SLS and avoided the bias
- But... this approach just approximates jackknife IV!
  - This own observation issue is exactly the bias that came up from overidentified IV in finite samples
  - When the number of judges is large-ish within a court, this can be an issue

## Use (U)JIVE!

- Given the leniency approach is exactly solved with a known technique, there is no great reason to use leniency directly
- It is more transparent to use the judge variation directly
- If you want to do graphical visualization, just use the first stage coefficients!
- A key issue raised in Kolesar (2013) is the issue of many controls (e.g. many fixed effects) which creates the same problem as many instruments under certain settings
  - Provides a solution, through UJIVE. See his website for code.

# How to test these exclusion and monotonicity?

- Testing exclusion is always challenging
- However, like standard treatment designs and RCTs, we can test for balance
- Use the predicted first stage (e.g. the propensity score) and test of excluded covariates across p-scores  $\hat{\mu}_D(q)$
- See Aronow and Miller for discussion on balance tests

Table 1  
Descriptive Statistics and Randomization Balance

	All Credit Users		Judge Sample		
	Full Sample	Bankruptcy Filers	Harsh Judge	Lenient Judge	p-value
<i>Panel A: Judge Leniency</i>	(1)	(2)	(3)	(4)	(5)
Judge Leniency	-	-	-0.013	0.012	0.000
<i>Panel B: Baseline Characteristics</i>					
Age	48.549	43.699	44.843	44.863	0.229
Homeowner	0.470	0.520	0.668	0.643	0.175
<i>Panel C: Baseline Financial Events</i>					
Delinquency	0.148	0.413	0.681	0.675	0.962
Collection	0.137	0.296	0.460	0.467	0.897
Charge-off	0.065	0.188	0.308	0.310	0.630
Bankruptcy	0.010	0.007	0.046	0.048	0.318
Judgment	0.009	0.034	0.067	0.060	0.403
Foreclosure	0.003	0.010	0.055	0.048	0.632
Lien	0.004	0.011	0.021	0.021	0.445
Repossession	0.003	0.012	0.022	0.020	0.491

# How to test these exclusion and monotonicity?

- Testing (and believing) monotonicity is also challenging

EXAMPLE 2 (Administrative Screening).<sup>5</sup> Suppose applicants for a social program are screened by two officials. The two officials are likely to have different admission rates, even if the stated admission criteria are identical. Since the identity of the official is probably immaterial to the response, it seems plausible that Condition 1 is satisfied. The instrument is binary so Condition 3 is trivially satisfied. However, Condition 2 requires that if official A accepts applicants with probability  $P(0)$ , and official B accepts people with probability  $P(1) > P(0)$ , official B must accept *any* applicant who would have been accepted by official A. This is unlikely to hold if admission is based on a number of criteria. Therefore, in this example we *cannot* use Theorem 1 to identify a local average treatment effect nonparametrically despite the presence of an instrument satisfying Condition 1.

# How to test these exclusion and monotonicity?

- Testing (and believing) monotonicity is also challenging
- Kitagawa (2015) and Frandsen et al. (2019) discuss ways to test this for binary outcomes
- Current limitation is finite sample approximation
- However, provide useful intuition for tests

Judging Judge Fixed Effects  
Brigham R. Frandsen, Lars J. Lefgren, and Emily C. Leslie  
NBER Working Paper No. 25528  
February 2019  
JEL No. C26,K14

## 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.

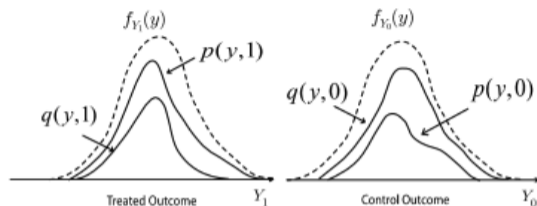
Brigham R. Frandsen  
Department of Economics  
Brigham Young University  
Provo, UT 84602  
frandsen@byu.edu

Emily C. Leslie  
Department of Economics  
Brigham Young University  
435 Crabtree Technology Building  
Provo, UT 84602  
emily.c.leslie@gmail.com

Lars J. Lefgren  
Department of Economics  
Brigham Young University  
130 Faculty Office Building  
Provo, UT 84602  
and NBER  
lars\_lefgren@byu.edu

## Kitagawa (2015) result

- Consider binary endogeneous treatment  $D$  and a discrete instrument  $Z$ .
- Joint *testable* assumption: instrument is valid and monotonicity. Why?
  - Consider
$$p(y, d) = \Pr(Y = y, D = d | Z = 1),$$
$$q(y, d) = \Pr(Y = y, D = d | Z = 0)$$
  - Let  $P$  and  $B$  be the probability over sets
  - Imbens and Rubin (1997) show
$$P(B, 1) - Q(B, 1) = \Pr(Y_1 \in B, D_1 > D_0)$$
$$P(B, 0) - Q(B, 0) = \Pr(Y_0 \in B, D_1 > D_0)$$
- Testable implication:
$$P(B, 1) - Q(B, 1) \geq 0$$
$$P(B, 0) - Q(B, 0) \geq 0$$



## Kitagawa (2015) result

- Consider binary endogeneous treatment  $D$  and a discrete instrument  $Z$ .
- Joint *testable* assumption: instrument is valid and monotonicity. Why?

- Consider

$$p(y, d) = \Pr(Y = y, D = d | Z = 1),$$

$$q(y, d) = \Pr(Y = y, D = d | Z = 0)$$

- Let  $P$  and  $B$  be the probability over sets
- Imbens and Rubin (1997) show

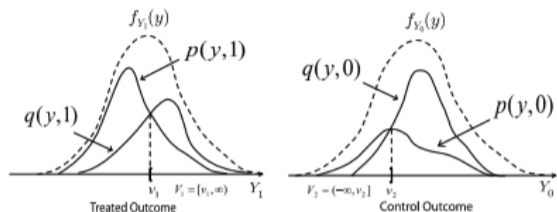
$$P(B, 1) - Q(B, 1) = \Pr(Y_1 \in B, D_1 > D_0)$$

$$P(B, 0) - Q(B, 0) = \Pr(Y_0 \in B, D_1 > D_0)$$

- Testable implication:

$$P(B, 1) - Q(B, 1) \geq 0$$

$$P(B, 0) - Q(B, 0) \geq 0$$



## Extension to Examiner designs (Frandsen et al.)

- Now, with multiple judges, we don't know the “true”  $z$  values
- Need to instead consider how this moves with noise
- Mapping from judges to “leniency” and then use Kitagawa test
- Package available: `testjfe` in Stata



## Brief aside on inference

- The many instruments field is still working on heteroskedasticity and complicated inference settings
- This stuff is challenging!
- However, given that the random assignment is effectively a set of judges to a given individual, robust standard errors seems most appropriate
- However, the norm (requested by referees, etc.) appears to be judge clustering
  - Not totally clear why

## Further Readings

- Cunningham (2021) chapter 7.8.2
- Judging Judge Fixed Effects (2020)
- Mueller-Smith (2015)
- Dobbie and Song (2015)