

# Unpacking the Black-Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies

**Kosuke Imai**

Princeton University

September 8, 2011

SREE-Fall 2011 Conference

# Project Reference

My talk is based on the collaborative project with L. Keele (Penn State), D. Tingley (Harvard), and T. Yamamoto (MIT)

- “Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies.” *American Political Science Review*
- “Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects.” *Statistical Science*
- “A General Approach to Causal Mediation Analysis.” *Psychological Methods*
- “Experimental Designs for Identifying Causal Mechanisms.”
- “Causal Mediation Analysis Using R.” *Advances in Social Science Research Using R*

All methods can be implemented by our R package `mediation`

# Identification of Causal Mechanisms

- Causal inference is a central goal of scientific research
- Scientists care about causal **mechanisms**, not just about causal effects
- Randomized experiments often only determine **whether** the treatment causes changes in the outcome
- Not **how** and **why** the treatment affects the outcome
- Common criticism of experiments and statistics:

**black box** view of causality

- Question: How can we learn about causal mechanisms from experimental and observational studies?

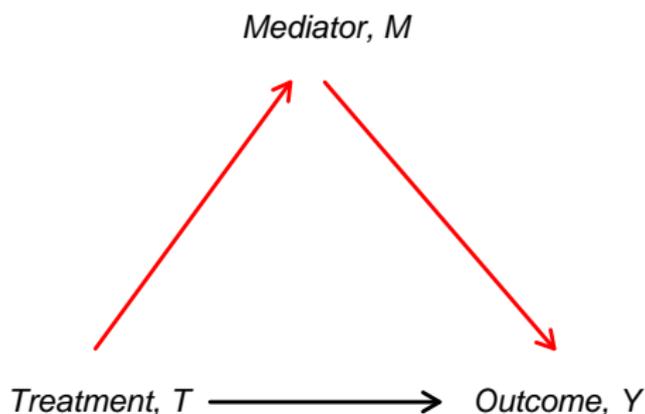
# Goals of the Talk

Present a general framework for statistical design and analysis of causal mechanisms

- 1 Show that the **sequential ignorability** assumption is required to identify mechanisms even in experiments
- 2 Offer a flexible **estimation strategy** under this assumption
- 3 Propose a **sensitivity analysis** to probe this assumption
- 4 Illustrate how to use the R package `mediation`
- 5 Propose **new experimental designs** that do not rely on sequential ignorability
- 6 Cover both experiments and observational studies under the same principle

# What Is a Causal Mechanism?

- Mechanisms as **alternative causal pathways**
- Cochran (1957)'s example:  
soil fumigants increase farm crops by reducing eel-worms
- **Causal mediation analysis**



- Quantities of interest: Direct and indirect effects
- Fast growing methodological literature

# Potential Outcomes Framework

Framework: Potential outcomes model of causal inference

- Binary treatment:  $T_i \in \{0, 1\}$
- Mediator:  $M_i \in \mathcal{M}$
- Outcome:  $Y_i \in \mathcal{Y}$
- Observed pre-treatment covariates:  $X_i \in \mathcal{X}$
  
- Potential mediators:  $M_i(t)$ , where  $M_i = M_i(T_i)$  observed
- Potential outcomes:  $Y_i(t, m)$ , where  $Y_i = Y_i(T_i, M_i(T_i))$  observed
- In a standard experiment, **only one potential outcome** can be observed for each  $i$

# Causal Mediation Effects

- Total causal effect:

$$\tau_i \equiv Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

- Causal mediation (Indirect) effects:

$$\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

- Causal effect of the change in  $M_i$  on  $Y_i$  that would be induced by treatment
- Change the mediator from  $M_i(0)$  to  $M_i(1)$  while holding the treatment constant at  $t$
- Represents the mechanism through  $M_i$

