Testing Mechanisms

Soonwoo Kwon Jonathan Roth

Brown University

August 29, 2024

Motivation

- Over the last few decades, a lot of progress has been made on more-credibly estimating causal effects of a treatment D on an outcome Y using (quasi-)experimental variation
- Once the effect of D on Y is established, the natural next question is why?
  I.e. what are the mechanisms?

Motivating Example: Bursztyn et al (2020)

- Bursztyn, Gonzalez, Yanagizawa-Drott (2020 AER) conduct an RCT in Saudi Arabia focused on women's economic outcomes.

- They provide descriptive evidence that men under-estimate how open other men are about women working outside the home. They randomly assign the treated group information about other men's opinions.

- At the end of the experiment, men can sign their wives up for a job-finding service or take a gift card.

- Bursztyn et al. (2020) find that the treatment has a positive effect on both enrollment in the job-search service and longer-run economic outcomes for women (e.g. apply/interview for jobs)

Motivating Example: Bursztyn et al (2020)

- Key Q: are the long-run effects the mechanical impact of the job-search service, or do they also reflect longer-run changes in attitudes?

C. *Interpreting the Results*

Understanding the Longer-Term Effects.—It is difficult to separate the extent to which the longer-term effects are driven by the higher rate of access to the job service versus a persistent change in perceptions of the stigma associated with WWOH.

- Bursztyn et al. (2020) are unsure, but speculate that there may be non-mechanical effects based on longer-run follow-ups about men's beliefs

Existing approaches for examining mechanisms

- Formal methods (many from biostats and polisci) exist for estimating how much of the treatment effect is explained by a mediator M. Typically,
  - Estimate effect of D on M
  - Estimate effect of M on Y (conditional on D)
  - Multiply these effects to obtain the average "indirect effect" of D on Y through M
- However, these typically require strong assumptions to identify the effect of M on Y
  - M is randomly assigned conditional on D and observable characteristics (e.g. Imai et al., 2010; Huber, 2014; Acharya et al., 2016; Huber et al., 2017)
  - Alternative approaches using DID or IV (e.g. Frölich and Huber, 2017; Deuchert et al., 2019; Schenk, 2023)
- These tools are rarely used in empirical economics. Instead, testing of mechanisms is typically done more informally
  - Examine effects of D on intermediate outcomes
  - Heterogeneity analysis: do groups with larger effects of D on M have larger effects on Y?

This paper

- Goal: can we say something formal about mechanisms while avoiding strong assumptions needed to identify the effect of M on Y?
- We make progress on this by trying to answer an easier (but hopefully still informative) Q
- Instead of estimating the average indirect effect, we consider what we call the sharp null hypothesis of full mediation:

  Can the effect of D on Y be explained fully by a candidate mechanism (or set of mechanisms) M?

- If we reject the null, then we have learned that other mechanisms must also matter (at least for some people)
  - And we provide tools for lower-bounding the magnitude of the other mechanisms

First observation

- Suppose we want to evaluate the sharp null that the effect of D on Y operates only through a candidate mechanism M
  - E.g. in Bursztyn et al, does the effect operate entirely through job service signup?
- Assume that D is (as good as) randomly assigned, and has a monotone effect on M
  - In Bursztyn et al, random assignment is by design
  - Monotonicity says that learning about others' beliefs only increases job service signup
- Observe that under the sharp null, D is a valid instrumental variable for the LATE of M on Y
- But the IV model is known to have testable implications!

