150C Causal Inference

Randomized Experiments 2

Jonathan Mummolo

Outline

1. Hypothesis Testing in Small Samples
2. Using Covariates in Experiments
3. Examples
   - Social Pressure Experiment
   - Tax Compliance
   - Labor Market Discrimination
   - Job Training Partnership Act (JTPA)
4. Threats to Validity and Ethics

Regression to Estimate the Average Treatment Effect

R Code

```
> library(sandwich)
> library(lmtest)
>
> lout <- lm(earnings~assignmt,data=d)
> coeftest(lout,vcov = vcovHC(lout, type = "HC1")) # matches Stata

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 15040.50     265.38 56.6752 < 2.2e-16 ***
assignmt     1159.43     330.46  3.5085 0.0004524 ***
---
```

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0,$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of } \textit{no} \text{ effect)}$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of \textit{no} effect)}$$

- Let Ω be the set of all possible randomization realizations.
- We only observe the outcomes, Y_i, for one realization of the experiment. We calculate $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$.

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of } \textit{no} \text{ effect)}$$

- Let Ω be the set of all possible randomization realizations.
- We only observe the outcomes, Y_i, for one realization of the experiment. We calculate $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$.
- Under the sharp null hypothesis, we can compute the value that the difference in means estimator would have taken under any other realization, $\hat{\tau}(\omega)$, for $\omega \in \Omega$.

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i
1	3	?	1
2	1	?	1
3	?	0	0
4	?	1	0
$\widehat{\tau}_{ATE}$			1.5

What do we know given the sharp null $H_0 : Y_1 = Y_0$?

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i
1	3	3	1
2	1	1	1
3	0	0	0
4	1	1	0
$\widehat{\tau}_{ATE}$			1.5
$\hat{\tau}(\omega)$			1.5

Given the full schedule of potential outcomes under the sharp null, we can compute the null distribution of ATE_{H_0} across all possible randomization.

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i
1	3	3	1	1
2	1	1	1	0
3	0	0	0	1
4	1	1	0	0
$\widehat{\tau}_{ATE}$			1.5	
$\hat{\tau}(\omega)$			1.5	0.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i
1	3	3	1	1	1
2	1	1	1	0	0
3	0	0	0	1	0
4	1	1	0	0	1
$\widehat{\tau}_{ATE}$			1.5		
$\hat{\tau}(\omega)$			1.5	0.5	1.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0
2	1	1	1	0	0	1
3	0	0	0	1	0	1
4	1	1	0	0	1	0
$\widehat{\tau}_{ATE}$			1.5			
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0
2	1	1	1	0	0	1	1
3	0	0	0	1	0	1	0
4	1	1	0	0	1	0	1
$\widehat{\tau}_{ATE}$			1.5				
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0	0
2	1	1	1	0	0	1	1	0
3	0	0	0	1	0	1	0	1
4	1	1	0	0	1	0	1	1
$\widehat{\tau}_{ATE}$			1.5					
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5	-1.5

So $\Pr(\hat{\tau}(\omega) \geq \widehat{\tau}_{ATE}) = 2/6 \approx .33$.

Which assumptions are needed?

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0	0
2	1	1	1	0	0	1	1	0
3	0	0	0	1	0	1	0	1
4	1	1	0	0	1	0	1	1
$\widehat{\tau}_{ATE}$			1.5					
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5	-1.5

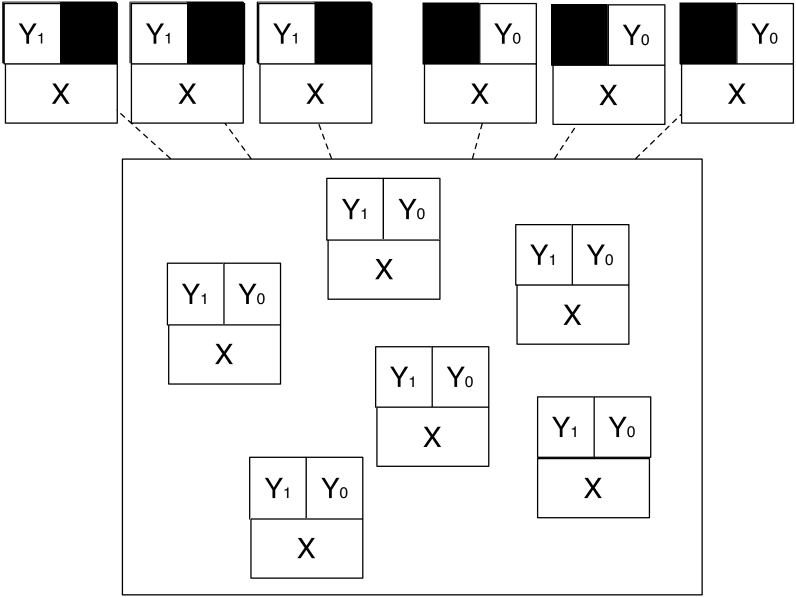
So $\Pr(\hat{\tau}(\omega) \geq \widehat{\tau}_{ATE}) = 2/6 \approx .33$.

Which assumptions are needed? None!

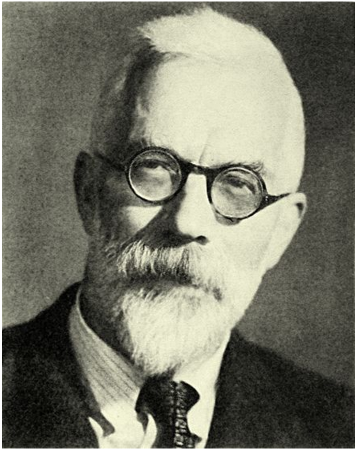
Outline

1. Hypothesis Testing in Small Samples
2. Using Covariates in Experiments
3. Examples
    - Social Pressure Experiment
    - Tax Compliance
    - Labor Market Discrimination
    - Job Training Partnership Act (JTPA)
4. Threats to Validity and Ethics

Covariates and Experiments



Controlling for All Confounders, Seen and Unseen



Randomization “relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which [their] data may be disturbed.” -RA Fisher

Covariates for Balance Checks

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.

Covariates for Balance Checks

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.
- Unlike potential outcomes, you observe baseline covariates for all units. Covariate values are predetermined with respect to the treatment and do not depend on D_i.

