Priming bias versus post-treatment bias in experimental designs

Jacob R. Brown (joint work with Matt Blackwell, Sophie Hill, Kosuke Imai, and Teppei Yamamoto)

Department of Government & Institute for Quantitative Social Sciences, Harvard University
https://jakerbrown.github.io/

Motivation

The problem

- Randomization of treatment gives unbiased estimation of ATE
- For conditional ATEs, we need to measure covariates, but when?
  - Before treatment: possibility of priming bias.
  - After treatment: possibility of post-treatment bias.
- **Priming bias**: measurement before treatment alters responses.
- **Post-treatment bias**: conditioning on covariates affected by treatment biases conditional ATEs via selection.

Our contribution

- Sharp bounds for CATEs using a principal strata approach.
  - Under just design assumptions (randomization), we learn almost nothing about interactions.
  - Adding stronger substantive assumptions narrows them.
- Bounds can incorporate pre/post design and covariates

Notation and setup

- (Y_i, T_i, D_i): outcome, treatment, moderator (all binary)
- Indicator for post-treatment measurement of moderator: Z_i
- Potential outcomes per covariate measurement timing: $Y_i(t, z)$.
  - Priming bias occurs when $Y_i(t,0) \neq Y_i(t,1)$.
- Potential values of the covariate: $D_i(t, z)$
  - Potential for post-treatment bias if $D_i(0,1) \neq D_i(1,1)$
- $\rightsquigarrow$ Post-test causal effects conditional on pre-test values of the moderator:

$$\tau(d) \equiv \mathbb{E}(Y_i(1,1) - Y_i(0,1) \mid D_i(0) = d)$$
$$\delta \equiv \tau(1) - \tau(0)$$

- Principal strata defined by the joint distribution of the post-test potential outcomes, conditional on the pre-test moderator:

$$\phi_{ydtd^*} \equiv \Pr(Y_i(t,1) = y, D_i(t,1) = d \mid D_i(0) = d^*)$$

- Key element of a principal strata approach is to connect the unobserved principal strata to the observed strata of the data:

$$P_{ydt} = \Pr(Y_i(t,1) = y, D_i(t,1) = d)$$
$$= Q\phi_{ydt1} + (1-Q)\phi_{ydt0}$$

for all y, t, $d \in \{0,1\}$, where $Q = \Pr(D_i(0) = 1)$.

- Quantity of interest in terms of the principal strata:

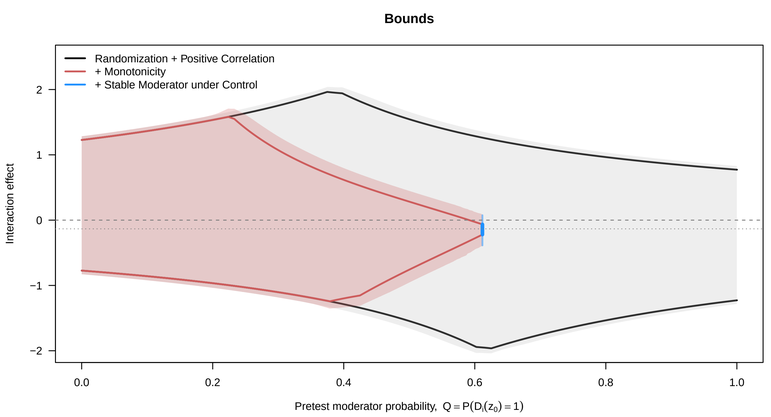
$$\delta = \mathbf{E}(Y_i(1,1) - Y_i(0,1) \mid D_i(0) = 1) - \mathbf{E}(Y_i(1,1) - Y_i(0,1) \mid D_i(0) = 0)$$
$$= \sum_{d=0}^{1} \{\phi_{1d11} - \phi_{1d01} - \phi_{1d10} + \phi_{1d00}\}$$

Substantive assumptions

- We derive sharp bounds under two substantive assumptions.
- **Monotonicity of post-test effect**: $D_i(t,1) \geq D_i(0)$ for all $t = 0, 1$.
  - No one has their value of the moderator lowered by measuring the covariate after treatment.
  - We can derive bounds on δ in terms of (unidentified) $Q = \Pr(D_i(0) = 1)$.
- **Stable moderator under control**: $D_i(0) = D_i(0,1)$
  - "True" pre-test moderator observed in control condition.
  - Plausible when control is neutral or similar to pre-test environment.
  - Identifies Q, substantially narrowing bounds.

Application

- Press et al (2013) randomly vary information about the effectiveness of a nuclear strike versus conventional weapons against a hypothetical Al Qaeda lab in Syria.
  - Sample size: 1,615
  - Treatment: news article describing a military report on the effectiveness of a nuclear strike versus a conventional strike ("control" condition = nuclear more effective; "treatment" condition = equally effective)
  - Outcome: whether respondent prefers a nuclear strike or a conventional strike
  - Moderators: religiosity (dummy for scoring higher than midpoint on 6-point scale about the importance of religion in daily life)
  - Covariates measured after treatment $\rightsquigarrow$ post-test design.



Pre/post design

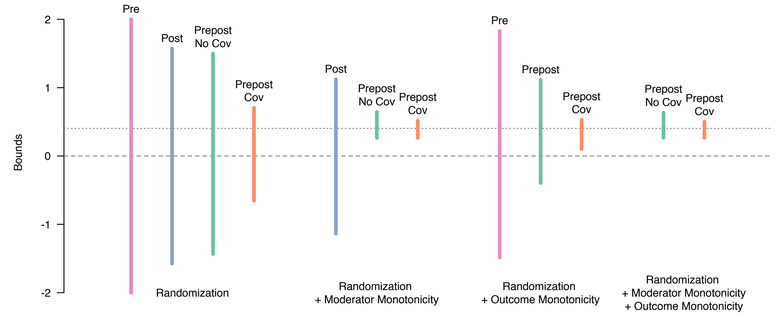
- Randomizing measurement of covariate pre- vs post-treatment allows for estimation of Q
- Avoids stable moderator, but have to make assumptions about priming effects
  - **Priming monotonicity:** $Y_i(t,0) \geq Y_i(t,1)$ for all $t = 0, 1$.
  - Implies that $\psi_{10d_1td_0} = 0$ for any value d_1, t, and d_0

Narrowing bounds using covariates

- Covariates that are unaffected by the question ordering and the treatment provide information about the joint distribution of the covariates and the various potential outcomes
- Augment the principal strata to be conditional on the level of covariate:

$$\psi_{y_1y_0d_1td_0x} = \Pr(Y_i(t,1) = y_1, Y_i(t,0) = y_0, D_i(t,1) = d_1 \mid D_i(0) = d_0, X_i = x)$$

Simulation:



Next steps

- Other results in the paper:
  - Bias formulas under a linear parametric models (especially useful for non-binary moderators).
  - Sensitivity analysis varying how correlated pre- and post-test moderators are.
- Future directions:
  - Adapt Bayesian framework to impute principal strata and estimate posterior distribution of the CATE