  (Balke and Pearl, 1997; Kitagawa, 2015; Huber and Mellace, 2015; Mourifié and Wan, 2017)

Overview

- First key observation:
  IV testing methods can be used "off-the-shelf" to test sharp null of full mediation with binary D, M and a monotonicity assumption

- Extend this insight to develop sharp testable implications:

  - When M can be non-binary and/or multi-dimensional (w/finite support)

  - Under relaxations of the monotonicity assumption

  $\hookrightarrow$ Results imply sharp testable implications for IV settings with multi-valued treatment, which may be of indep. interest (building on non-sharp implications in Sun, 2023)

- We also show how one can quantify the magnitude of alternative mechanisms when the sharp null is violated
  - Lower bounds on the fraction of "always-takers" and "never-takers" who are affected by treatment despite having no effect on M, as well as the average effect for such ATs/NTs

Set-up

- Binary treatment of interest D
- Potential mediator $M(d)$ with finite support
  - Dimension is irrelevant, but finite support is crucial
- Potential outcomes $Y(d, m)$
- We observe $(Y, M, D) = (Y(D, M(D)), M(D), D)$
- Assume throughout that D is as good as randomly assigned: $D \perp\!\!\!\perp (Y(\cdot, \cdot), M(\cdot))$ and $0 < P(D = 1) < 1$.
  - All identification arguments go through if assignment is random conditional on X

Type shares

- Let M have support $\{m_0, ..., m_{K-1}\}$
  - Write $G = lk$ to denote the event that $M(0) = m_l$ and $M(1) = m_k$
  - Refer to individuals with $G = kk$ as *k-always takers* and $G = lk$ as *lk-compliers*
  - Denote by $\theta_{lk} := P(G = lk)$ the share of group lk

- Allow for arbitrary restrictions on the type shares:
  $\theta \in R \subseteq \Delta$, where Δ is the K^2-dimensional simplex
  - Full monotonicity: $R = \{\theta : \theta_{lk} = 0 \text{ if } m_l > m_k\}$
  - Bounded share of defiers: $R = \{\theta : \sum_{l,k:m_l > m_k} \theta_{lk} \leq d\}$
  - Elementwise monotonicity: can impose that $M(d)$ is elementwise increasing in d by setting $R = \{\theta : \theta_{lk} = 0 \text{ if } m_l \not\preceq m_k\}$ for $\preceq$ the element-wise partial order
  - No restrictions: $R = \Delta$

Sharp null of full mediation

- We say the sharp null of full mediation is satisfied if

$$Y(d, m) = Y(m) \text{ (a.s.) for all } d, m$$

- If the sharp null is satisfied, then M is the only mechanism that matters
- If the data is inconsistent with the sharp null, then we have evidence that mechanisms other than M matter (for at least some people)
- E.g.: in motivating example, if reject the sharp null, we can conclude that the information treatment changes behavior through channels other than job service sign-up

Deriving testable implications

- We define ν_k to be fraction of k-ATs who are affected by the treatment,

$$\nu_k := P(Y(1,k) \neq Y(0,k) \mid G = kk)$$

- In other words, ν_k is the fraction of k-always takers for whom there is a direct effect of the treatment
- Note that the sharp null implies that $\nu_k = 0$ for all k
- In what follows, we will derive lower-bounds on the ν_k: the sharp null is violated if any of the lower-bounds are non-zero

Derivation of the lower bounds on the ν_k

- For any Borel set A, we have

$$P(Y \in A, M = k \mid D = 0) =$$

$$\underbrace{\theta_{kk} \cdot P(Y(0,k) \in A \mid G = kk)}_{\text{Probability for } k\text{-ATs}} + \sum_{l: l \neq k} \underbrace{\theta_{kl} P(Y(0,k) \in A \mid G = kl)}_{\text{Sum of prob.s for } kl\text{-compliers}}.$$

Derivation of the lower bounds on the ν_k

- For any Borel set A, we have

$$P(Y \in A, M = k \mid D = 0) =$$
$$\underbrace{\theta_{kk} \cdot P(Y(0,k) \in A \mid G = kk)}_{\text{Probability for } k\text{-ATs}} + \sum_{l:l \neq k} \underbrace{\theta_{kl} P(Y(0,k) \in A \mid G = kl)}_{\text{Sum of prob.s for } kl\text{-compliers}}.$$

- Likewise,

$$P(Y \in A, M = k \mid D = 1) =$$
$$\underbrace{\theta_{kk} \cdot P(Y(1,k) \in A \mid G = kk)}_{\text{Probability for } k\text{-ATs}} + \sum_{l:l \neq k} \underbrace{\theta_{lk} P(Y(1,k) \in A \mid G = lk)}_{\text{Sum of prob.s for } lk\text{-compliers}}.$$

