

Experimental Methods: Lecture 4

Heterogeneous Treatment Effects

Raymond Duch

May 23, 2023

University of Oxford

Road Map to Lecture 4

- Heterogenous Treatment Effects
- Mediation
- Measurement Error

Heterogenous Treatment Effects

Constant Treatment Effects

- Recall the fundamental assumption about treatment effects
- What does “constant treatment effects” really mean?
- More importantly, is the average treatment effect the same for every single observation in the sample?
- Furthermore, we are often interested in the “generalizability” of experimental findings and their policy relevance
- Treatment effect heterogeneity is one way to address these issues

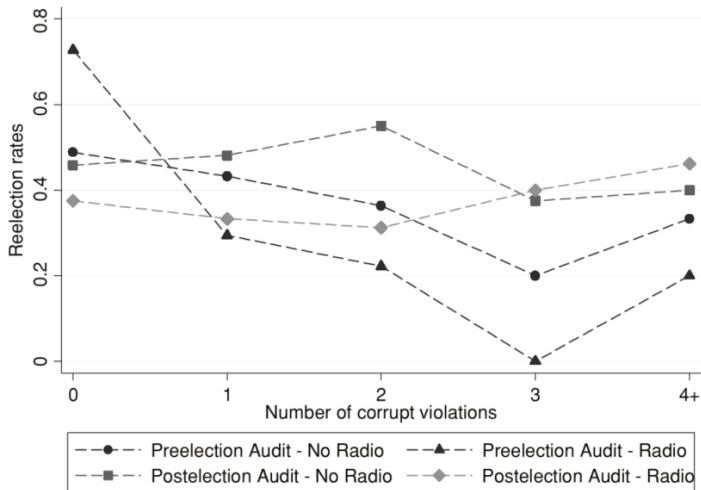


FIGURE IV
Relationship between Reelection Rates and Corruption Levels

Theory

We move away from constant treatment effects and therefore define

$$\tau_i \equiv Y_i(1) - Y_i(0) \quad (1)$$

The fundamental interest under treatment effect heterogeneity is in

$$\begin{aligned} \text{Var}(\tau_i) &= \text{Var}(Y_i(1) - Y_i(0)) \\ &= \text{Var}(Y_i(1)) + \text{Var}(Y_i(0)) + 2\text{Cov}(Y_i(1), Y_i(0)) \end{aligned} \quad (2)$$

Informally, we define treatment effect heterogeneity as *variance of the treatment effect τ_i across subjects*.

What is the problem with Eq. 2?

- This is an old and now for us very familiar problem:
- Any experiment does not allow us to estimate every component of $Var(\tau_i)$
- We have information about the marginal distributions of $Y_i(1)$ and $Y_i(0)$, but not about the joint distribution of these potential outcomes
- So what should we do?

Bounding $Var(\tau_i)$

- Recall that by randomization,
 $E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$
- We can pair each observed $Y_i(1)$ with one of the observed $Y_i(0)$
- But which one? Many combinations possible
- We place bounds suggesting how large or small $Var(\tau_i)$ may be
- Pair values of $Y_i(0)$ and $Y_i(1)$ such that implied $Cov(Y_i(0), Y_i(1))$ is as large (upper bound) or as small (lower bound) as possible
- Sort values in ascending-ascending / ascending-descending order

Testing for heterogeneity

Suppose $H_0 : \text{Var}(\tau_i) = 0$ What if we compared $\text{Var}(Y_i(1))$ and $\text{Var}(Y_i(0))$?

Note that

$$\begin{aligned}\text{Var}(Y_i(1)) &= \text{Var}(Y_i(0) + \tau_i) \\ &= \text{Var}(Y_i(0)) + \text{Var}(\tau_i) + 2\text{Cov}(Y_i(0), \tau_i)\end{aligned}\tag{3}$$

Then, the Null of constant τ_i implies that

$$\text{Var}(\tau_i) = -2\text{Cov}(Y_i(0), \tau_i) = 0\tag{4}$$

These two terms therefore cancel in Eq. 3 and we have shown that testing $H_0 : \text{Var}(\tau_i) = 0$ is the same as testing $\text{Var}(Y_i(1)) = \text{Var}(Y_i(0))$

Observed Outcome Local Budget

We can test this with randomization inference

	Budget share if village head is male	Budget share if village head is female
Village 1	?	15
Village 2	15	?
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	?
Village 7	?	30
Mean	16	22.5
Variance	17.5	112.5

Variance in control:

$$\frac{1}{7-2-1} 2(15 - 16)^2 + 2(20 - 16)^2 + (10 - 16)^2 = 17.5$$

$$\text{Variance in treatment: } \frac{1}{2-1} (15 - 22.5)^2 + (30 - 22.5)^2 = 112.5$$

Interaction

- These approaches test *whether* τ_i varies
- But we want to know more: conditions under which τ_i varies
- We are interested in a different estimand: Conditional Average Treatment Effect (CATE) = ATE for a defined subset of subjects $\tau_i(x) = E[Y_i(1) - Y_i(0)|X_i = x]$ (individual), and, if distribution of X_i is known, $E[\tau_i(X_i)]$ is identified (average)
- Change in treatment effect that occurs from one subgroups to the next is the difference between 2 CATEs
- These subgroups can either be defined by covariate values (*treatment-by-covariate interactions*) or by design (*treatment-by-treatment interactions*)

Treatment-by-covariate interactions

- What is the H_0 here?
- We can test the difference in CATEs with randomization inference or in a regression framework

$$Y_i = a + bl_i + cP_i + dl_iP_i + u_i \quad (5)$$

When $P_i = 0$, the CATE is b :

$$Y_i = a + bl_i + u_i \quad (6)$$

When $P_i = 1$, the CATE is $b + d$:

$$Y_i = a + bl_i + c + dl_i + u_i = (a + c) + (b + d)l_i + u_i \quad (7)$$

where d yields the change in CATEs that occurs when P_i changes

Treatment-by-covariate interactions

- An alternative is to conduct an F test using randomization inference
- Compares sum of squared residuals from the two nested models (alternative model is Eq. 5 and null model is $Y_i = a + bI_i + u_i$)
- If there are interaction effects, Eq. 5 should reduce SSR
- Simulate random assignments and calculate fraction of F-statistics at least as large as the observed F-statistic
- H_0 is that 2 CATEs are the same

Caveats

- Multiple comparisons problem:
 - With 20 covariates, the probability of finding at least 1 that significantly interacts with the treatment at $\alpha = 0.05$ is $1 - (1 - 0.05)^{20} = 0.642$
 - Bonferroni correction (divide target p-value by number of hypothesis tests h)
 - Pre-register your design! (lab)
- Subgroup analysis is non-experimental: groups that are not formed by random assignment, but pre-assignment
- Teacher incentives and teacher education