# Total Effect = Indirect Effect + Direct Effect

- **Direct effects:**

$$\zeta_i(t) \equiv Y_i(1, M_i(t)) - Y_i(0, M_i(t))$$

- Causal effect of  $T_i$  on  $Y_i$ , holding mediator constant at its potential value that would realize when  $T_i = t$
- Change the treatment from 0 to 1 while holding the mediator constant at  $M_i(t)$
- Represents all mechanisms other than through  $M_i$
- Total effect = mediation (indirect) effect + direct effect:

$$\tau_i = \delta_i(t) + \zeta_i(1 - t) = \frac{1}{2} \{ \delta_i(0) + \delta_i(1) + \zeta_i(0) + \zeta_i(1) \}$$

# Mechanisms, Manipulations, and Interactions

## Mechanisms

- **Indirect effects:**  $\delta_i(t) \equiv Y_i(t, M_i(1)) - Y_i(t, M_i(0))$
- Counterfactuals about treatment-induced mediator values

## Manipulations

- **Controlled direct effects:**  $\xi_j(t, m, m') \equiv Y_j(t, m) - Y_j(t, m')$
- Causal effect of directly manipulating the mediator under  $T_j = t$

## Interactions

- **Interaction effects:**  $\xi(1, m, m') - \xi(0, m, m') \neq 0$
- Doesn't imply the existence of a mechanism

# What Does the Observed Data Tell Us?

- Quantity of Interest: **Average causal mediation effects**

$$\bar{\delta}(t) \equiv \mathbb{E}(\delta_i(t)) = \mathbb{E}\{Y_i(t, M_i(1)) - Y_i(t, M_i(0))\}$$

- **Average direct effects** ( $\bar{\zeta}(t)$ ) are defined similarly
- Problem:  $Y_i(t, M_i(t))$  is observed but  $Y_i(t, M_i(t'))$  can never be observed
- We have an **identification problem**

⇒ Need additional assumptions to make progress

# Identification under Sequential Ignorability

- Proposed identification assumption: **Sequential Ignorability**

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x, \quad (1)$$

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x \quad (2)$$

- (1) is guaranteed to hold in a standard experiment
- (2) does **not** hold unless  $X_i$  includes all confounders

Under sequential ignorability, both ACME and average direct effects are **nonparametrically identified**  
 (= consistently estimated from observed data)

# Nonparametric Identification

**Theorem:** Under SI, both ACME and average direct effects are given by,

- ACME  $\bar{\delta}(t)$

$$\int \int \mathbb{E}(Y_i | M_i, T_i = t, X_i) \{dP(M_i | T_i = 1, X_i) - dP(M_i | T_i = 0, X_i)\} dP(X_i)$$

- Average direct effects  $\bar{\zeta}(t)$

$$\int \int \{\mathbb{E}(Y_i | M_i, T_i = 1, X_i) - \mathbb{E}(Y_i | M_i, T_i = 0, X_i)\} dP(M_i | T_i = t, X_i) dP(X_i)$$

# Traditional Estimation Method

- **Linear structural equation model (LSEM):**

$$M_i = \alpha_2 + \beta_2 T_i + \xi_2^\top X_i + \epsilon_{i2},$$

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \xi_3^\top X_i + \epsilon_{i3}.$$

- Fit two least squares regressions separately
- Use **product of coefficients** ( $\hat{\beta}_2 \hat{\gamma}$ ) to estimate ACME
- Use asymptotic variance to test significance (Sobel test)
- Under SI and the **no-interaction assumption** ( $\bar{\delta}(1) \neq \bar{\delta}(0)$ ),  $\hat{\beta}_2 \hat{\gamma}$  consistently estimates ACME
- Can be extended to LSEM with interaction terms
- Problem: Only valid for the simplest LSEM

# Proposed General Estimation Algorithm

- 1 Model outcome and mediator
  - Outcome model:  $p(Y_i | T_i, M_i, X_i)$
  - Mediator model:  $p(M_i | T_i, X_i)$
  - These models can be of **any form** (linear or nonlinear, semi- or nonparametric, with or without interactions)
- 2 Predict mediator for both treatment values ( $M_i(1), M_i(0)$ )
- 3 Predict outcome by first setting  $T_i = 1$  and  $M_i = M_i(0)$ , and then  $T_i = 1$  and  $M_i = M_i(1)$
- 4 Compute the average difference between two outcomes to obtain a consistent estimate of ACME
- 5 Monte-Carlo or bootstrapping to estimate uncertainty