Covariates for Balance Checks

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.
- Unlike potential outcomes, you observe baseline covariates for all units. Covariate values are predetermined with respect to the treatment and do not depend on D_i.
- Under randomization, $f_{X|D}(X|D=1) \overset{d}{=} f_{X|D}(X|D=0)$ (equality in distribution).

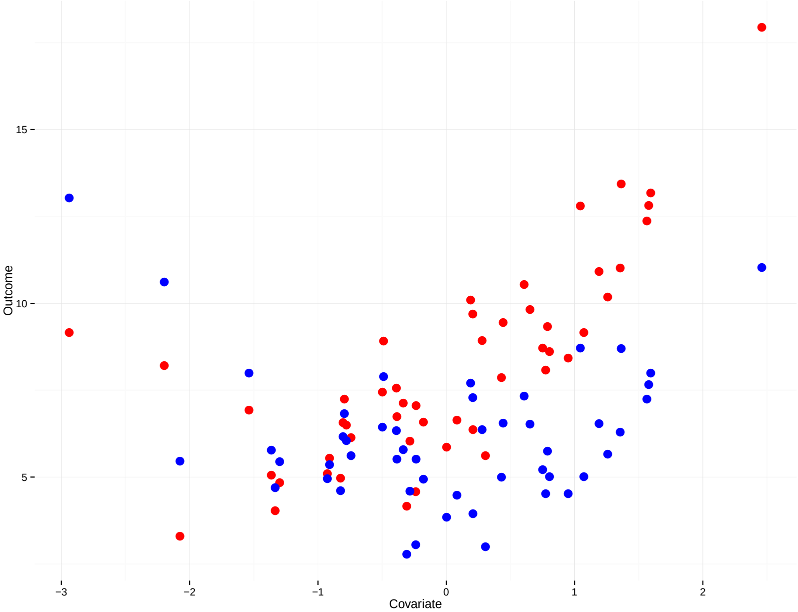
Covariates for Balance Checks

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.
- Unlike potential outcomes, you observe baseline covariates for all units. Covariate values are predetermined with respect to the treatment and do not depend on D_i.
- Under randomization, $f_{X|D}(X|D=1) \overset{d}{=} f_{X|D}(X|D=0)$ (equality in distribution).
- Similarity in distributions of covariates is known as covariate balance.
- If this is not the case, then one of two possibilities:

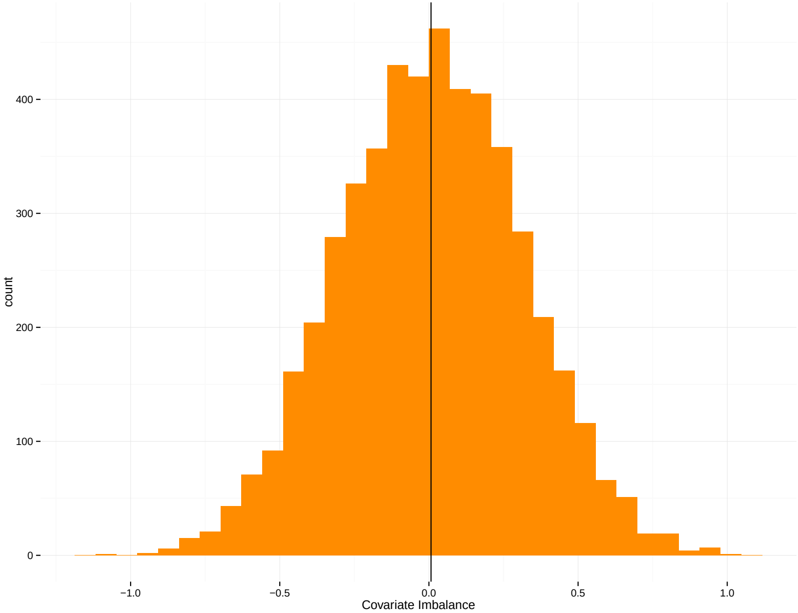
Covariates for Balance Checks

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.
- Unlike potential outcomes, you observe baseline covariates for all units. Covariate values are predetermined with respect to the treatment and do not depend on D_i.
- Under randomization, $f_{X|D}(X|D=1) \overset{d}{=} f_{X|D}(X|D=0)$ (equality in distribution).
- Similarity in distributions of covariates is known as covariate balance.
- If this is not the case, then one of two possibilities:
    - Randomization was compromised.
    - Sampling error (bad luck)
- One should always test for covariate balance on important covariates, using so called "balance checks" (eg. t-tests, F-tests, etc.)

Covariates for Balance Checks



Covariates for Balance Checks



Effect of Training on Earnings

Balance Table: Mean Values of Pre-Training Characteristics

	Treatment Group	Control Group
Pre-Training Earnings ($)	3251	3177
Fraction Males	0.46	0.45
Age (in years)	33	33
Fraction Married	0.26	0.28
Fraction High School Degree	0.69	0.71

Regression Adjusted Estimator for ATE

Definition (Regression Estimator)

We can use the following regression to estimate the ATE while adjusting for the covariates

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Correct for chance covariate imbalances

Regression Adjusted Estimator for ATE

Definition (Regression Estimator)

We can use the following regression to estimate the ATE while adjusting for the covariates

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Correct for chance covariate imbalances
- Increase precision: remove variation in the outcome accounted for by pre-treatment characteristics

Regression Adjusted Estimator for ATE

Definition (Regression Estimator)

We can use the following regression to estimate the ATE while adjusting for the covariates

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Correct for chance covariate imbalances
- Increase precision: remove variation in the outcome accounted for by pre-treatment characteristics
- ATE estimates are robust to model specification (with sufficient N), but best if covariate adjustment is pre-specified

Regression Adjusted Estimator for ATE

Definition (Regression Estimator)

We can use the following regression to estimate the ATE while adjusting for the covariates

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Correct for chance covariate imbalances
- Increase precision: remove variation in the outcome accounted for by pre-treatment characteristics
- ATE estimates are robust to model specification (with sufficient N), but best if covariate adjustment is pre-specified
- Never control for post-treatment covariates (i.e. covariates causally affected by the treatment)!