- Can use these equations to solve for $P(Y(1,k) \in A \mid G = kk) - P(Y(0,k) \in A \mid G = kk)$
  Taking a sup over A yields TV distance btwn $Y(1,k)$ & $Y(0,k)$ for k-ATs

TV Bounds

- Using the fact that probs are between 0 and 1, it follows that

$$\underbrace{\theta_{kk}}_{\text{Share of } k\text{-ATs}} \quad TV_k \geq \sup_A \Delta_k(A) - \sum_{l:l \neq k} \underbrace{\theta_{lk}}_{\text{Share of } lk\text{-Cs}}$$

  where
  1) TV_k is the TV distance between $Y(1,k) \mid G = kk$ and $Y(0,k) \mid G = kk$, and
  2) $\sup_A \Delta_k(A) = \sup_A [P(Y \in A, M = k \mid D = 1) - P(Y \in A, M = k \mid D = 0)]$ measures the distance between $Y, M = k \mid D = 1$ and $Y, M = k \mid D = 0$.

- However, as shown in Borusyak (2015), $\nu_k \geq TV_k$

- Replacing ν_k with TV_k in the previous display yields our lower-bound

$$\theta_{kk}\nu_k \geq \sup_A \Delta_k(A) - \sum_{l:l \neq k} \theta_{lk}$$

Unknown θ

- The bounds above on ν_k involved the complier/AT shares θ_{lk}
- With binary M & monotonicity, these shares are point-identified. In general, θ is only partially identified. Why?
- However, the identified set Θ_I for θ is characterized by linear inequalities when R is a polyhedron. Details
    - Intuitively, sum of θs must match the marginal distributions of $M \mid D$, and $\theta \in R$
- It is thus straightforward to compute sharp bounds on ν_k that optimize over the identified set for θ via linear programming (LP)
    - Lower bound on ν_k corresponds to minimum value of θ_{kk} in the ID set.
    - If M is fully-ordered & impose monotonicity, the resulting bounds have a closed-form solution

Formal bounds on ν_k with partially identified θ

Proposition: Let θ_{kk}^{min} be the minimum value of θ_{kk} among $\theta \in \Theta_I$. Assume $\theta_{kk}^{min} > 0$. Then

$$\nu_k \geq \max\left\{ \frac{1}{\theta_{kk}^{min}} \left[\sup_A \Delta_k(A) - P(M = k \mid D = 1) + \theta_{kk}^{min} \right], 0 \right\}.$$

Moreover, this bound is **sharp**: there exists a distribution of POs consistent with the observed data such that the bound holds with equality.

Testable Implications of the Sharp Null

- We have the bound

$$\theta_{kk}\nu_k \geq \sup_A \Delta_k(A) - \sum_{l:l\neq k} \theta_{lk}$$

- Therefore, if the sharp null is satisfied, then there exists some $\theta \in \Theta_I$ such that for all k,

$$\sup_A \Delta_k(A) \leq \sum_{l:l\neq k} \theta_{lk}$$

- When R is a polyhedron, the ID set is defined by linear inequalities, and so this is equivalent to checking whether an LP is feasible

- These testable implications are sharp!
  - Equivalent to the implications of Kitagawa (2015) when M is binary

Inference

- This testing problem is non-standard for mainly two reasons:
- The bounds involve quantities of the form

$$\sup_A \Delta_k(A) = \int_{\mathcal{Y}} \left(f_{Y,M=1|D=1} - f_{Y,M=1|D=0}\right)_+$$

which are potentially non-differentiable in the underlying distributions in the DGP

- With multi-valued and/or non-monotone M, the bounds involve the solution to a linear program, which are also potentially non-differentiable in the underlying DGP

One solution - moment inequalities Details

- When Y is discrete, the implications of the sharp null of full mediation can be written as a system of moment inequalities with linear nuisance parameters
  - If Y is continuous, discretizing preserves the validity of the test but at the potential loss of sharpness