# Need for Sensitivity Analysis

- Standard experiments require sequential ignorability to identify mechanisms
- The sequential ignorability assumption is often too strong
- Need to assess the robustness of findings via sensitivity analysis
- **Question:** How large a departure from the key assumption must occur for the conclusions to no longer hold?
- Parametric sensitivity analysis by assuming

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x$$

but not

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x$$

- Possible existence of unobserved *pre-treatment* confounder

# Parametric Sensitivity Analysis

- **Sensitivity parameter:**  $\rho \equiv \text{Corr}(\epsilon_{i2}, \epsilon_{i3})$
- Sequential ignorability implies  $\rho = 0$
- Set  $\rho$  to different values and see how ACME changes

- **Result:**

$$\bar{\delta}(0) = \bar{\delta}(1) = \frac{\beta_2 \sigma_1}{\sigma_2} \left\{ \tilde{\rho} - \rho \sqrt{(1 - \tilde{\rho}^2)/(1 - \rho^2)} \right\},$$

where  $\sigma_j^2 \equiv \text{var}(\epsilon_{ij})$  for  $j = 1, 2$  and  $\tilde{\rho} \equiv \text{Corr}(\epsilon_{i1}, \epsilon_{i2})$ .

- When do my results go away completely?
- $\bar{\delta}(t) = 0$  if and only if  $\rho = \tilde{\rho}$
- Easy to estimate from the regression of  $Y_i$  on  $T_i$ :

$$Y_i = \alpha_1 + \beta_1 T_i + \epsilon_{i1}$$

# Interpreting Sensitivity Analysis with R squares

- Interpreting  $\rho$ : how small is too small?
- An unobserved (pre-treatment) confounder formulation:

$$\epsilon_{i2} = \lambda_2 U_i + \epsilon'_{i2} \quad \text{and} \quad \epsilon_{i3} = \lambda_3 U_i + \epsilon'_{i3}$$

- How much does  $U_i$  have to explain for our results to go away?
- Sensitivity parameters: **R squares**

- 1 Proportion of **previously unexplained variance** explained by  $U_i$

$$R_M^{2*} \equiv 1 - \frac{\text{var}(\epsilon'_{i2})}{\text{var}(\epsilon_{i2})} \quad \text{and} \quad R_Y^{2*} \equiv 1 - \frac{\text{var}(\epsilon'_{i3})}{\text{var}(\epsilon_{i3})}$$

- 2 Proportion of **original variance** explained by  $U_i$

$$\tilde{R}_M^2 \equiv \frac{\text{var}(\epsilon_{i2}) - \text{var}(\epsilon'_{i2})}{\text{var}(M_i)} \quad \text{and} \quad \tilde{R}_Y^2 \equiv \frac{\text{var}(\epsilon_{i3}) - \text{var}(\epsilon'_{i3})}{\text{var}(Y_i)}$$

- Then reparameterize  $\rho$  using  $(R_M^{2*}, R_Y^{2*})$  (or  $(\tilde{R}_M^2, \tilde{R}_Y^2)$ ):

$$\rho = \text{sgn}(\lambda_2 \lambda_3) R_M^* R_Y^* = \frac{\text{sgn}(\lambda_2 \lambda_3) \tilde{R}_M \tilde{R}_Y}{\sqrt{(1 - R_M^2)(1 - R_Y^2)}},$$

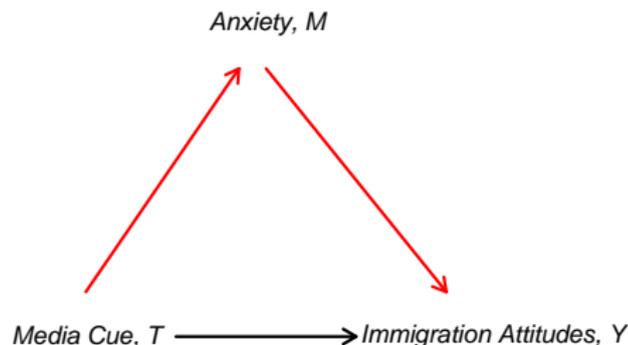
where  $R_M^2$  and  $R_Y^2$  are from the original mediator and outcome models