Treatment-by-treatment interactions

- Manipulate treatment *and* contextual factor / personal characteristic (e.g. COVID and community infection levels)
- Define a factorial experiment as an experiment involving factors 1 and 2, with factor 1 conditions being A and B, and factor 2 conditions being C and D and E
- Then, allocate subjects at random to every possible combination of experimental conditions
- $\{AC, AD, AE, BC, BD, BE\}$

Gottlieb et al. 2018: EGAP Metaketa II: Taxation

Jessica Gottlieb, Adrienne LeBas, Nonso Obikili: “Formalization, Tax Appeals, and Social Intermediaries in Lagos, Nigeria”

T1. Control condition, not encouraged

T2. Encouraged, but not receiving a follow-up visit

T3. Encouraged, and receiving one of the following four follow-up visit combinations:

T3a. Public goods message from state representative

T3b. Enforcement message from state representative

T3c. Public goods message from marketplace representative

T3d. Enforcement message from marketplace representative

Figure 2: Research Design and Assignment Probabilities

				Message Type	
				Public Goods	Enforcement
Control	Formalization Intervention only	Delivery Type	State Rep.	T3a: 5/36	T3b: 5/36
T1: 1/6	T2: 5/18		Market Association	T3c: 5/36	T3d: 5/36

Multiple treatment arms

From Rosen 2010

	Colin		Jose	
	Good grammar	Bad grammar	Good grammar	bad grammar
% Received reply	52	29	37	34
(N)	(100)	(100)	(100)	(100)

This design requires us to be especially careful with defining the causal estimand – what quantity are we interested in in this application?

Multiple treatment arms

Quiz: Why would these two models estimate the same quantities from the Rosen 2010 experiment?

$\{NG, HG, NB, HB\}$ are indicator variables for each of the 4 treatment groups

$J_i = 1$ if Jose Ramirez; $G_i = 1$ if good grammar

$$Y_i = b_1 CG + b_2 JG + b_3 CB + b_4 JB + u_i$$

$$Y_i = a + bJ_i + cG_i + d(J_i G_i) + u_i$$

What quantity in the table do each of the coefficients represent?

Machine Learning and Heterogeneity Illustrated: Lying Experiment (Duch Laroze Zakharov 2018)

- Outcome of interest: Lying about income from RET
- Treatment: Deduction rate that make it more expensive to lie
- Expectation: Lying declines if deduction rates rise

Lying Experiment Design (Duch Laroze Zakharov 2018)

- 3 different tax rates (10%, 20% and 30%)
- Fixed at the group level
- Taxes are redistributed equally among group members
- Public good
- No excludability
- No social gains/losses
- No audits or fines
- 10 rounds
- Paid for one of them at random
- Fixed groups of 4 participants
- Random matching at the beginning

Design: each round

- RET: solve as many additions as possible in 60 sec
- two random two-digit numbers
- Information individual gross profit (before tax)
- Declare their income (to be taxed)
- Information individual net profit (after tax and redistribution)
- Differentiated by profit, tax and redistribution

Lying Experiments



	Mode			
	Lab	Online Lab	Online UK	Mturk
Ability Rank	-0.500*** (0.036)	-0.163*** (0.045)	-0.163** (0.071)	-0.120*** (0.037)
20% Deduction	-0.123*** (0.024)			
30% Deduction	-0.128*** (0.025)	-0.184*** (0.025)	0.042 (0.038)	0.018 (0.021)
No Audit	-0.334*** (0.023)	-0.127*** (0.026)	-0.155*** (0.036)	0.011 (0.024)
Age	0.012*** (0.002)	0.007** (0.003)	-0.0002 (0.001)	0.002** (0.001)
Gender	0.002 (0.022)	0.100*** (0.025)	-0.022 (0.035)	-0.004 (0.020)
Constant	0.715*** (0.066)	0.476*** (0.089)	0.880*** (0.070)	0.576*** (0.043)

Note:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Standard errors clustered by participant

Table 1: GLM estimation on percent declared

BART Estimation

- Bayesian estimation strategy using tree-logic
- Highly flexible estimation strategy

To recover individual estimates of treatment effect:

- Assume binary treatment
- Run BART on experimental data (the training set) to generate both model and predicted outcomes for observed data
- Invert treatment assignment of all observations, and pass through model (test set) to generate set of counterfactual predictions
- For each individual, i , $CATE = Y_{i,D=1} - Y_{i,D=0}$

BART: R Code

```
# Separate outcome and training data
y <- df$report.rate
train <- df[,-1]

# Gen. test data where those treated become untreated, for use in calculating ITT
test <- train
test$treat.het <- ifelse(test$treat.het == 1,0,ifelse(test$treat.het == 0,1,NA))

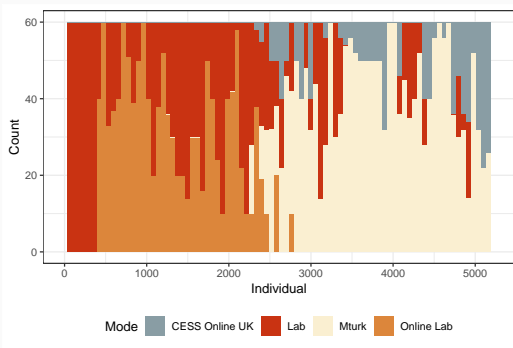
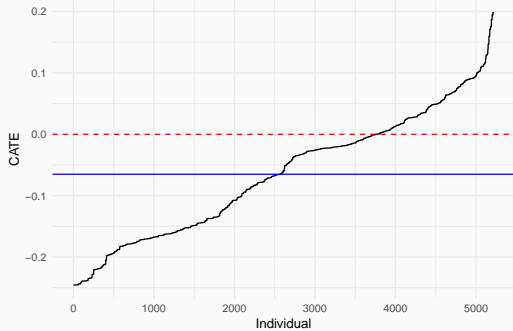
# Run BART for predicted values of observed and synthetic observations
bart.out <- bart(x.train = train, y.train = y, x.test = test)

# Recover CATE estimates and format into dataframe
CATE <- c(bart.out$yhat.train.mean[train$treat.het == 1] - bart.out$yhat.test.mean[test$treat.het == 0],
        bart.out$yhat.test.mean[test$treat.het == 1] - bart.out$yhat.train.mean[train$treat.het == 0])

CATE_df <- data.frame(CATE = CATE)
covars <- rbind(train[train$treat.het == 1,c(2:5)], test[test$treat.het==1,c(2:5)])

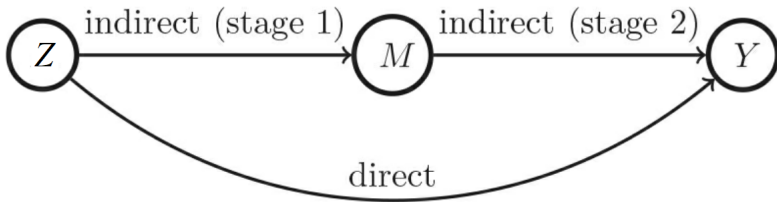
CATE_df <- cbind(CATE_df,covars)
CATE_df <- CATE_df[order(CATE_df$CATE),]
CATE_df$id <- c(1:length(CATE))
```