- The nuisance parameters correspond to compliers shares θ and positive differences between partial densities

$$\delta_{qk} = (P(Y = q, M = k \mid D = 1) - P(Y = q, M = k \mid D = 0))_+$$

- Tractable tests for moment inequalities with linear nuisance parameters have been developed recently by Fang et al. (2023); Andrews et al. (2023); Cox and Shi (2022); Cho and Russell (2024)
  - Tentatively recommend Cox and Shi (2022) based on simulations

Bounds on Average Direct Effects

Back

- In addition to bounds for the fraction of ATs affected, we can also bound the average direct effect on k-ATs

$$ADE_k = E[Y(1,k) - Y(0,k) \mid G = kk]$$

- Intuition: the distribution of $Y \mid M = k, D = 1$ is a mixture of the $Y(1,k)$ potential outcome for k-always takers and lk-compliers with weight proportional to θ_{kk} on the ATs.

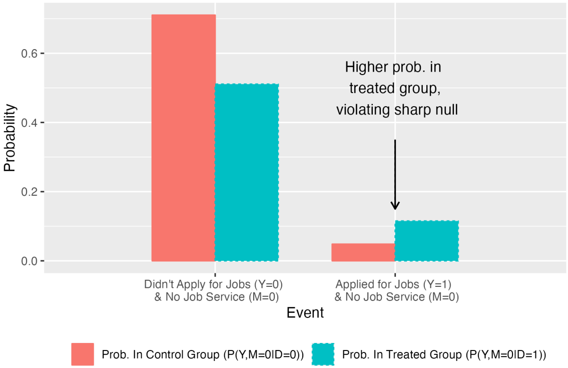
- The lowest/highest possible values of $E[Y(1,k) \mid G = kk]$ correspond to the means of the least/most-favorable subdistributions of $Y \mid M = k, D = 1$

- In the special case of binary M, the bounds on treatment effects for ATs correspond to Lee (2009) bounds treating M as the sample selection
  - This was observed by Flores and Flores-Lagunes (2010) for binary M

Example 1: Bursztyn et al

- In Bursztyn et al. (2020), we would like to know whether the effect of treatment D on economic outcomes Y is explained by increase in job service signup M?
  - If not, we can conclude that the information treatment has some economic impact through changes in behavior other than job service sign-up
- The inequalities we derived above imply that

$$
\begin{aligned}
&P(\text{apply for job \& don't use job service} \mid \text{control}) \geq \\
&P(\text{apply for job \& don't use job service} \mid \text{treated})
\end{aligned}
$$



- We see that $\hat{P}(Y = 1, M = 0 \mid D = 1) > \hat{P}(Y = 1, M = 0 \mid D = 0)$, contradicting the sharp null. Reject the sharp null at the 5% level.
- Thus, some NTs who never enroll in the job service are affected by treatment – evidence the treatment effect on LR outcomes is not purely thru the job service!
- Bounds on ν_k suggest at least 11% of NTs are affected by treatment (ADE: $[0.11, 0.18]$)

Example 2: Baranov et al (2020, AER)

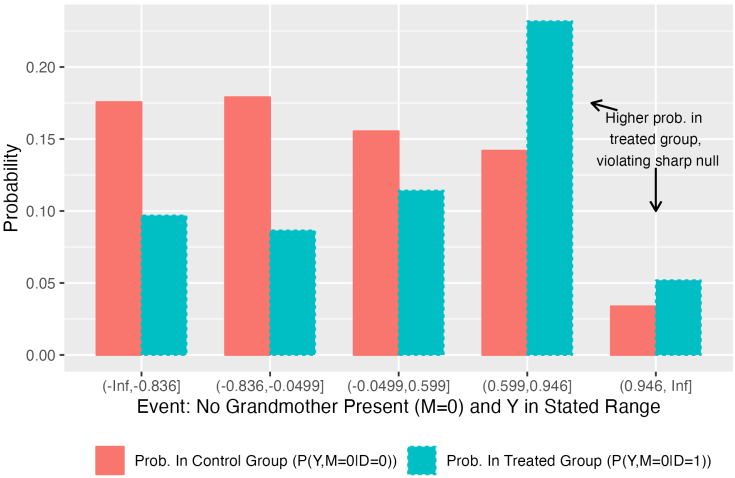
- Baranov et al. (2020) study an RCT that randomized access to cognitive behavioral therapy (CBT) for depression for mothers in Pakistan
- They find that CBT substantially reduces rates of depression, and increases mother's financial empowerment (e.g. work outside the home, control over finances)
- They would like to know the mechanisms by which CBT affects financial empowerment.
- To explore mechanisms, they look for impacts on a variety of intermediate outcomes