Regression Adjusted Estimator for ATE

Definition (Regression Estimator)

We can use the following regression to estimate the ATE while adjusting for the covariates

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Correct for chance covariate imbalances
- Increase precision: remove variation in the outcome accounted for by pre-treatment characteristics
- ATE estimates are robust to model specification (with sufficient N), but best if covariate adjustment is pre-specified
- Never control for post-treatment covariates (i.e. covariates causally affected by the treatment)!
- β have no causal interpretation!

Precision Gain in Regression Adjustment

$$
\begin{aligned}
Y_i &= \alpha + \tau_{ATE} D_i + \varepsilon_i && (1)\\
Y_i &= \alpha + \tau_{ATEReg} D_i + \mathbf{X}_i \beta + \varepsilon_i^* && (2)
\end{aligned}
$$

where $\mathbf{X}_i$ is vector of k covariates. Then given iid sampling:

$$
\begin{aligned}
V[\widehat{\tau}_{ATE}] &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^N (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-2} = \frac{SSR_{\widehat{\varepsilon}}}{N-2}\\
V[\widehat{\tau}_{ATEReg}] &= \frac{\sigma_{\varepsilon *}^2}{\sum_{i=1}^N (D_i - \bar{D})^2 (1 - R_D^2)} \quad \text{with } \widehat{\sigma}_{\varepsilon *}^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon *}_i^2}{N-k-1} = \frac{SSR_{\widehat{\varepsilon *}}}{N-k-1}
\end{aligned}
$$

where R_D^2 is R^2 from regression of D of covariates in $\mathbf{X}_i$ and a constant.
So when is $V[\widehat{\tau}_{ATEReg}] < V[\widehat{\tau}_{ATE}]$?

Precision Gain in Regression Adjustment

$$
\begin{align}
Y_i &= \alpha + \tau_{ATE} D_i + \varepsilon_i \tag{1}\\
Y_i &= \alpha + \tau_{ATEReg} D_i + \mathbf{X}_i \beta + \varepsilon_i^* \tag{2}
\end{align}
$$

where $\mathbf{X}_i$ is vector of k covariates. Then given iid sampling:

$$
\begin{aligned}
V[\widehat{\tau}_{ATE}] &= \frac{\sigma^2_{\varepsilon}}{\sum_{i=1}^N (D_i - \bar{D})^2} \qquad \text{with } \widehat{\sigma}^2_{\varepsilon} = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-2} = \frac{SSR_{\widehat{\varepsilon}}}{N-2}\\
V[\widehat{\tau}_{ATEReg}] &= \frac{\sigma^2_{\varepsilon^*}}{\sum_{i=1}^N (D_i - \bar{D})^2 (1 - R_D^2)} \text{ with } \widehat{\sigma}^2_{\varepsilon^*} = \frac{\sum_{i=1}^N \widehat{\varepsilon^*}_i^2}{N-k-1} = \frac{SSR_{\widehat{\varepsilon^*}}}{N-k-1}
\end{aligned}
$$

where R_D^2 is R^2 from regression of D of covariates in $\mathbf{X}_i$ and a constant.
Since $R_D^2 \approx 0$ $V[\widehat{\tau}_{ATEReg}] < V[\widehat{\tau}_{ATE}]$ if $\frac{SSR_{\widehat{\varepsilon^*}}}{n-k-1} < \frac{SSR_{\widehat{\varepsilon}}}{n-2}$

Regression Adjusted Estimator for ATE

R Code

```
> lout <- lm(earnings~assignmt,data=d)
> coeftest(lout,vcov = vcovHC(lout, type = "HC1"))

t test of coefficients:

            Estimate Std. Error t value             Pr(>|t|)    
(Intercept) 15040.50     265.38 56.6752 < 0.0000000000000022 ***
 assignmt    1159.43     330.46  3.5085            0.0004524 ***
---

> lout <- lm(earnings~assignmt+prevearn+sex+age+married+hsorged,data=d)
> coeftest(lout,vcov = vcovHC(lout, type = "HC1"))

t test of coefficients:

               Estimate  Std. Error t value              Pr(>|t|)    
(Intercept) 9289.801525  579.370545 16.0343 < 0.0000000000000022 ***
assignmt    1161.026601  307.031813  3.7815            0.0001567 ***
prevearn       1.232860    0.058648 21.0214 < 0.0000000000000022 ***
sex         3835.020883  308.124891 12.4463 < 0.0000000000000022 ***
age          -94.034325   13.678179 -6.8748  0.00000000000653927 ***
married     2906.269212  373.568794  7.7797  0.00000000000007905 ***
hsorged     3330.175626  315.216182 10.5647 < 0.0000000000000022 ***
---
```

Outline

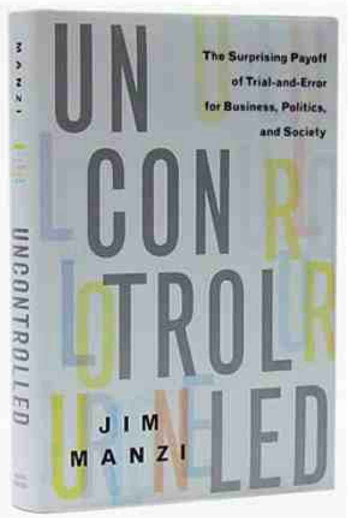
1. Hypothesis Testing in Small Samples

2. Using Covariates in Experiments

3. Examples
   - Social Pressure Experiment
   - Tax Compliance
   - Labor Market Discrimination
   - Job Training Partnership Act (JTPA)

4. Threats to Validity and Ethics

Experiments in Popular Culture


MANZI
UNCONTROLLED
UN
CON
TROL
LED
The Surprising Payoff of Trial-and-Error for Business, Politics, and Society
JIM MANZI

The Rise of Experiments

Large increase in the use of experiments in the social sciences: laboratory, survey, and field experiments (see syllabus)

The Rise of Experiments

Large increase in the use of experiments in the social sciences: laboratory, survey, and field experiments (see syllabus) Abbreviated list of examples:

- *Program Evaluation*: development programs, education programs, weight loss programs, fundraising, deliberative polls, virginity pledging, advertising campaigns, mental exercise for elderly

The Rise of Experiments

Large increase in the use of experiments in the social sciences: laboratory, survey, and field experiments (see syllabus) Abbreviated list of examples:

- *Program Evaluation*: development programs, education programs, weight loss programs, fundraising, deliberative polls, virginity pledging, advertising campaigns, mental exercise for elderly
- *Public policy evaluations*: teacher pay, class size, speed traps, vouchers, alternative sentencing, job training, health insurance subsidies, tax compliance, public housing, jury selection, police interventions

The Rise of Experiments

Large increase in the use of experiments in the social sciences: laboratory, survey, and field experiments (see syllabus) Abbreviated list of examples:

- *Program Evaluation*: development programs, education programs, weight loss programs, fundraising, deliberative polls, virginity pledging, advertising campaigns, mental exercise for elderly
- *Public policy evaluations*: teacher pay, class size, speed traps, vouchers, alternative sentencing, job training, health insurance subsidies, tax compliance, public housing, jury selection, police interventions
- *Behavioral Research*: persuasion, mobilization, education, income, interpersonal influence, conscientious health behaviors, media exposure, deliberation, discrimination

The Rise of Experiments

Large increase in the use of experiments in the social sciences: laboratory, survey, and field experiments (see syllabus) Abbreviated list of examples:

- *Program Evaluation*: development programs, education programs, weight loss programs, fundraising, deliberative polls, virginity pledging, advertising campaigns, mental exercise for elderly
- *Public policy evaluations*: teacher pay, class size, speed traps, vouchers, alternative sentencing, job training, health insurance subsidies, tax compliance, public housing, jury selection, police interventions
- *Behavioral Research*: persuasion, mobilization, education, income, interpersonal influence, conscientious health behaviors, media exposure, deliberation, discrimination
- *Research on Institutions*: rules for authorizing decisions, rules of succession, monitoring performance, transparency, corruption auditing, electoral systems

Outline

1. Hypothesis Testing in Small Samples

2. Using Covariates in Experiments

3. Examples
   - Social Pressure Experiment
   - Tax Compliance
   - Labor Market Discrimination
   - Job Training Partnership Act (JTPA)

4. Threats to Validity and Ethics

Social Pressure Experiment: Design

- Voter turnout theories based on rational self-interested behavior generally fail to predict significant turnout unless they account for the utility that citizens receive from performing their civic duty.

Social Pressure Experiment: Design

- Voter turnout theories based on rational self-interested behavior generally fail to predict significant turnout unless they account for the utility that citizens receive from performing their civic duty.
- Two aspects of this type of utility: intrinsic satisfaction from behaving in accordance with a norm and extrinsic incentives to comply.

Social Pressure Experiment: Design

- Voter turnout theories based on rational self-interested behavior generally fail to predict significant turnout unless they account for the utility that citizens receive from performing their civic duty.

- Two aspects of this type of utility: intrinsic satisfaction from behaving in accordance with a norm and extrinsic incentives to comply.

- Gerber, Green, and Larimer (2008) test these motives in a large scale field experiment by applying varying degrees of intrinsic and extrinsic pressure on voters using a series of mailings to 180,002 households before the August 2006 primary election in Michigan.

Social Pressure Experiment: Treatments

- **Civic Duty**
    - Encouraged to vote.

Social Pressure Experiment: Treatments

- **Civic Duty**
  - Encouraged to vote.
- **Hawthorne**
  - Encouraged to vote.
  - Told that researchers would be checking on whether they voted: “YOU ARE BEING STUDIED!”

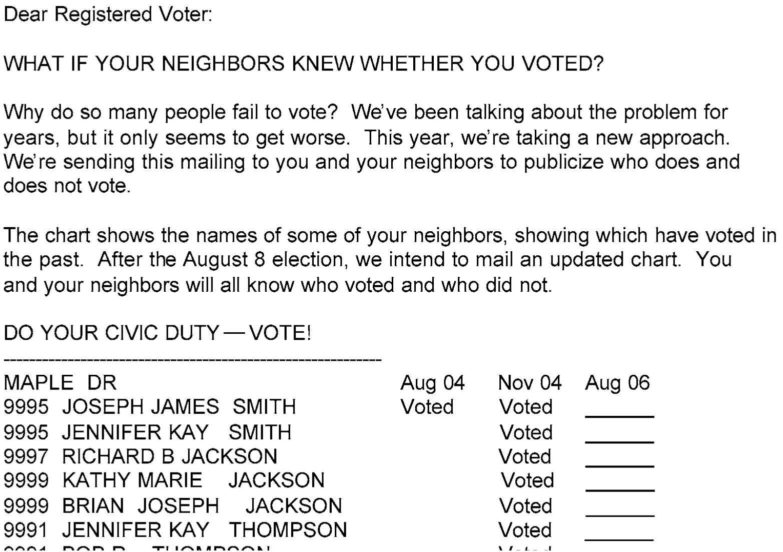
Social Pressure Experiment: Treatments

- **Civic Duty**
  - Encouraged to vote.
- **Hawthorne**
  - Encouraged to vote.
  - Told that researchers would be checking on whether they voted: "YOU ARE BEING STUDIED!"
- **Self**
  - Encouraged to vote.
  - Told that whether one votes is a matter of public record.
  - Shown whether members of their own household voted in the last two elections and promised to send post-card after election indicating whether or not they voted.

Social Pressure Experiment: Treatments

- **Civic Duty**
  - Encouraged to vote.
- **Hawthorne**
  - Encouraged to vote.
  - Told that researchers would be checking on whether they voted: “YOU ARE BEING STUDIED!”
- **Self**
  - Encouraged to vote.
  - Told that whether one votes is a matter of public record.
  - Shown whether members of their own household voted in the last two elections and promised to send post-card after election indicating whether or not they voted.
- **Neighbors**
  - Like **Self** treatment but in addition recipients are shown whether the neighbors on the block voted in the last two elections.
  - Promised to inform neighbors whether or not subject voted after election.