All replication code available at <https://github.com/rayduch/Experimental-Modes-and-Heterogeneity>



Mediation

Mediation



Classic approaches to mediation

$$\text{Total Effect : } Y_i = \alpha_1 + \beta Z_i + e_{1i} \quad (8)$$

$$\text{Direct Effect : } Y_i = \alpha_2 + \gamma Z_i + \omega M_i + e_{2i} \quad (9)$$

$$\text{Indirect Effect : } (\beta - \gamma) \quad (10)$$

Mediation and Potential Outcomes

- Define $M_i(z)$ as the potential value of M_i when $Z_i = z$
- Define $Y_i(m, z)$ as potential outcome when $M_i = m$ and $Z_i = z$
- $Y_i(M_i(1), 1)$ thus expresses potential outcome when $Z_i = 1$ and M_i takes on potential outcome that occurs when $Z_i = 1$
- Total effect of Z_i on Y_i is $Y_i(M_i(1), 1) - Y_i(M_i(0), 0)$
- What is the direct effect of Z_i on Y_i controlling for M_i ?
 - There is more than one definition
 - $Y_i(M_i(0), 1) - Y_i(M_i(0), 0)$ is direct effect of Z_i on Y_i holding m constant at $M_i(0)$
 - $Y_i(M_i(1), 1) - Y_i(M_i(1), 0)$ is direct effect of Z_i on Y_i holding m constant at $M_i(1)$
 - $Y_i(M_i(0), 1)$ and $Y_i(M_i(1), 0)$ are complex potential outcomes, so named because they are purely imaginary and never occur empirically

Mediation and Potential Outcomes

- What is the direct effect of Z_i on Y_i through M_i ?
 - This is the effect on Y_i of changing from $M_i(0)$ to $M_i(1)$ while holding Z_i constant
 - So again, depending on Z_i , we get two definitions of the indirect effect
 - $Y_i(M_i(1), 1) - Y_i(M_i(0), 1) | Z_i = 1$ and $Y_i(M_i(1), 0) - Y_i(M_i(0), 0) | Z_i = 0$
 - Again $Y_i(M_i(0), 1)$ and $Y_i(M_i(1), 0)$ are the earlier complex potential outcomes
- Each of these four equations involve a term that is fundamentally unobservable
- True even if we assume that both indirect effects are equal
- There is thus a fundamental limitation on what we can learn from an experiment while manipulating only Z_i without making further assumptions

Example: New Drug and Blood Pressure

- FDA Evidence
 1. New Drug (Z)
 2. Blood Pressure (Y)
 3. Aspirin (M)
- Total effect of drug on blood pressure
 - $Y(1) - Y(0)$
- Total effect of drug on aspirin use
 - $M(Z = 1) - M(Z = 0)$
- Total effect of aspirin use on blood pressure
 - $Y(M = 1) - Y(M = 0)$
- Joint effect of drug + aspirin use on blood pressure
 - $Y(11) - Y(00)$

Example: New Drug and Blood Pressure

- Effect of drug when individual forced to refrain from aspirin
 - $Y(10) - Y(00)$
- Effect of drug when individual forced to take aspirin
 - $Y(11) - Y(01)$

Summary

$$\begin{aligned} Y(1) - Y(0) &= Y(1M(1)) - Y(0M(0)) \\ &= \underbrace{Y(1M(1)) - Y(1M(0))}_{\text{indirect}} + \underbrace{Y(1M(0)) - Y(0M(0))}_{\text{direct}} \\ &= \underbrace{Y(1M(1)) - Y(0M(1))}_{\text{direct}} + \underbrace{Y(0M(1)) - Y(0M(0))}_{\text{indirect}} \end{aligned}$$

Ruling Out Mediators

- What if the sharp null hypothesis $M_i(0) = M_i(1)$ is true?
- $Y_i(M_i(1), 1) - Y_i(M_i(0), 1) | Z_i = 1$ and $Y_i(M_i(1), 0) - Y_i(M_i(0), 0) | Z_i = 0$
- Then both indirect effects equal 0. Experiments may indicate when mediation does not occur, but sometimes difficult to do in practice:
 - Need tight estimate around 0
 - Need sharp null to be true, not just $ATE = 0$
- Although sharp null cannot be proven, we can cite evidence suggesting whether this conjecture is a reasonable approximation
- We thus learn something useful about mediation when discovering a lack of causal relationship between Z_i and proposed mediator
- Conversely, if Z_i and M_i have a strong relationship, we cannot rule out M_i as a possible mediator

Manipulating the Mediators

- A fundamental problem is that M_i is not independently manipulated via random intervention
- Could we manipulate M_i as well to build the case for mediation? → In principle, yes, but difficult in practical situations
- Example Y_i is scurvy, Z_i is lemon, M_i is vitamin C
 - We want indirect effect $Y_i(M_i(1), 0) - Y_i(M_i(0), 0)$
 - $M_i(1)$ is vitamin C level of lemon, we feed pills without lemons
 - Still not perfect: Vitamin C in lemons consumed differently from pills, pills might have other effects on Y_i
- Manipulations of M_i are therefore instructive, but ability to provide empirical estimates inevitably requires additional assumptions
- In the Bhavnani example, possible M_i are number of female incumbents, voters' sense of whether it is appropriate or desirable to have women representatives, and turnout rate in local elections

Implicit Mediation

- Consider a treatment Z_i that contains multiple elements in it
- Rather than manipulating M_i , change the treatment to isolate particular elements of Z_i (i.e. Z^1, Z^2, Z^3) whose attributes affect M_i along the way
- Focus is not on demonstrating how a Z_i -induced change in M_i changes Y_i , but on the effect of different isolated treatments on Y_i
- In particular, no attempt to estimate the effects of observed changes in M_i at all