- Two intermediate outcomes for which they find an effect are presence of a grandmother giving help and relationship quality with the husband

> These results suggest that improved social support within the household, either through a better relationship with the husband or asking grandmothers for help, might be a mechanism underlying the effectiveness of this CBT intervention.

- Using our tools, we can test whether these intermediate outcomes can fully explain the effect, or whether there must be other mechanisms at play, too

Help from Grandmother



Point estimates suggest at least 16 percent of never-takers who never get help from grandma are affected by treatment (under monotonicity)

We reject the sharp null at $p = 0.02$ (CS)

LB positive allowing up to 11 percent defiers

Relationship Quality

- We can likewise test whether the effect is explained through relationship quality, which is measured on a 1-5 scale
- Using our results on multi-valued M, we estimate that 10% of all ATs are affected by treatment under monotonicity (pooling across different values of M)
- Tests of the sharp null significant using CS ($p = 0.03$)
- However, the test using $M = c(\text{grandmother}, \text{relationship quality})$ yield a p-value of 0.65.
  - Can't reject that these two mechanisms together explain the effect

To do list

- Incorporating additional restrictions to sharpen testable implications
  - In some settings, may be reasonable to impose monotonicity or smoothness of $Y(d, m)$ in m
  - May sometimes be reasonable to impose stochastic dominance relationships between compliers and ATs
  - Incorporate restrictions that allowing for testing w continuous M (a la D'Haultfœuille et al., 2021)
- Extension to non-experimental settings (e.g. IV)
  - Note that if Z is a valid instrument and D affects Y only thru M, then Z affects Y only through M
  - So can use the tools developed replacing D with Z
  - Conjecture that this is sharp

Acharya, Avidit, Matthew Blackwell, and Maya Sen, "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects," *American Political Science Review*, August 2016, *110* (3), 512–529.

Andrews, Isaiah, Jonathan Roth, and Ariel Pakes, "Inference for Linear Conditional Moment Inequalities," *The Review of Economic Studies*, January 2023, p. rdad004.

Balke, Alexander and Judea Pearl, "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, September 1997, *92* (439), 1171–1176. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.1997.10474074.

Baranov, Victoria, Sonia Bhalotra, Pietro Biroli, and Joanna Maselko, "Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial," *American Economic Review*, March 2020, *110* (3), 824–859.

Borusyak, Kirill, "Bounding the Population Shares Affected by Treatments," March 2015.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott, "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia," *American Economic Review*, October 2020, *110* (10), 2997–3029.

Cho, JoonHwan and Thomas M. Russell, “Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments,” *Journal of Business & Economic Statistics*, April 2024, *42* (2), 563–578.

Cox, Gregory and Xiaoxia Shi, “Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models,” *The Review of Economic Studies*, March 2022, p. rdac015.

Deuchert, Eva, Martin Huber, and Mark Schelker, “Direct and Indirect Effects Based on Difference-in-Differences With an Application to Political Preferences Following the Vietnam Draft Lottery,” *Journal of Business & Economic Statistics*, October 2019, *37* (4), 710–720.

D’Haultfœuille, Xavier, Stefan Hoderlein, and Yuya Sasaki, “Testing and relaxing the exclusion restriction in the control function approach,” *Journal of Econometrics*, 2021.

