# Examiner Designs

Paul Goldsmith-Pinkham
Yale SOM & NBER

Peter Hull
Brown University & NBER

Michal Kolesar
Princeton University

June 2022

# 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 (e.g. Dobbie, Goldin and Yang (2018))
  - banktrupcy (Dobbie et al. (2017))
  - getting a loan (Liberman et al. (2021))
  - parole (Green and Winik (2010))
  - disability insurance (French and Song (2014))
  - patent rights (Galasso and Schankerman (2015))
  - cancer screening ( Chan et al. (2022))

- In many cases, this examiner is randomly assigned conditional on a large set of controls, *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 (e.g. Dobbie, Goldin and Yang (2018))
  - banktrupcy (Dobbie et al. (2017))
  - getting a loan (Liberman et al. (2021))
  - parole (Green and Winik (2010))
  - disability insurance (French and Song (2014))
  - patent rights (Galasso and Schankerman (2015))
  - cancer screening ( Chan et al. (2022))

- In many cases, this examiner is randomly assigned conditional on a large set of controls, *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

# Outline of today's talk

1. Describe examiner design
   - leniency vs. 2SLS/JIVE
2. Consider design-based approach with no controls
   - Conditions for LATE style interpretation
   - Implicatinos for inference
3. Discuss challenges for examiner designs
   - Monotonicity + Exclusion

# Basic notation and set ideas

- We have a sample of $n$ units indexed by $i$
  - Random assignment of an examiner $Q_i \in \{0, \ldots, K\}$
  - Potential treatment $D_i(q)$ (assume binary for today)
  - Potential outcome $Y_i(d, q)$
  - A vector of covariates $W_i$ with finite support $\mathcal{W}$
- We *observe* $(Q_i, W_i, Y_i, D_i)$
- Random assignment of $Q_i$ is typically conditional on $W_i$

# Basic notation and set ideas

- In the context of Dobbie, Goldsmith-Pinkham and Yang (2017)
  - Random assignment of a **bankruptcy judge** $Q_i$
  - Potential treatment of **bankruptcy discharge** $D_i$
  - Potential outcome of **credit score** $Y_i$
  - A vector of covariates $W_i$ of **bankruptcy court-year fixed effects** $W_i$

- Random assignment of judge is considered random *within* court-year
  - For now, ignore covariates and assume $W_i = 1$

# Typical approach in literature: leniency

- Typical approach (including in DGY) is to instrument treatment $D_i$ using 2SLS with a measure of "leniency" of assigned judge $\pi_Q = E(D_i|Q_i)$:

$$\hat{L}_i = \frac{1}{|Q_{i'} = Q_i|} \sum_{Q_{i'}=Q_i} D_i \tag{1}$$

$$\hat{L}_i^{lo} = \frac{1}{|i' \neq i, Q_{i'} = Q_i|} \sum_{i' \neq i, Q_{i'}=Q_i} D_i \tag{2}$$

- This leniency instrument is then used in a just-identified 2SLS setup:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \tag{3}$$

$$D_i = \gamma_0 + \gamma_1 \hat{L}_i + \epsilon_i \tag{4}$$

- Note that the leave-out leniency is the commonly used approach

# A "folk theorem" of examiner designs: leniency = 2SLS/JIVE

- Consider the alternative estimation approach using examiner dummies

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \tag{5}$$

$$D_i = \tilde{\gamma}_0 + \sum_{k=1}^{K} \tilde{\gamma}_{1k} 1(Q_i = k) + \epsilon_i \tag{6}$$

- Point estimates from this are *identical* to using $\hat{L}_i$ in a just-identified 2SLS estimator:

$$\hat{\beta}_1 = \frac{D' \underbrace{Q(Q'Q)^{-1}Q'}_{P_Q} Y}{D' \underbrace{Q(Q'Q)^{-1}Q'}_{P_Q} D} = \frac{\hat{D}'Y}{\hat{D}'D}, \tag{7}$$

where $\hat{D}_i = \frac{1}{|Q_{i'}=Q_i|} \sum_{Q_{i'}=Q_i} D_i = \hat{L}_i$,

- first stage fitted values are the group means!