Example: Conditional Cash Transfers

- Interest in conditional cash transfers on poor to keep children in school and attend health clinics
- Field experiments find improved educational outcomes for children in developing countries from these transfers (Baird, McIntosh, and Ozler 2009)
- What could the causal mechanism be?
 1. Cash subsidies allow greater investment in children's welfare
 2. Imposed conditions improve children's welfare
- Baird, McIntosh and Ozler (2009) designed experiment with three groups
 - Control group with no subsidy, instructions, or conditions
 - One treatment group gets cash without conditions
 - Another treatment group gets cash with conditions
 - Finding: Null hypothesis of no difference between treatment groups cannot be rejected

Benefits of Implicit Mediation

1. Simple: Never strays from the unbiased statistical framework of comparing randomly assigned groups
2. By adding and subtracting elements from treatment, this approach lends itself to exploration and discovery of new treatments
 - Facilitates the process of testing basic propositions about what works by providing clues about the active elements that cause a treatment to work particularly well
3. Can gauge treatment effects on a wide array of outcome variables
 - Allows manipulation checks for establishing the empirical relationship between intended and actual treatments
 - Example: Does discussion in the classroom improve performance? Check if treatment increases discussion

Voter Turnout Example

- Gerber, Green, and Larimer (2008) interested in the effect of communication on turnout
- U.S. has voters files, anyone know what they are?
- 180,000 Michigan households in experiment
- 100,000 in control group (no postcards), other groups 20,000 each
- **Civic duty:** "It's your civic duty to vote"
- **Hawthorne:** "It's your civic duty to vote, we're doing a study and will check public records"
- **Self:** "You should vote, here's your recent voting record"
- **Neighbors:** "You should vote, here's your neighbors' voting records and your own"

Results

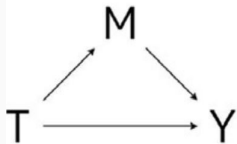
	Control	Civic	Hawthorne	Self	Neighbors
Pct Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N	191,243	38,218	38,204	38,218	38,201

Anyone here know how Gerber followed up on this study?

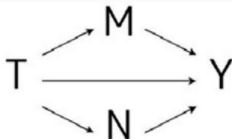
Summary

- We are often curious about the mechanisms by which an experimental treatment transmits its influence
- Adding mediators as right-hand variables to determine this is a flawed strategy that generally provides bias in favor of mediation
- Main issue here is that the mediator is not experimentally manipulated
- In theory we could manipulate mediators experimentally, but this is difficult for two reasons
 1. We never observe complex potential outcomes
 2. Manipulation of mediators directly is often impractical
- However, two lines of inquiry seem promising:
 1. We can rule out mediators easier than we can find them
 2. We can implicitly manipulate mediators

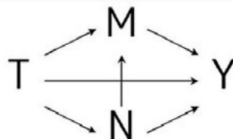
Causal Mechanisms



(a)



(b)



(c)

Potential Outcomes

Total unit effect:

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

Indirect effect:

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

Direct effect:

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

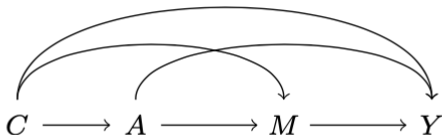
Model Based Causal Mediation Estimation

Sequential Ignorability

$$\{Y_i(t, m), M_i(t)\} \perp T_i | X_i = x$$

$$Y_i(t, m) \perp M_i(t) | T_i = t, X_i = x$$

Unconfoundness Holds:



- no mediator-outcome confounders are influenced by exposure
- a set of observed pre-exposure covariates C captures all confounding.

General Estimator Algorithm

- 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)
- Predict mediator for both treatment values $M_i(1), M_i(0)$
- Predict outcome by first setting $T_i = 1$ and $M_i = M_i(0)$ and then $T_i = 1$ and $M_i = M_i(1)$
- Compute the average difference between two outcomes to obtain a consistent estimate of ACME
- Monte-Carlo or bootstrapping to estimate uncertainty

Example: Continuous Mediator and Binary Outcome

- Estimate the following models:

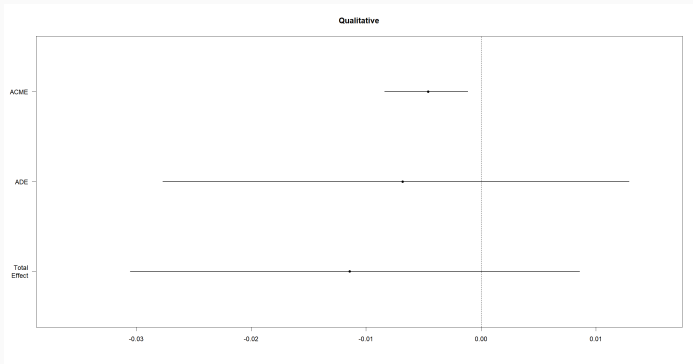
$$M_i = \alpha_2 + \beta_2 T_i + X_i + e_{2i}$$

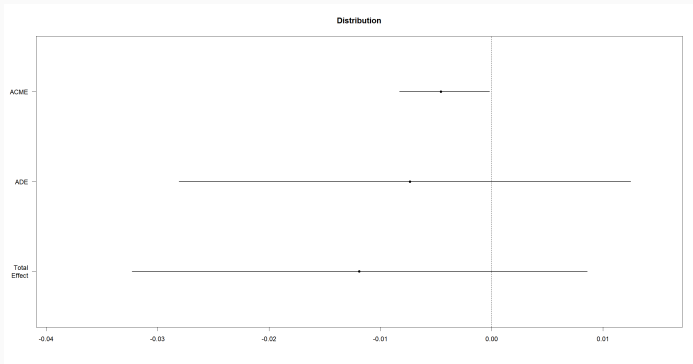
$$Pr(Y_i = 1) = \Phi(\alpha_3 + b_3 T_i + \gamma M_i + X_i + e_{3i})$$

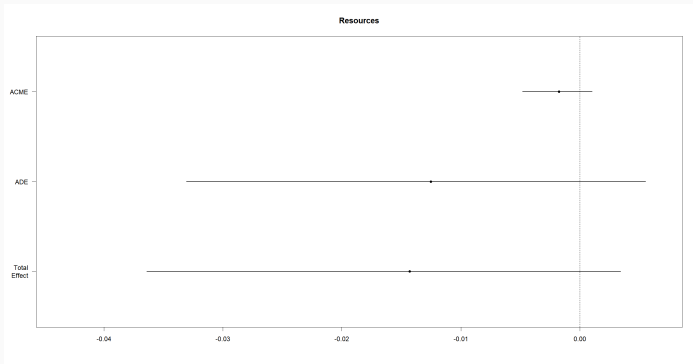
- Predict M_i for $T_i = 1$ and $T_i = 0$. This gives you $\hat{M}_i(1)$ and $\hat{M}_i(0)$
- Predict Y_i with $T_i = 1$ and $\hat{M}_i(0)$ and vice versa
- Take average of these two predictions