Social Pressure Experiment: Neighbors Treatment



Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______
9999 BRIAN JOSEPH JACKSON		Voted	______
9991 JENNIFER KAY THOMPSON		Voted	______
9991 BOB R THOMPSON		Voted	

Social Pressure Experiment: Balance Check

```
d <- read.dta("gerber.dta")
covars <- c("hh_size","g2002","g2000","p2004","p2002","p2000","sex","yob")
print(aggregate(d[,covars],by=list(d$treatment),mean),digits=3)
```

Social Pressure Experiment: Balance Check

```
d <- read.dta("gerber.dta")
covars <- c("hh_size","g2002","g2000","p2004","p2002","p2000","sex","yob")
print(aggregate(d[,covars],by=list(d$treatment),mean),digits=3)
```

```
     Group.1 hh_size g2002 g2000 p2004 p2002 p2000   sex  yob
1    Control    1.91 0.834 0.866 0.417 0.409 0.265 0.502 1955
2  Hawthorne    1.91 0.836 0.867 0.419 0.412 0.263 0.503 1955
3 Civic Duty    1.91 0.836 0.865 0.416 0.410 0.266 0.503 1955
4  Neighbors    1.91 0.835 0.865 0.423 0.406 0.263 0.505 1955
5       Self    1.91 0.835 0.863 0.421 0.410 0.263 0.501 1955
```

Social Pressure Experiment: Balance Check

```
print(aggregate(d[,covars],by=list(d$treatment),sd),digits=3)
```

Social Pressure Experiment: Balance Check

```
print(aggregate(d[,covars],by=list(d$treatment),sd),digits=3)
```

```
     Group.1 hh_size g2002 g2000 p2004 p2002 p2000   sex  yob
1    Control   0.720 0.294 0.271 0.444 0.435 0.395 0.273 12.9
2  Hawthorne   0.718 0.295 0.270 0.444 0.435 0.393 0.272 12.9
3 Civic Duty   0.729 0.293 0.270 0.444 0.435 0.396 0.275 12.9
4  Neighbors   0.728 0.295 0.273 0.445 0.434 0.393 0.274 13.0
5       Self   0.718 0.294 0.274 0.444 0.434 0.392 0.274 12.8
```

Social Pressure Experiment: Balance Check

```
print(aggregate(d[,c("yob")],by=list(d$treatment),quantile),digits=3)
```

Social Pressure Experiment: Balance Check

```
print(aggregate(d[,c("yob")],by=list(d$treatment),quantile),digits=3)
```

```
    Group.1 x.0% x.25% x.50% x.75% x.100%
1     Control 1900  1946  1957  1964   1986
2   Hawthorne 1908  1946  1957  1964   1984
3 Civic Duty 1906  1947  1957  1964   1986
4   Neighbors 1905  1946  1957  1964   1986
5        Self 1908  1946  1957  1964   1986
```

Social Pressure Experiment: Multivariate Balance Check

```
form <- as.formula(paste("treatment","~",paste(covars,collapse="+")))
form
treatment ~ hh_size + g2002 + g2000 + p2004 + p2002 + p2000 +
    sex + yob
summary(lm(form,data=d))
```

Social Pressure Experiment: Multivariate Balance Check

```
form <- as.formula(paste("treatment","~",paste(covars,collapse="+")))
form
treatment ~ hh_size + g2002 + g2000 + p2004 + p2002 + p2000 +
    sex + yob
summary(lm(form,data=d))
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.7944614  0.5496699   3.265   0.0011 **
hh_size     -0.0032727  0.0051836  -0.631   0.5278
g2002        0.0121818  0.0123389   0.987   0.3235
g2000       -0.0233410  0.0133489  -1.749   0.0804 .
p2004        0.0118147  0.0079130   1.493   0.1354
p2002        0.0018055  0.0081488   0.222   0.8247
p2000       -0.0031604  0.0087721  -0.360   0.7186
sex          0.0031331  0.0125052   0.251   0.8022
yob          0.0001671  0.0002815   0.594   0.5528
Residual standard error: 1.449 on 179993 degrees of freedom
Multiple R-squared: 4.004e-05,  Adjusted R-squared: -4.406e-06
F-statistic: 0.9009 on 8 and 179993 DF,  p-value: 0.5145
```

Social Pressure Experiment: Results

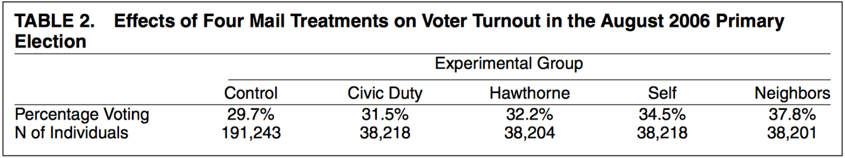


TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Social Pressure Experiment: Results

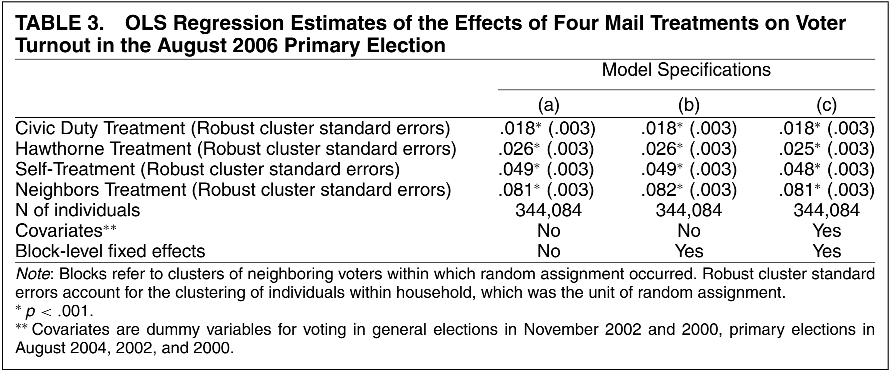


TABLE 3. OLS Regression Estimates of the Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Model Specifications		
	(a)	(b)	(c)
Civic Duty Treatment (Robust cluster standard errors)	.018* (.003)	.018* (.003)	.018* (.003)
Hawthorne Treatment (Robust cluster standard errors)	.026* (.003)	.026* (.003)	.025* (.003)
Self-Treatment (Robust cluster standard errors)	.049* (.003)	.049* (.003)	.048* (.003)
Neighbors Treatment (Robust cluster standard errors)	.081* (.003)	.082* (.003)	.081* (.003)
N of individuals	344,084	344,084	344,084
Covariates**	No	No	Yes
Block-level fixed effects	No	Yes	Yes

Note: Blocks refer to clusters of neighboring voters within which random assignment occurred. Robust cluster standard errors account for the clustering of individuals within household, which was the unit of random assignment.

* $p < .001$.

** Covariates are dummy variables for voting in general elections in November 2002 and 2000, primary elections in August 2004, 2002, and 2000.