Fang, Zheng, Andres Santos, Azeem M. Shaikh, and Alexander Torgovitsky, “Inference for Large-Scale Linear Systems With Known Coefficients,” *Econometrica*, 2023, *91* (1), 299–327. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18979.

Flores, Carlos and Alfonso Flores-Lagunes, “Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects,” Working paper January 2010.

Frölich, Markus and Martin Huber, "Direct and Indirect Treatment Effects–Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, November 2017, *79* (5), 1645–1666.

Gunsilius, F F, "Nontestability of instrument validity under continuous treatments," *Biometrika*, December 2021, *108* (4), 989–995.

Huber, Martin, "Identifying Causal Mechanisms (primarily) Based on Inverse Probability Weighting," *Journal of Applied Econometrics*, 2014, *29* (6), 920–943. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.2341.

_ **and Giovanni Mellace**, "Testing Instrument Validity for Late Identification Based on Inequality Moment Constraints," *The Review of Economics and Statistics*, 2015, *97* (2), 398–411. Publisher: The MIT Press.

_ **, Michael Lechner, and Giovanni Mellace**, "Why Do Tougher Caseworkers Increase Employment? The Role of Program Assignment as a Causal Mechanism," *The Review of Economics and Statistics*, March 2017, *99* (1), 180–183.

Imai, Kosuke, Luke Keele, and Dustin Tingley, "A general approach to causal mediation analysis.," *Psychological Methods*, 2010, *15* (4), 309–334.

Kitagawa, Toru, "A Test for Instrument Validity," *Econometrica*, 2015, *83* (5), 2043–2063.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, July 2009, *76* (3), 1071–1102.

Mourifié, Ismael and Yuanyuan Wan, "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, May 2017, *99* (2), 305–313.

Schenk, Timo, "Mediation Analysis in Difference-in-Differences Designs," Technical Report 2023.

Sun, Zhenting, "Instrument validity for heterogeneous causal effects," *Journal of Econometrics*, 2023, *237* (2), 105523.

Identified set for θ Back

The shares θ are consistent with the observed data if they satisfy the following inequalities:

$$\sum_l \theta_{kl} = P(M = k \mid D = 0) \text{ for } k = 0, .., K-1 \qquad \text{(Match marginals for } D = 0\text{)}$$

$$\sum_l \theta_{lk} = P(M = k \mid D = 1) \text{ for } k = 0, .., K-1 \qquad \text{(Match marginals for } D = 1\text{)}$$

$$\theta_{kk'} = 0 \text{ for } k \npreceq k' \qquad \text{(Monotonicity)}$$

$$0 \leq \theta_{kk'} \leq 1 \text{ for all } k, k' \qquad \text{(Probabilities in unit interval)}$$

$$\theta \in R \qquad \text{(Additional restrictions)}$$

We denote by Θ_I the identified set for θ

Moment Inequalities Details Back

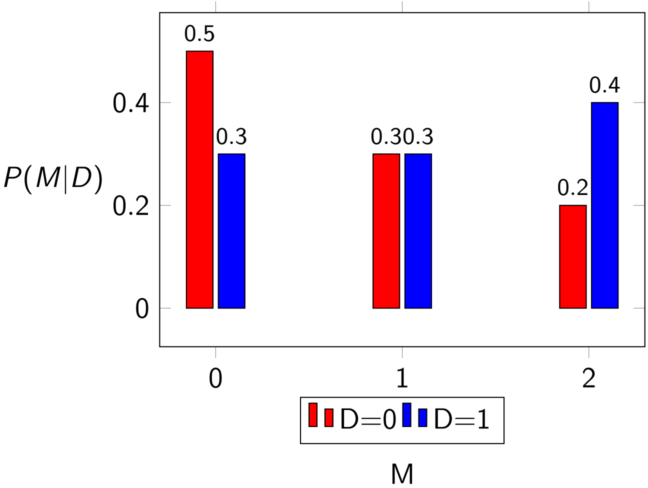
- Note that for discrete Y,

$$\sup_A \Delta_k(A) = \sum_q \max\{P(Y = q, M = k \mid D = 1) - P(Y = q, M = k \mid D = 0), 0\}$$