Table 5—: Likelihood of Making a Public Donation

Panel A: Regression Analysis							
	Baseline	Malfeasance	Positive	Negative	Qualitative	Resources	Distribution
Intercept	-2.757 (0.557)	-2.761 (0.557)	-4.919 (1.596)	-2.489 (0.600)	-2.836 (0.553)	-2.839 (0.554)	-2.819 (0.553)
Treat	-0.114 (0.107)		0.790 (0.341)	-0.227 (0.114)			
Previous	2.133 (0.108)	2.135 (0.108)	3.095 (0.349)	2.022 (0.115)	2.150 (0.108)	2.139 (0.108)	2.148 (0.108)
IV Malfeasance		-0.007 (0.007)					
Updating					-0.050 (0.021)	-0.029 (0.020)	-0.135 (0.060)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num.Obs.	3439	3439	498	2941	3439	3439	3439
AIC	2494.4	2494.5	336.4	2176.3	2490.0	2493.5	2490.4
BIC	2672.5	2672.6	445.8	2350.0	2668.1	2671.6	2668.5
Log.Lik.	-1218.191	-1218.244	-142.175	-1059.173	-1215.993	-1217.740	-1216.182
Panel B: Mediation Analysis							
	Qualitative		Resources		Distribution		
	ACME	ADE	ACME	ADE	ACME	ADE	
Belief Updating	-0.004 (-0.008, 0.00) [0.21]	-0.006 (-0.030, 0.02)	-0.001 (-0.005, 0.00) [0.06]	-0.010 (-0.033, 0.01)	-0.003 (-0.007, 0.00) [0.17]	-0.008 (-0.029, 0.01)	







Sensitivity: R Code

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

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

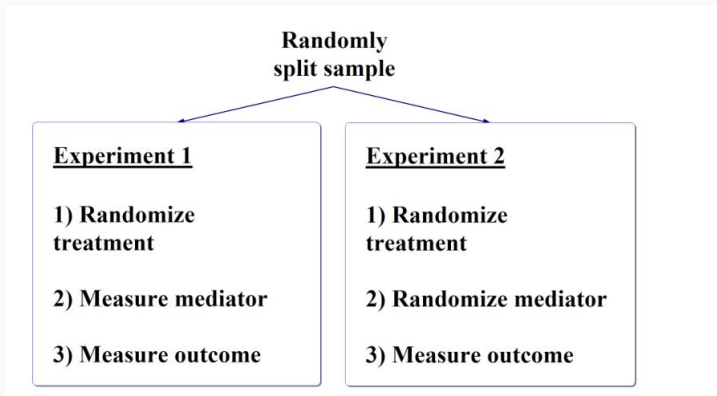
- 2 **Mediation analysis:** Feed model objects into the `mediate()` function. Call a summary of results.

```
> 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")
```

Parallel Design



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) legislator

Encouragement Design

- Randomly *encourage* subjects to take particular values of the mediator M_i
- Standard *instrumental variable* assumptions (Angrist et al.)
- Use a 2×3 factorial design:
 - Randomly assign T_i
 - Also randomly decide whether to positively encourage, negatively encourage, or do nothing
 - 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

Cross-over Design

- Recall *ACME* can be identified if we observe $Y_i(t_0; M_i(t))$
- Get $M_i(t)$, then switch T_i to t_0 while holding $M_i = M_i(t)$
- Crossover design:
 - Round 1: Conduct a standard experiment
 - 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

Cross-over Encouragement Design

- Cross-over encouragement design:
 - Round 1: Conduct a standard experiment
 - 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
 - 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

Measurement Error

Measurement Error

- Empirical measures are error prone
- Measurement error: discrepancy between true value of a variable of interest and its measured value
- It can impact the validity and reliability of study findings (e.g. bias parameter estimates)
- Potentially ubiquitous in experimental and survey research

Types of Measurement Error






- Two main types of measurement error
- Systematic error
 - Deviations from the true value in the same direction across measurements
- Random error
 - Random fluctuations across measurements
- What sources of measurement error?

Sources of Measurement Error

- Fluctuation in respondents' attention
- Rounding due to finite choice menus
- Questionnaire design (e.g. question order effect, anchoring effect)
- Consistency pressure in pre-post designs
- Social desirability bias
- Experimenter demand effect
- How do you know if there is a measurement error?

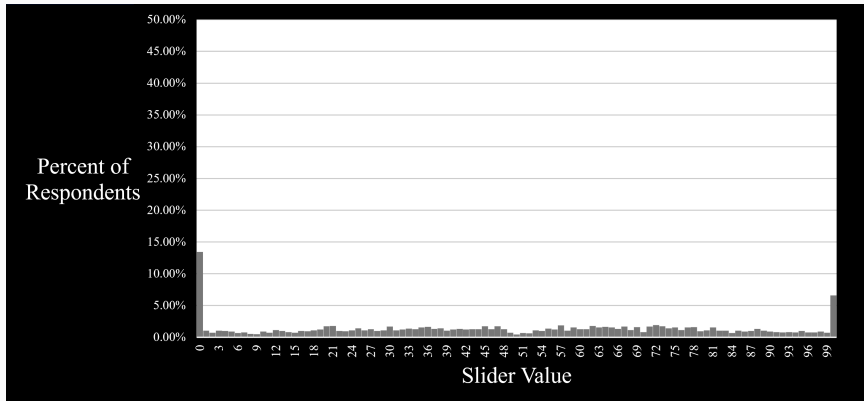
Anchoring effect in Questionnaires

How does what you have just seen make you feel? Please move the slider up or down to indicate to the location that shows how you feel.