Outline

1. Hypothesis Testing in Small Samples

2. Using Covariates in Experiments

3. Examples
   - Social Pressure Experiment
   - Tax Compliance
   - Labor Market Discrimination
   - Job Training Partnership Act (JTPA)

4. Threats to Validity and Ethics

Tax Compliance Experiment

- Can tax evasion be reduced by appeals to taxpayers' conscience?

Tax Compliance Experiment

- Can tax evasion be reduced by appeals to taxpayers' conscience?
- Slemrod, Blumenthal, and Christian (2001, JPubE) worked with Minnesota Department of Revenue to conduct income tax compliance experiments to test alternative strategies for improving voluntary compliance

Tax Compliance Experiment

- Can tax evasion be reduced by appeals to taxpayers' conscience?
- Slemrod, Blumenthal, and Christian (2001, JPubE) worked with Minnesota Department of Revenue to conduct income tax compliance experiments to test alternative strategies for improving voluntary compliance
- In 1994, group of 1724 randomly selected taxpayers was informed by letter that the returns they were about to file, both state and federal, would be "closely examined"
    - D: Educational letter
    - Y: Changes in reported income and taxed paid between 1994 and 1993 (from federal and state returns)
    - Stratify by income and high/low opportunity to evade

Tax Compliance Experiment

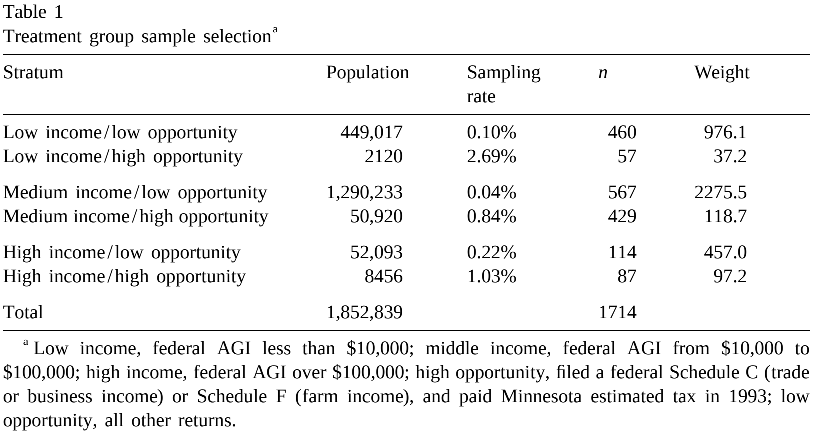


Table 1
Treatment group sample selection[a]

Stratum	Population	Sampling rate	*n*	Weight
Low income/low opportunity	449,017	0.10%	460	976.1
Low income/high opportunity	2120	2.69%	57	37.2
Medium income/low opportunity	1,290,233	0.04%	567	2275.5
Medium income/high opportunity	50,920	0.84%	429	118.7
High income/low opportunity	52,093	0.22%	114	457.0
High income/high opportunity	8456	1.03%	87	97.2
Total	1,852,839		1714	

[a] Low income, federal AGI less than $10,000; middle income, federal AGI from $10,000 to $100,000; high income, federal AGI over $100,000; high opportunity, filed a federal Schedule C (trade or business income) or Schedule F (farm income), and paid Minnesota estimated tax in 1993; low opportunity, all other returns.

Tax Compliance Experiment

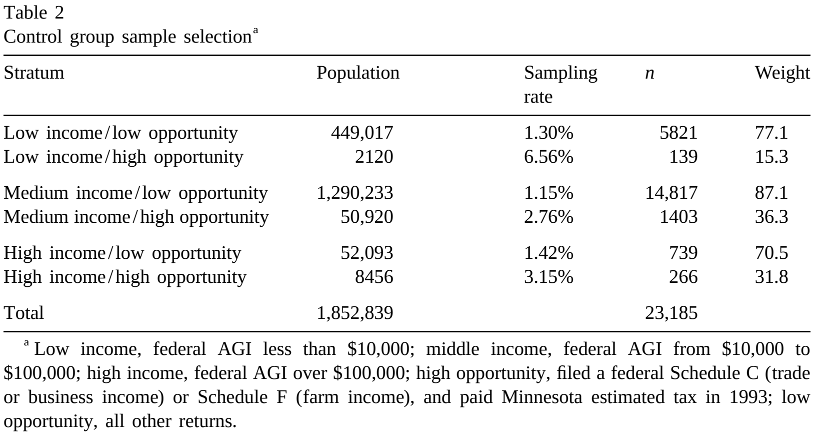


Table 2
Control group sample selection[a]

Stratum	Population	Sampling rate	*n*	Weight
Low income/low opportunity	449,017	1.30%	5821	77.1
Low income/high opportunity	2120	6.56%	139	15.3
Medium income/low opportunity	1,290,233	1.15%	14,817	87.1
Medium income/high opportunity	50,920	2.76%	1403	36.3
High income/low opportunity	52,093	1.42%	739	70.5
High income/high opportunity	8456	3.15%	266	31.8
Total	1,852,839		23,185	

[a] Low income, federal AGI less than $10,000; middle income, federal AGI from $10,000 to $100,000; high income, federal AGI over $100,000; high opportunity, filed a federal Schedule C (trade or business income) or Schedule F (farm income), and paid Minnesota estimated tax in 1993; low opportunity, all other returns.