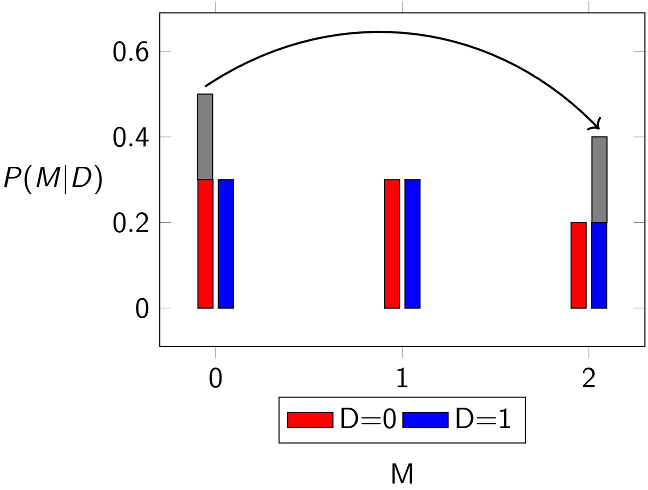
- Thus, $\sum_{l:l \neq k} \theta_{lk} \geq \sup_A \Delta_k(A)$ if and only if there exists δ_{qk} such that

$$\sum_{l:l \neq k} \theta_{lk} \geq \sum_q \delta_{qk}$$
$$\delta_{qk} \geq P(Y = q, M = k \mid D = 1) - P(Y = q, M = k \mid D = 0)$$
$$\delta_{qk} \geq 0$$

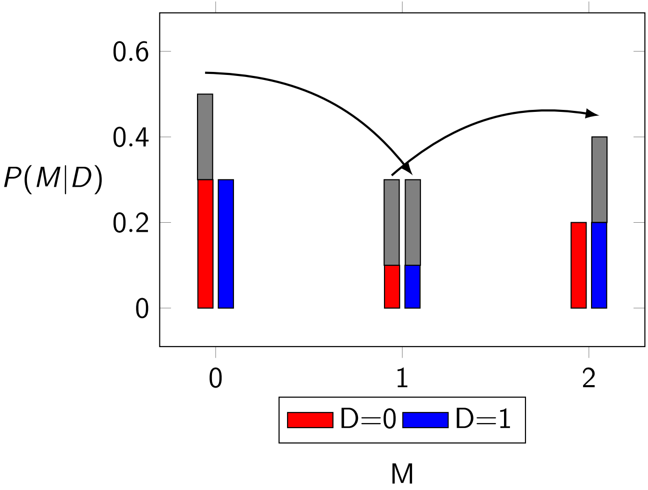
- We can thus test the sharp null by testing the moment inequalities above, along with the additional moments implied by the constraint that $\theta \in \Theta_I$ ID Set for θ



- Consider the following distributions of $M \mid D$. The $M \mid D = 0$ distribution has more mass at 0 and less mass at 2. Back



- This is consistent with $\theta_{02} = 0.2$ and $\theta_{01} = \theta_{12} = 0$. Back



- But it is also consistent with a cascade: $\theta_{01} = \theta_{12} = 0.2$, and $\theta_{02} = 0$. Back

Monte Carlo Design Back

- We conduct Monte Carlo simulations calibrated to our empirical applications
  - Bursztyn et al. (2020) with a binary M, Baranov et al. (2020) where M takes 5 values
- To evaluate size control, we draw (Y, M) for both treated and untreated units from the empirical distribution of control units in the data
  - This ensures null holds and all moments are binding
- To evaluate power, we draw $(Y, M)|D$ from the empirical distribution in the data. We also consider mixtures between this and the DGP above
- Sample sizes in simulations match those in the data:
  Bursztyn et al (284)
  Baranov et al (40 clusters, $\sim$ 600 obs)
  $\hookrightarrow$ also consider designs with 80, 200 clusters
- When outcome is discrete, consider discretizations based on 2,5,10 bins