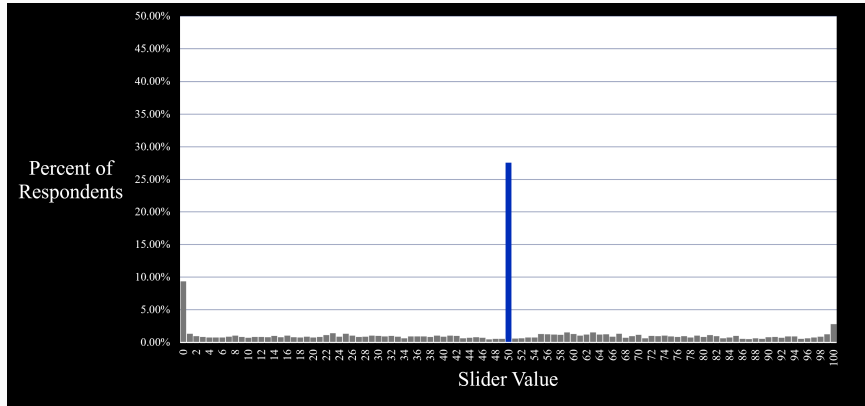
Bitter	Angry	Worried	Enthusiastic	Scared
				
Not Bitter	Not Angry	Not Worried	Not Enthusiastic	Not Scared

Next

Anchoring effect in Questionnaires



Anchoring effect in Questionnaires



Duplicating Noisy Elicitations

- Gillen et al “Experimenting with Measurement Error”
Forthcoming Journal of Political Economy
- Niederle, Muriel and Lise Vesterlund. 2007. “Do Women Shy Away from Competition? Do Men Compete too Much?” Quarterly Journal of Economics

Error as a percent of $\text{Var}[X]$:	0	10%	20%	30%	40%	50%
Panel A: $N = 100$						
$\hat{\alpha}$	0.00 (0.11)	0.06 (0.11)	0.11 (0.12)	0.16 (0.12)	0.21* (0.12)	0.26*** (0.12)
$\hat{\beta}$	1.00*** (0.12)	0.87*** (0.11)	0.75*** (0.11)	0.64*** (0.10)	0.54*** (0.10)	0.44*** (0.09)
Percent of time $\alpha = 0$ rejected at the 5% level with:						
1 noisy measure of X^*	5%	8%	15%	25%	37%	50%
5 noisy measures of X^*	5%	6%	6%	7%	9%	11%
10 noisy measures of X^*	5%	5%	5%	5%	6%	7%
20 noisy measures of X^*	5%	5%	5%	5%	5%	6%
Panel B: $N = 1,000$						
$\hat{\alpha}$	0.00 (0.03)	0.06* (0.04)	0.11*** (0.04)	0.16*** (0.04)	0.21*** (0.04)	0.26*** (0.04)
$\hat{\beta}$	1.00*** (0.04)	0.87*** (0.04)	0.75*** (0.03)	0.64*** (0.03)	0.54*** (0.03)	0.43*** (0.03)
Percent of time $\alpha = 0$ rejected at the 5% level with:						
1 noisy measure of X^*	5%	31%	81%	98%	100%	100%
5 noisy measures of X^*	5%	6%	11%	23%	42%	66%
10 noisy measures of X^*	5%	5%	7%	10%	16%	28%
20 noisy measures of X^*	5%	5%	5%	6%	8%	11%

Table 2: Correlations with X and Y measured with error. True model: $\text{Corr}[X^*, Y^*] = 1$.

Error as a percent of $\text{Var}[X]$ and $\text{Var}[Y]$:	0	10%	20%	30%	40%	50%
Panel A: $N = 100$						
$\widehat{\text{Corr}}[X, Y]$	1.00 (0.00)	0.90*** (0.02)	0.80*** (0.04)	0.70*** (0.05)	0.60*** (0.06)	0.50*** (0.08)
$\widehat{\text{Corr}}[\mathbb{E}[X], \mathbb{E}[Y]]$	1.00 (0.00)	0.95*** (0.01)	0.89*** (0.02)	0.82*** (0.03)	0.75*** (0.04)	0.66*** (0.06)
ORIV $\widehat{\text{Corr}}[X, Y]$	1.00 (0.00)	1.00 (0.01)	1.00 (0.02)	1.00 (0.04)	1.00 (0.06)	1.00 (0.10)
Panel B: $N = 1,000$						
$\widehat{\text{Corr}}[X, Y]$	1.00 (0.00)	0.90*** (0.01)	0.80*** (0.01)	0.70*** (0.02)	0.60*** (0.02)	0.50*** (0.02)
$\widehat{\text{Corr}}[\mathbb{E}[X], \mathbb{E}[Y]]$	1.00 (0.00)	0.95*** (0.00)	0.89*** (0.01)	0.82*** (0.01)	0.75*** (0.01)	0.66*** (0.02)
ORIV $\widehat{\text{Corr}}[X, Y]$	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	1.00 (0.01)	1.00 (0.02)	1.00 (0.03)

Table 3: Gender, competition, and controls

Dependent Variable	Chose to Compete ($N = 783$)							
Male	0.19*** (.034)	0.13*** (.030)	0.11*** (.031)	0.11*** (.031)	0.048 (.033)	0.050 (.034)	0.041 (.033)	0.0063 (.054)
Guessed Tournament Rank	-0.15*** (.017)	$F = 29$ $p = 0.00$	$F = 28$ $p = 0.00$	$F = 23$ $p = 0.00$	$F = 21$ $p = 0.00$			$\chi^2_3 = 8.8$ $p = 0.04$
Tournament Performance	0.086*** (.020)	$F = 1.6$ $p = 0.02$	$F = 1.6$ $p = 0.03$	$F = 1.6$ $p = 0.03$	$F = 1.5$ $p = 0.05$			$\chi^2_{30} = 36$ $p = 0.21$
Performance Difference	-0.021 (.017)	$F = 1.4$ $p = 0.09$	$F = 1.5$ $p = 0.07$	$F = 1.4$ $p = 0.11$	$F = 1.4$ $p = 0.11$			$\chi^2_{25} = 27$ $p = 0.34$
Risk Aversion: MPL #1			0.042*** (.015)					
Overplacement: CRT			0.026* (.015)					
Risk Aversion: Project #2					0.067*** (.016)			
Perceived Performance (ptile.): CRT					-0.042*** (.016)			
All Risk Aversion Controls						$F = 4.9$ $p = 0.00$		
All Overconfidence Controls						$F = 1.8$ $p = 0.05$		
First 5 Principal Components							$F = 37$ $p = 0.00$	
Instrumental Variables (IV)								$\chi^2_7 = 24$ $p = 0.00$
Adjusted R^2	0.038	0.23	0.26	0.27	0.28	0.29	0.22	

Noisy Measures: Risk Elicitation
















	Option A		Option B	
	\$2.00 with a probability of 1/10 , \$1.60 otherwise	 	\$3.85 with a probability of 1/10 , \$0.10 otherwise	
	\$2.00 with a probability of 2/10 , \$1.60 otherwise	 	\$3.85 with a probability of 2/10 , \$0.10 otherwise	
	\$2.00 with a probability of 3/10 , \$1.60 otherwise	 	\$3.85 with a probability of 3/10 , \$0.10 otherwise	
	\$2.00 with a probability of 4/10 , \$1.60 otherwise	 	\$3.85 with a probability of 4/10 , \$0.10 otherwise	
	\$2.00 with a probability of 5/10 , \$1.60 otherwise	 	\$3.85 with a probability of 5/10 , \$0.10 otherwise	
	\$2.00 with a probability of 6/10 , \$1.60 otherwise	 	\$3.85 with a probability of 6/10 , \$0.10 otherwise	
	\$2.00 with a probability of 7/10 , \$1.60 otherwise	 	\$3.85 with a probability of 7/10 , \$0.10 otherwise	
	\$2.00 with a probability of 8/10 , \$1.60 otherwise	 	\$3.85 with a probability of 8/10 , \$0.10 otherwise	
	\$2.00 with a probability of 9/10 , \$1.60 otherwise	 	\$3.85 with a probability of 9/10 , \$0.10 otherwise	
	\$2.00 with a probability of 10/10 , \$1.60 otherwise	 	\$3.85 with a probability of 10/10 , \$0.10 otherwise	