- $\text{sgn}(\lambda_2 \lambda_3)$  indicates the direction of the effects of  $U_i$  on  $Y_i$  and  $M_i$
- Set  $(R_M^{2*}, R_Y^{2*})$  (or  $(\tilde{R}_M^2, \tilde{R}_Y^2)$ ) to different values and see how mediation effects change

# Example: Anxiety, Group Cues and Immigration

Brader, Valentino & Suhart (2008, AJPS)

- **How** and **why** do ethnic cues affect immigration attitudes?
- Theory: Anxiety transmits the effect of cues on attitudes



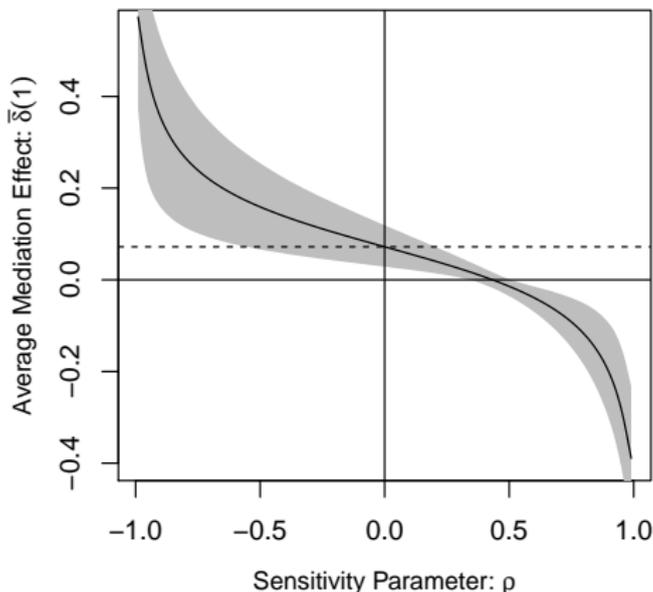
- ACME = Average difference in immigration attitudes due to the change in anxiety induced by the media cue treatment
- Sequential ignorability = No unobserved covariate affecting both anxiety and immigration attitudes

# Reanalysis: Estimates under Sequential Ignorability

- Original method: **Product of coefficients** with the **Sobel test**
  - Valid only when both models are linear w/o  $T-M$  interaction (which they are not)
- Our method: Calculate ACME using our general algorithm

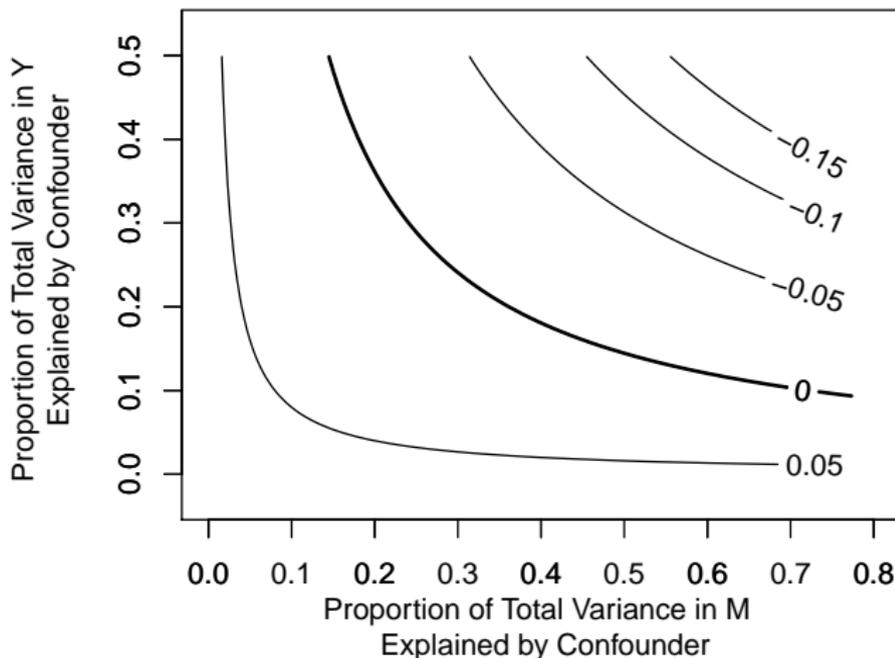
Outcome variables	Product of Coefficients	Average Causal Mediation Effect ( $\delta$ )
Decrease Immigration $\bar{\delta}(1)$	.347 [0.146, 0.548]	.105 [0.048, 0.170]
Support English Only Laws $\bar{\delta}(1)$	.204 [0.069, 0.339]	.074 [0.027, 0.132]
Request Anti-Immigration Information $\bar{\delta}(1)$	.277 [0.084, 0.469]	.029 [0.007, 0.063]
Send Anti-Immigration Message $\bar{\delta}(1)$	.276 [0.102, 0.450]	.086 [0.035, 0.144]