Results for Full Sample

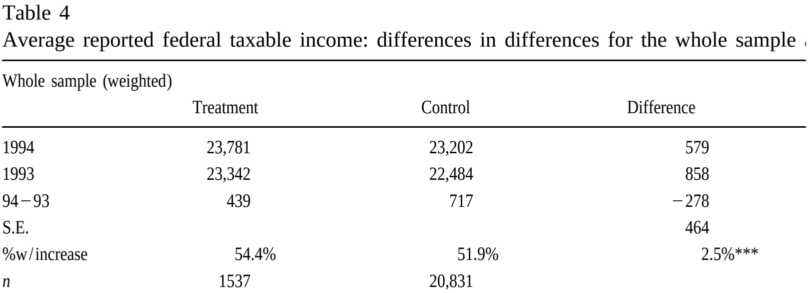


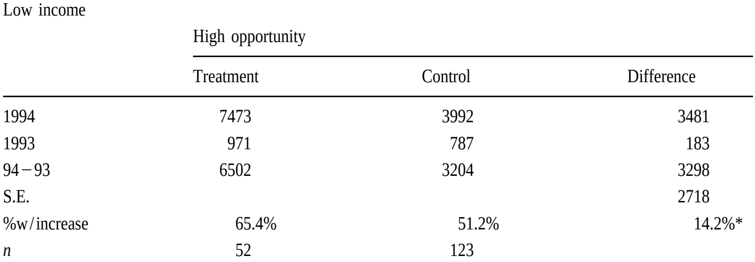
Table 4

Average reported federal taxable income: differences in differences for the whole sample

Whole sample (weighted)

	Treatment	Control	Difference
1994	23,781	23,202	579
1993	23,342	22,484	858
94 − 93	439	717	−278
S.E.			464
%w/increase	54.4%	51.9%	2.5%***
n	1537	20,831	

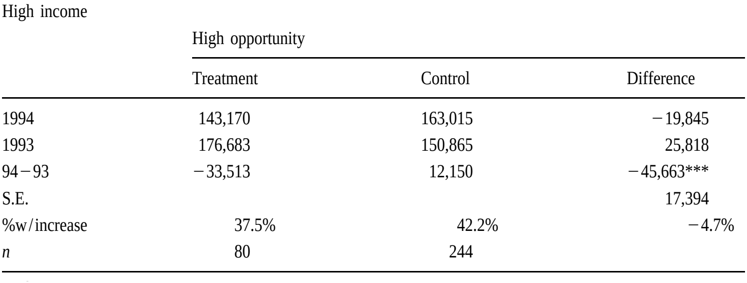
Results for Low Income/High Opportunity Group



Low income

	High opportunity		
	Treatment	Control	Difference
1994	7473	3992	3481
1993	971	787	183
94 − 93	6502	3204	3298
S.E.			2718
%w/increase	65.4%	51.2%	14.2%*
n	52	123	

Results for High Income/High Opportunity Group



High income

	High opportunity		
	Treatment	Control	Difference
1994	143,170	163,015	−19,845
1993	176,683	150,865	25,818
94−93	−33,513	12,150	−45,663***
S.E.			17,394
%w/increase	37.5%	42.2%	−4.7%
n	80	244	

Outline

1. Hypothesis Testing in Small Samples

2. Using Covariates in Experiments

3. Examples
   - Social Pressure Experiment
   - Tax Compliance
   - Labor Market Discrimination
   - Job Training Partnership Act (JTPA)

4. Threats to Validity and Ethics

CV Experiment

- To measure race based labor market discrimination Bertrand and Mullainathan (2004) sent fictional resumes to help-wanted ads in Boston and Chicago newspapers

CV Experiment

- To measure race based labor market discrimination Bertrand and Mullainathan (2004) sent fictional resumes to help-wanted ads in Boston and Chicago newspapers
- Sample: 1,300 employments ads in sales, administrative support, customer service job categories

CV Experiment

- To measure race based labor market discrimination Bertrand and Mullainathan (2004) sent fictional resumes to help-wanted ads in Boston and Chicago newspapers
- Sample: 1,300 employments ads in sales, administrative support, customer service job categories
- D: to manipulate perceived race, otherwise identical resumes are randomly assigned African-American sounding names (Lakisha, Jamal, etc.) or White sounding names (Emily, Greg, etc.)
- Four CVs are send to each ad (two high- and two low-quality resumes, one of each CV is treatment/control)
- Y: callback rates

Example: CV Experiment Results

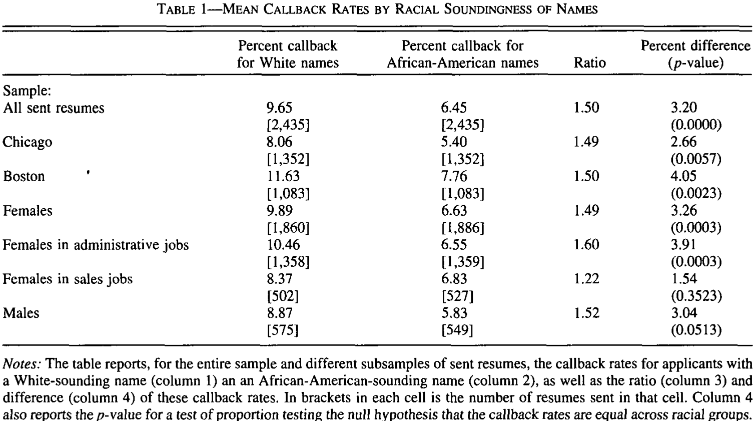


TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65 [2,435]	6.45 [2,435]	1.50	3.20 (0.0000)
Chicago	8.06 [1,352]	5.40 [1,352]	1.49	2.66 (0.0057)
Boston	11.63 [1,083]	7.76 [1,083]	1.50	4.05 (0.0023)
Females	9.89 [1,860]	6.63 [1,886]	1.49	3.26 (0.0003)
Females in administrative jobs	10.46 [1,358]	6.55 [1,359]	1.60	3.91 (0.0003)
Females in sales jobs	8.37 [502]	6.83 [527]	1.22	1.54 (0.3523)
Males	8.87 [575]	5.83 [549]	1.52	3.04 (0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the p-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

Outline

1. Hypothesis Testing in Small Samples
2. Using Covariates in Experiments
3. Examples
   - Social Pressure Experiment
   - Tax Compliance
   - Labor Market Discrimination
   - Job Training Partnership Act (JTPA)
4. Threats to Validity and Ethics

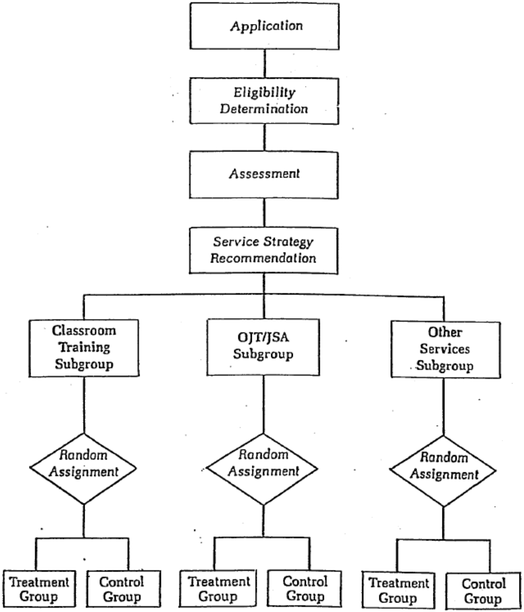
Job Training Partnership Act (JTPA): Design