Table 3: Gender, competition, and controls

Dependent Variable	Chose to Compete ($N = 783$)							
Male	0.19*** (.034)	0.13*** (.030)	0.11*** (.031)	0.11*** (.031)	0.048 (.033)	0.050 (.034)	0.041 (.033)	0.0063 (.054)
Guessed Tournament Rank	-0.15*** (.017)	$F = 29$ $p = 0.00$	$F = 28$ $p = 0.00$	$F = 23$ $p = 0.00$	$F = 21$ $p = 0.00$			$\chi^2_3 = 8.8$ $p = 0.04$
Tournament Performance	0.086*** (.020)	$F = 1.6$ $p = 0.02$	$F = 1.6$ $p = 0.03$	$F = 1.6$ $p = 0.03$	$F = 1.5$ $p = 0.05$			$\chi^2_{30} = 36$ $p = 0.21$
Performance Difference	-0.021 (.017)	$F = 1.4$ $p = 0.09$	$F = 1.5$ $p = 0.07$	$F = 1.4$ $p = 0.11$	$F = 1.4$ $p = 0.11$			$\chi^2_{25} = 27$ $p = 0.34$
Risk Aversion: MPL #1			0.042*** (.015)					
Overplacement: CRT			0.026* (.015)					
Risk Aversion: Project #2					0.067*** (.016)			
Perceived Performance (ptile.): CRT					-0.042*** (.016)			
All Risk Aversion Controls						$F = 4.9$ $p = 0.00$		
All Overconfidence Controls						$F = 1.8$ $p = 0.05$		
First 5 Principal Components							$F = 37$ $p = 0.00$	
Instrumental Variables (IV)								$\chi^2_7 = 24$ $p = 0.00$
Adjusted R^2	0.038	0.23	0.26	0.27	0.28	0.29	0.22	

Bounding Experimenter Effect

- Participants may try to infer the experimenter's objective from the experimental environment, and then act accordingly
- Latent demand might be driven by treatment, hence not be orthogonal to treatment variation
- Threat internal and external validity

Bounding Experimenter Effect

- de Quidt et al. (2018)
- Novel technique for assessing robustness to demand effects of findings from experiments and surveys
- Deliberately inducing experimenter demand via "demand treatments"
- Use results to construct plausible upper and lower bounds on demand-free behavior

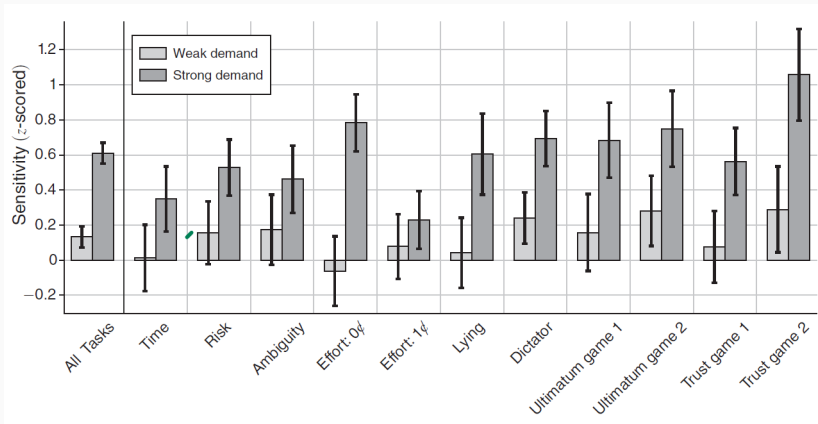
- “Weak” demand treatments: “We expect that participants who are shown these instructions will [work, invest, . . .] more/less than they normally would.”
- “Strong” demand treatments: “You will do us a favor if you [work, invest, . . .] more/less than you normally would.”

- Bounding via "demand treatments" yields useful insights under the assumptions of
 - monotone demand effect
 - bounding
 - monotone sensitivity
- Bounds can be used to obtain demand-robust point estimates of treatment effect

TABLE 2—RESPONSE TO STRONG DEMAND TREATMENTS, ALL INCENTIVIZED TASKS

	Time	Risk	Ambiguity Aversion	Effort 0 cent bonus	Effort 1 cent bonus	Lying	Dictator Game	Ultimatum Game 1	Ultimatum Game 2	Trust Game 1	Trust Game 2
<i>Panel A. Unconditional means</i>											
Positive demand	0.795 (0.024)	0.550 (0.020)	0.583 (0.024)	0.405 (0.011)	0.492 (0.011)	0.606 (0.013)	0.434 (0.015)	0.520 (0.013)	0.474 (0.014)	0.535 (0.024)	0.469 (0.017)
No demand	0.786 (0.025)	0.466 (0.022)		0.341 (0.012)	0.476 (0.012)		0.282 (0.015)				
Negative demand	0.659 (0.028)	0.373 (0.019)	0.428 (0.023)	0.255 (0.011)	0.449 (0.011)	0.510 (0.014)	0.251 (0.014)	0.404 (0.014)	0.337 (0.015)	0.350 (0.022)	0.288 (0.015)
<i>Panel B. Sensitivity (positive – negative)</i>											
Raw data	0.137 (0.037)	0.177 (0.027)	0.155 (0.033)	0.150 (0.016)	0.043 (0.016)	0.096 (0.019)	0.183 (0.021)	0.116 (0.018)	0.136 (0.020)	0.185 (0.032)	0.181 (0.023)
z-score	0.349 (0.095) [0.001]	0.528 (0.082) [0.001]	0.462 (0.098)	0.783 (0.083) [0.001]	0.229 (0.084) [0.020]	0.604 (0.118)	0.694 (0.080) [0.001]	0.684 (0.109)	0.750 (0.111)	0.563 (0.097)	1.058 (0.133)
<i>Panel C. Monotonicity</i>											
Positive – neutral (z-score)	0.022 (0.088) [0.363]	0.252 (0.088) [0.001]		0.333 (0.085) [0.001]	0.084 (0.088) [0.159]		0.574 (0.082) [0.001]				
Negative – neutral (z-score)	–0.327 (0.097) [0.001]	–0.276 (0.086) [0.001]		–0.450 (0.085) [0.001]	–0.145 (0.086) [0.101]		–0.120 (0.080) [0.046]				
Observations	727	728	404	731	714	365	770	409	421	382	371

Notes: This table uses data from incentivized MTurk respondents with strong demand treatments. Panel A displays mean actions with standard errors in the positive, negative, and no-demand conditions, respectively. Panel B presents the raw and z-scored sensitivity of behavior to our demand treatments. Panel C displays the response to our positive and negative demand treatments separately, when “no demand” choices were also collected. Robust standard errors are in parentheses. False-discovery rate adjusted *p*-values are in brackets, adjusting across tests within each task when testing the Monotonicity assumption.