# Reanalysis: Sensitivity Analysis w.r.t. $\rho$



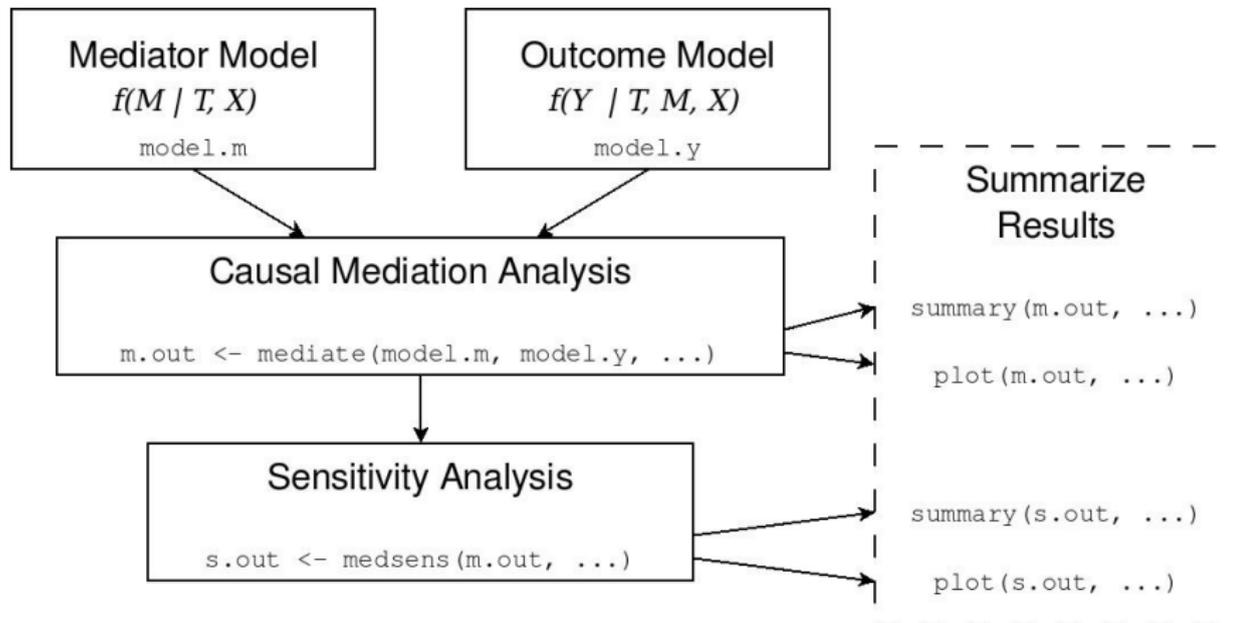
- ACME  $> 0$  as long as the error correlation is less than 0.39 (0.30 with 95% CI)

# Reanalysis: Sensitivity Analysis w.r.t. $\tilde{R}_M^2$ and $\tilde{R}_Y^2$



- An unobserved confounder can account for up to 26.5% of the variation in both  $Y_i$  and  $M_i$  before ACME becomes zero

# Overview of R Package **mediation**



- 1 Fit models for the mediator and outcome variable and store these models.

```
> m <- lm(Mediator ~ Treat + X)
> y <- lm(Y ~ Treat + Mediator + X)
```