Monte Carlo Results - Binary M (Bursztyn et al) Back

Panel A: Bursztyn et al

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.038	0.032	0.030	0.078	0.070
t=0.5	0.036	0.196	0.190	0.116	0.214	0.194
t=1	0.077	0.626	0.632	0.386	0.620	0.584

ARP = Andrews et al (2023); FSST = Fang et al (2023), cs = Cox & Shi (2022), K = Kitagawa (2015)

- All tests reasonably well-sized
- Power similar for ARP, CS, FSST; all better than K

Monte Carlo Results - Binary *M* (Baranov et al) Back

Panel B: Baranov et al, 40 clusters

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.056	0.154	0.050	0.232	0.212
t=0.5	0.134	0.194	0.206	0.064	0.314	0.270
t=1	0.283	0.570	0.668	0.422	0.750	0.680

Panel C: Baranov et al, 80 clusters

	$\bar{\nu}$ LB	ARP	CS	K	FSSTdd	FSSTndd
t=0	0	0.044	0.064	0.040	0.132	0.112
t=0.5	0.134	0.322	0.340	0.160	0.410	0.322
t=1	0.283	0.836	0.936	0.846	0.956	0.936

- Size control good for ARP, K; CS moderately over-sized with small # of clusters but OK w/80 clusters; FSST somewhat over-sized even w/80 clusters

Monte Carlo Results - Multivalued M (Baranov et al) Back

Panel A: Baranov et al, 40 clusters

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.052	0.088	0.274	0.178
t=0.5	0.119	0.066	0.228	0.438	0.374
t=1	0.255	0.166	0.754	0.864	0.828

Panel B: Baranov et al, 80 clusters

	$\bar{\nu}$ LB	ARP	CS	FSSTdd	FSSTndd
t=0	0	0.066	0.048	0.188	0.128
t=0.5	0.119	0.066	0.314	0.582	0.500
t=1	0.255	0.164	0.962	0.994	0.990

- CS and ARP reasonably well-sized, and in terms of power, $CS \gg ARP$
- FSST somewhat oversized (but good power)

Alternative tests in applications

Back - Bursztyn

Application	M	CS	ARP	FSSTdd	FSSTndd
Bursztyn et al (main sample)	Job-search Sign-up	0.020	0.030	0.018	0.018
Bursztyn et al (full sample)	Job-search Sign-up	0.019	0.020	0.019	0.019
Baranov et al	Grandmother	0.023	0.030	0.011	0.015
Baranov et al	Relationship	0.028	0.650	0.037	0.049
Baranov et al	Grandmother + Relationship	0.654	0.550	0.115	0.256

Table: *p*-values for tests for the sharp null using alternative procedures

Note on identification "power"

- The sharp null implies that there should be no effect of D on Y for k-ATs, for all k.
- If the data is consistent with there being no ATs for any k (i.e. everyone is a complier), then there are no testable implications of the sharp null!
- When M is fully-ordered and impose monotonicity, LB on the fraction of ATs is positive iff

$$\underbrace{P(M = k \mid D = 1)}_{\text{Point mass at M=k when D=1}} \quad > \quad \underbrace{P(M \geq k \mid D = 1) - P(M \geq k \mid D = 0)}_{\text{Treatment effect on survival fn of } M \text{ at } k}$$

- Heuristically, we thus only have identifying power when there is (a) substantial point mass in M, or (b) little effect of D on M in some region
- Relates to results in Gunsilius (2021) on non-testability of IV model with cts treatment (w/o monotonicity)

Monte Carlo Summary Details

- We calibrate sims to our empirical applications and consider the tests of: Cox & Shi (CS), Fang et al (FSST), Andrews et al (ARP); and Kitagawa (K) for the binary M case
- Tradeoffs between finite-sample size control and power
- On balance, tentatively recommend Cox and Shi test for most practical situations
  - Controls size in most simulation designs (except with small number of clusters) and relatively good power (dominates ARP and K)
- ARP has better size control with small # of clusters, but at a big loss of power
  FSST offers power improvements w/large N, but can be over-sized w small/moderate N