Repeated Measure Design

- Standard post-only experimental designs measure outcome variables only after treatment
- Causal effect is estimated by comparing the outcomes of treated and untreated participants
- Conversely, repeated measure designs measures outcome both before and after exposure to a treatment
- Track participants' attitudes change during the study and whether they vary across experimental conditions
- Increase precision of estimates

Repeated Measure Design

- Does measuring outcomes pre-treatment alter estimates of treatment effect?
- Possible sources of biases include:
 - demand effects
 - consistency pressures
 - attitude strengthening

Repeated Measure Design

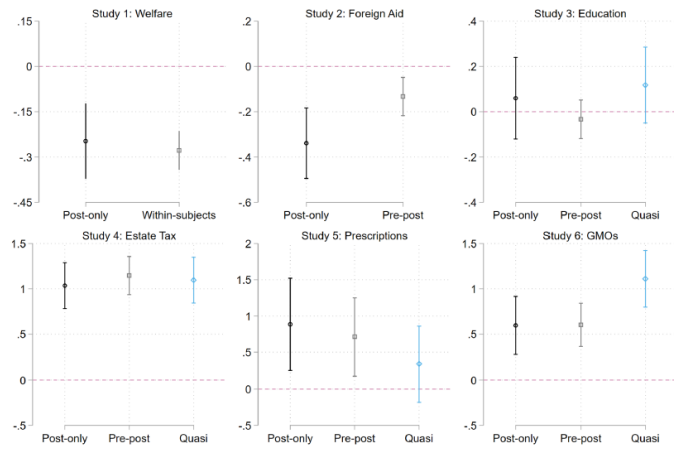
- Clifford et al. (2021) conducts 6 experiments randomly assigning participants to a post-only or a pre-test design
- Study 3: Education spending information experiment
 - Information: average annual per pupil spending on public schools
 - Outcome: support for in/decreasing taxes financing public schools measured on a five-point scale
- Study 4: Estate Tax information experiment
 - Information: federal estate tax applies only to the the wealthiest 0.0006% of Americans
 - Outcome: favor/opposition of the estate tax measured on a seven-point scale

TABLE 1. Comparisons of Experimental Designs

		T ₁		T ₂	
Post-only	R		X	O ₁	
	R			O ₂	
Pre-post	R	O ₁	X	O ₂	
	R	O ₃		O ₄	
Quasi-Pre-post	R	Q ₁	X	O ₁	
	R	Q ₂		O ₂	
Within	R	X	O ₁		O ₂
	R		O ₃	X	O ₄

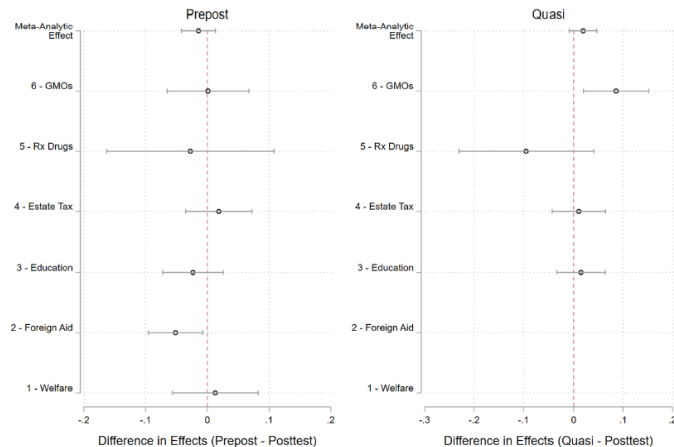
Note: R = Randomization assignment to a group. O = observation of the dependent variable. Q = observation of a variable closely related to the dependent variable. X = exposure to a treatment. T = time of implementation.

FIGURE 1. Treatment Effects by Experimental Design



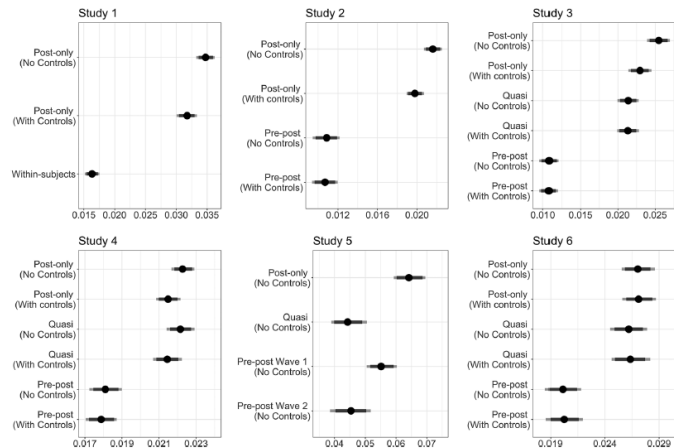
Note: The figures display estimated average treatment effect within each design in each study. In Study 5, the displayed effect is the interaction term between the treatment and respondent partisan identity. The effects in each panel are unstandardized and plotted on the scale of the dependent variable. The bars around the estimates are 95% confidence intervals.

FIGURE 2. Meta-Analysis of Design Effects



Note: The left panel displays the difference between the estimated effect in the pre-post design and the corresponding post-only study. The right panel displays the difference between the quasi and post-only study. All dependent variables were rescaled to range 0–1 and coded so that all treatment effects are positive. The meta-analytic effect (top row) represents the precision-weighted average of all studies. The bars around the point estimates correspond to 95% confidence intervals.

FIGURE 4. Standard Errors of the Estimate by Study, Design, and Analysis



Note: The figure plots the standard error of the treatment effect for each experimental group generated from 1,000 bootstrap estimates. The points correspond to the median of these samples. The dark gray bars correspond to the interval that contains 90% of the samples, and the light gray bars correspond to the interval that contains 95% of the samples.