- 2 **Mediation analysis:** Feed model objects into the `mediate()` function. Call a summary of results.

```
> m.out <- mediate(m, y, treat = "Treat",
                  mediator = "Mediator")
> summary(m.out)
```

- 3 **Sensitivity analysis:** Feed the output into the `medsens()` function. Summarize and plot.

```
> s.out <- medsens(m.out)
> summary(s.out)
> plot(s.out, "rho")
> plot(s.out, "R2")
```

## Data Types Available via **mediation** (For Now)

- For the `mediate()` function:

<i>Mediator Model Types</i>	<i>Outcome Model Types</i>				
	Linear	GLM	Ordered	Censored	Quantile
Linear ( <code>lm</code> )	✓	✓	✓*	✓	✓
GLM (probit etc.)	✓	✓	✓*	✓	✓
Ordered	✓	✓	✓*	✓	✓
Censored (tobit)	-	-	-	-	-
Quantile	✓*	✓*	✓*	✓*	✓*

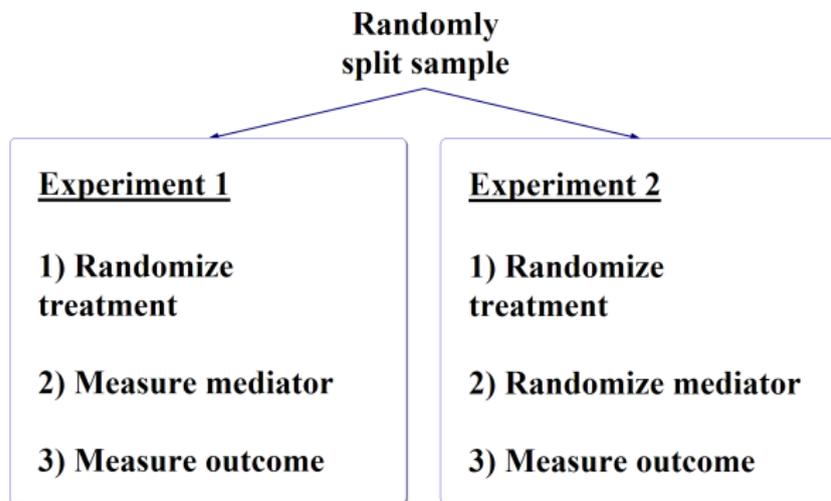
- Can also deal with interaction between treatment and mediator and semiparametric regression
- For the `medsens()` function:

<i>Mediator</i>	<i>Outcome</i>		
	Continuous	Ordered	Binary
Continuous	Yes	No	Yes
Ordered	No	No	No
Binary	Yes	No	No

# Beyond Sequential Ignorability

- Without sequential ignorability, standard experimental design lacks identification power
- Even the sign of ACME is not identified
- Need to develop **alternative experimental designs** for more credible inference
- Possible when the mediator can be directly or indirectly manipulated

# Parallel Design



- Must assume **no direct effect of manipulation** on outcome
- More informative than standard single experiment
- If we assume no  $T-M$  interaction, ACME is point identified

# Example from Behavioral Neuroscience

**Why study brain?:** Social scientists' search for causal mechanisms underlying human behavior

- Psychologists, economists, and even political scientists

**Question:** What mechanism links low offers in an ultimatum game with "irrational" rejections?

- A brain region known to be related to fairness becomes more active when unfair offer received (single experiment design)

**Design solution:** manipulate mechanisms with TMS

- Knoch et al. use TMS to manipulate — turn off — one of these regions, and then observes choices (parallel design)

# Limitations

- Difference between manipulation and mechanism

Prop.	$M_i(1)$	$M_i(0)$	$Y_i(t, 1)$	$Y_i(t, 0)$	$\delta_i(t)$
0.3	1	0	0	1	-1
0.3	0	0	1	0	0
0.1	0	1	0	1	1
0.3	1	1	1	0	0

- Here,  $\mathbb{E}(M_i(1) - M_i(0)) = \mathbb{E}(Y_i(t, 1) - Y_i(t, 0)) = 0.2$ , but  $\bar{\delta}(t) = -0.2$
- **Limitations:**
  - Direct manipulation of the mediator is often impossible
  - Even if possible, manipulation can directly affect outcome
- Need to allow for subtle and indirect manipulations

# Encouragement Design

- Randomly **encourage** subjects to take particular values of the mediator  $M_i$
- Standard **instrumental variable** assumptions (Angrist et al.)