# A “folk theorem” of examiner designs: leniency = 2SLS/JIVE

- Additionally, the just-identified 2SLS approach using $\hat{L}_i^{lo}$ is numerically identical to using jackknife IV with $Q_i$ (Angrist et al. 1999)

- JIVE is designed to solve many-weak instrument bias – many judges can cause “overfitting”, but leaving out own observation can help avoid issue
  - Implication: Leniency design with leave-out also solves many-weak instrument problem

- However,
  1. Leniency design vs. examiner fixed effects are not using different empirical strategies, but different estimation approaches
  2. the single instrument $L_i$ is estimated with error from many instruments, and using leniency shrouds this fact; estimating the underlying first stage is noisy!
  3. JIVE is insufficient to solve bias once you introduce many covariates – requires other estimators such as UJIVE (Kolesar (2013))

# Outstanding questions in examiner designs

- Despite the fact that there appears to be random assignment, clarity is missing on a number of questions:
  1. Some papers cluster on location; others two-way cluster on location and examiner. What is right?
  2. If constructing leniency, what is the right dimension to be aggregating over?
  3. How do the monotonicity and exclusion assumptions generalize to the examiner design setting?

- A design-based framework sheds light on these questions

# Design-based approach using oracle estimator (no controls)

- Consider a design-based approach to thinking about estimation. We *observe* $(Q_i, W_i, Y_i, D_i)$, and assume that the potential outcomes and controls are fixed, while $Q_i$ is random.
  - Let $W_i = 1$ (no controls)
  - Let $E_n$ denote the sample average

- We focus on a benchmark case of an *oracle* estimator that uses a single instrument

$$D_i^*(Q_i) = \sum_{k=0}^{K} 1\{Q_i = k\} d_k^*, \tag{8}$$

where $d_k^*$ is a demeaned single instrument value for each examiner value $k$.

  - E.g. in this no controls case, this is the demeaned first stage predicted value using known coefficients

# Design-based approach using oracle estimator (no controls)

- This gives a Wald estimator of

$$\tau^* = \frac{E_n(D_i^* Y_i)}{E_n(D_i^* D_i)}, \tag{9}$$

  which under regularity conditions converges to

$$\tau = \frac{E_n E(D_i^* Y_i)}{E_n E(D_i^* D_i)}. \tag{10}$$

- So far, we've only assumed random assignment of $Q_i$. Two important questions:
  1. What can we say about inference about $\hat{\tau}^* - \tau$?
  2. What is the causal interpretation of $\tau$? For this, will need more assumptions
     - exclusion restriction ($Y_i(d, q) = Y_i(d)$)
     - monotonicity (exists an ordering of the judges' strictness that holds for all individuals)

# Implications for inference

- Under reasonable regularity assumptions, can show that

$$\left( \frac{(\sum_i E[D_i^* D_i])^2}{\sum_i \text{var}(D_i^*(Y_i - D_i\tau))} \right)^{1/2} (\hat{\tau}^* - \tau) \xrightarrow{d} \mathcal{N}(0,1). \tag{11}$$

- This implies that Eicker-Huber-White (EHW) s.e. are slightly conservative (and exact under constant TE)
  - Analogous to standard ATE inference results under RCTs

- More importantly, clustering is not appropriate! There is no reason to cluster on judge
  - Analogy to an RCT - you would not cluster on each treatment arm
  - This result can allow the number of examiners to grow with sample size as well

# Exclusion and monotonicity restrictions

- It is briefly worth considering the economic content of these two assumptions
  - Exclusion restriction assumes that bankruptcy judge assignment *only* affects the outcome through the decision to discharge debt
  - Monotonicity assumes that judges have a common ordering of bankruptcy filers for who gets discharge

# Exclusion and monotonicity restrictions

- It is briefly worth considering the economic content of these two assumptions
  - Exclusion restriction assumes that bankruptcy judge assignment *only* affects the outcome through the decision to discharge debt
  - Monotonicity assumes that judges have a common ordering of bankruptcy filers for who gets discharge

- Several recent papers attempting to test and weaken these assumptions:
  - Frandsen et al. (2019) formalize a weaker "average monotonicity" condition
    - Also propose test of monotonicity + exclusion
  - Papers also parametrize variation in examiner differences and estimate jointly with treatment effects (Chan et al. (2021), Arnold et al. (2021))
  - Directly model violations of exclusion restricton
    - e.g. Kline and Walters (2016) and Kirkeboen et al. (2016)
    - Challenge: IV with multiple treatments is difficult to interpret as LATE!
  - Potentially weaken exclusion restriction to "on average" (Kolesár et al. (2015), Angrist et al. (2021))
    - Exclusion restrictions are uncorrealted with leniency variation
    - Need many examiners

# Exclusion and monotonicity restrictions

We will assume monotonicity and exclusion, but these are strong empirical assumptions in examiner settings (as flagged in Imbens and Angrist (1994)

EXAMPLE 2 (Administrative Screening):[5] 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.

# Interpretation of estimand under monotonicity + exclusion

$$\tau = \frac{E_n E(D_i^* Y_i)}{E_n E(D_i^* D_i)}. \tag{12}$$

- Formally, monotonicity is defined as the existence of an ordering $\{\pi(q)\}$ such that $D_i(\pi^{-1}(q)) \geq D_i(\pi^{-1}(q-1))$.
  - If $\pi(k) = l$, the $k$ is $l+1$ least strict judge

- With both monotonicity and exclusion, the estimand is the weighted combination of treatment effects for $K$ different complier groups.

- These weights are positive if $D_i^*$ is increasing in $\pi(k)$
  - This is satisfied with no controls under the oracle because the first stage prediction, $D_i^*$, correctly identifies the right ordering amongst judges

# Generalizing setup to include controls

- Let assignment probabilities depending on fixed controls $W_i$:

$$p_k(W_i) = Pr(Q_i = k|W_i)$$

- Controls typically matter in examiner designs, and there can be many of them
  - If judges are randomly assigned within an office within a given year or month, this may lead to thousands of strata
  - In Dobbie et al. (2017), there are 1,477 fixed effects for 173k observations

- Hence, important to think of asymptotics and empirical settings as having many controls and many instruments

- *Knowledge or assumption on $p_k(W_i)$ is key*
  - This choice is where the decision over time period to aggregate over is made

# The oracle instrument with controls

- The setup for our oracle instrument now matters quite a bit
  - We consider our oracle estimator using instrument

$$D_i^* = \sum_{k=0}^{K} 1\{Q_i = k\} d_k^*(W_i), \tag{13}$$

  where we assume $D_i^*$ is demeaned wrt $W_i$

- Now, we require monotonicity to hold *conditional* on $W_i$ (sometimes referred to as weak monotonicity (Słoczyński (2022))):

$$D_i(\pi_w^{-1}(k)) \geq D_i(\pi_w^{-1}(k-1))$$

  - Recall that we need $d_k^*(w)$ to increase in $\pi_w(k)$

- If $d_k^*(w)$ is incorrectly specified, this condition may not hold
  - This is even if monotonicity holds for $\{\pi_w\}$!

# Two concrete ways in which the oracle could be misspecified

- First, misspecification of the necessary controls for identification.
  - Concretely: assume judges are randomly assigned each office-month
  - Then the appropriate control is an office-month fixed effect, not controlling for office fixed effects and year-month fixed effects
  - This issue is discussed in Blandhol et al. (2021)

- Second, contamination bias can occur in $d^*(w)$ if the model does not consider a fully interacted model (Goldsmith-Pinkham, Hull and Kolesár (2022))
  - This implies needing to interact $Q_i$ with $W_i$ as a set of controls
  - This is an issue even in cases when the true ordering $\{\pi_k(w)\}$ does not depend on $w$ (strong monotonicity)!

- Full interactions significantly increases the number of necessary controls, which suggests important estimation challenges

# Inference with controls

- The previous result on inference holds even in the case of controls
- This suggests, again, that clustering is inappropriate, even with many offices
  - Potential caveats about the correlation of the assignment of treatments
- Still a work in progress to implement inference for non-oracle estimators

# Estimation solutions (UJIVE)

- Kolesár (2013) shows that many weak instruments with many controls can suffer from additional biases even with JIVE.
  - Proposes UJIVE solution which adjusts for leave-out means of controls
- Note: some “leave-out” leniency approaches do this! (e.g. Dobbie et al. (2017))
  - However, with the need to “fully interact” to solve contamination bias, UJIVE is a more functional approach
- We are working to provide a Stata and R package that does UJIVE effectively
  - Provide you with a beta testing opportunity!

# UJIVE package (work in progress)

`version 0.5.0 14Nov2021` | Installation | Usage | Examples | Compiling

## Installation

From the command line:

```
git clone git@github.com:mcaceresb/stata-manyiv
```

(or download the code manually and unzip). From Stata:

```
cap noi net uninstall manyiv
net install manyiv, from(`c(pwd)'/stata-manyiv)
```

(Change `stata-manyiv` if you download the package to a different folder; e.g. `stata-manyiv-main`.) Note if the repo were public, this could be installed directly from Stata:

```
local github "https://raw.githubusercontent.com"
net install manyiv, from(`github'/mcaceresb/stata-manyiv/master/)
```

## Usage

```
manyiv depvar (endogenous = instrument) [exogenous], options
help manyiv
```

# Checklist and future direction

- Examiner design is a many (potentially weak) instrument setting,
  - leniency measures are a convenient (but potentially problematic) approach
- Work so far suggests that robust SE are the appropriate approach to inference
  - At odds with norm to double cluster standard errors
- Accounting for contamination bias in the first stage (in order to correctly satisfy monotonicity) requires interacting instruments with controls
- This introduces many weak instrument + many control problems
  - One solution is UJIVE
  - Stata package here: https://github.com/gphk-metrics/stata-manyiv