Use a  $2 \times 3$  factorial design:

- ① Randomly assign  $T_i$
  - ② Also randomly decide whether to **positively encourage**, **negatively encourage**, or do nothing
  - ③ Measure mediator and outcome
- Informative inference about the “complier” ACME
  - Reduces to the parallel design if encouragement is perfect
  - Application to the immigration experiment:  
Use autobiographical writing tasks to encourage anxiety

# Crossover Design

- Recall ACME can be identified if we observe  $Y_i(t', M_i(t))$
- Get  $M_i(t)$ , then switch  $T_i$  to  $t'$  while holding  $M_i = M_i(t)$
- **Crossover design:**
  - 1 Round 1: Conduct a standard experiment
  - 2 Round 2: Change the treatment to the opposite status but fix the mediator to the value observed in the first round
- Very powerful – identifies mediation effects for each subject
- Must assume **no carryover effect**: Round 1 must not affect Round 2
- Can be made plausible by design

# Example from Labor Economics

Bertrand & Mullainathan (2004, AER)

- Treatment: Black vs. White names on CVs
- Mediator: Perceived qualifications of applicants
- Outcome: Callback from employers
  
- Quantity of interest: Direct effects of (perceived) race
- Would Jamal get a callback if his name were Greg but his qualifications stayed the same?
  
- Round 1: Send Jamal's actual CV and record the outcome
- Round 2: Send his CV as Greg and record the outcome
  
- Assumptions are plausible

# Crossover Encouragement Design

- **Crossover encouragement design:**
  - ① Round 1: Conduct a standard experiment
  - ② Round 2: Same as crossover, except encourage subjects to take the mediator values

## EXAMPLE Hainmueller & Hiscox (2010, APSR)

- Treatment: Framing immigrants as low or high skilled
- Outcome: Preferences over immigration policy
- Possible mechanism: Low income subjects may expect higher competition from low skill immigrants
- Manipulate expectation using a news story
- Round 1: Original experiment but measure expectation
- Round 2: Flip treatment, but encourage expectation in the same direction as Round 1

# Designing Observational Studies

- Key difference between experimental and observational studies: treatment assignment
- Sequential ignorability:
  - ① Ignorability of treatment given covariates
  - ② Ignorability of mediator given treatment and covariates
- Both (1) and (2) are suspect in observational studies
  
- Statistical control: matching, propensity scores, etc.
- Search for quasi-randomized treatments: “natural” experiments
  
- How can we design observational studies?
- Experiments can serve as templates for observational studies

# Example from Political Science

## EXAMPLE Incumbency advantage

- Estimation of incumbency advantages goes back to 1960s
- Why incumbency advantage? Scaring off quality challenger
- Use of cross-over design (Levitt and Wolfram)
  - ① 1st Round: two non-incumbents in an open seat
  - ② 2nd Round: same candidates with one being an incumbent
- Assume challenger quality (mediator) stays the same
- Estimation of direct effect is possible
- Redistricting as natural experiments (Ansolabehere et al.)
  - ① 1st Round: incumbent in the old part of the district
  - ② 2nd Round: incumbent in the new part of the district
- Challenger quality is the same but treatment is different
- Estimation of direct effect is possible

## Concluding Remarks

- Even in a randomized experiment, a strong assumption is needed to identify causal mechanisms
- However, progress can be made toward this fundamental goal of scientific research with modern statistical tools
- A general, flexible estimation method is available once we assume sequential ignorability
- Sequential ignorability can be probed via sensitivity analysis
- More credible inferences are possible using clever experimental designs
- Insights from new experimental designs can be directly applied when designing observational studies

The project website for papers and software:

<http://imai.princeton.edu/projects/mechanisms.html>

Email for comments and suggestions:

[kimai@Princeton.Edu](mailto:kimai@Princeton.Edu)