- Largest randomized training evaluation ever undertaken in the U.S.; started in 1983 at 649 sites throughout the country
- Sample: Disadvantaged persons in the labor market (previously unemployed or low earnings)

Job Training Partnership Act (JTPA): Design

- Largest randomized training evaluation ever undertaken in the U.S.; started in 1983 at 649 sites throughout the country
- Sample: Disadvantaged persons in the labor market (previously unemployed or low earnings)
- D: Assignment to one of three general service strategies
    - classroom training in occupational skills
    - on-the-job training and/or job search assistance
    - other services (eg. probationary employment)
- Y: Earnings 30 months following assignment
- X: Characteristics measured before assignment (age, gender, previous earnings, race, etc.)

Random Assignment Model for JTPA Experiment


Application

Eligibility Determination

Assessment

Service Strategy Recommendation

Classroom Training Subgroup

OJT/JSA Subgroup

Other Services Subgroup

Random Assignment

Random Assignment

Random Assignment

Treatment Group

Control Group

Treatment Group

Control Group

Treatment Group

Control Group

Subgroup Effects for JTPA Experiment

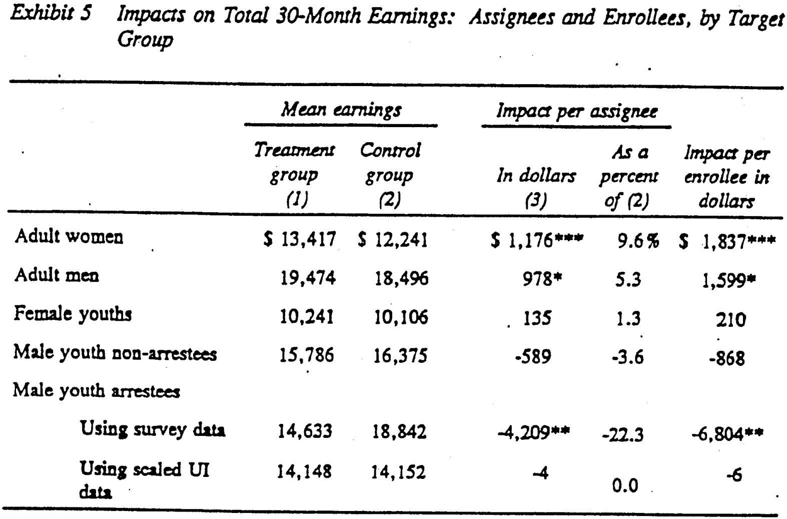


Exhibit 5 *Impacts on Total 30-Month Earnings: Assignees and Enrollees, by Target Group*

	Mean earnings		*Impact per assignee*		
	Treatment group (1)	*Control group (2)*	*In dollars (3)*	*As a percent of (2)*	*Impact per enrollee in dollars*
Adult women	$ 13,417	$ 12,241	$ 1,176***	9.6%	$ 1,837***
Adult men	19,474	18,496	978*	5.3	1,599*
Female youths	10,241	10,106	135	1.3	210
Male youth non-arrestees	15,786	16,375	-589	-3.6	-868
Male youth arrestees					
Using survey data	14,633	18,842	-4,209**	-22.3	-6,804**
Using scaled UI data	14,148	14,152	-4	0.0	-6

A Word about Policy Implications

After the results of the National JTPA study were released, in 1994, funding for JTPA training for the youth were drastically cut:

Spending on JTPA Programs

Year	Youth Training Grants	Adult Training Grants
1993	677	1015
1994	609	988
1995	127	996
1996	127	850
1997	127	895

Outline

1. Hypothesis Testing in Small Samples

2. Using Covariates in Experiments

3. Examples
   - Social Pressure Experiment
   - Tax Compliance
   - Labor Market Discrimination
   - Job Training Partnership Act (JTPA)

4. Threats to Validity and Ethics

Threats to Internal and External Validity

- Internal validity: can we estimate the treatment effect for our particular sample?
    - Fails when there are differences between treated and controls (other than the treatment itself) that affect the outcome and that we cannot control for

Threats to Internal and External Validity

- Internal validity: can we estimate the treatment effect for our particular sample?
    - Fails when there are differences between treated and controls (other than the treatment itself) that affect the outcome and that we cannot control for

- External validity: can we extrapolate our estimates to other populations?
    - Fails when outside the experimental environment the treatment has a different effect

Most Common Threats to Internal Validity

- Failure of randomization
  - E.g. implementing partners assign their favorites to treatment group, small samples, etc.
    - JTPA: Good balance

Most Common Threats to Internal Validity

- Failure of randomization
  - E.g. implementing partners assign their favorites to treatment group, small samples, etc.
    - JTPA: Good balance
- Non-compliance with experimental protocol
  - Failure to treat or "crossover": Some members of the control group receive the treatment and some in the treatment group go untreated
  - Can reduce power significantly
    - JTPA: only about 65% of those assigned to treatment actually enrolled in training (compliance was almost perfect in the control group)

Most Common Threats to Internal Validity

- Failure of randomization
    - E.g. implementing partners assign their favorites to treatment group, small samples, etc.
        - JTPA: Good balance

- Non-compliance with experimental protocol
    - Failure to treat or "crossover": Some members of the control group receive the treatment and some in the treatment group go untreated
    - Can reduce power significantly
        - JTPA: only about 65% of those assigned to treatment actually enrolled in training (compliance was almost perfect in the control group)

- Attrition
    - Can destroy validity if observed potential outcomes are not representative of all potential outcomes even with randomization
    - E.g. control group subjects are more likely to drop out of a study
        - JTPA: only 3 percent dropped out

Most Common Threats to Internal Validity

- Failure of randomization
  - E.g. implementing partners assign their favorites to treatment group, small samples, etc.
    - JTPA: Good balance
- Non-compliance with experimental protocol
  - Failure to treat or "crossover": Some members of the control group receive the treatment and some in the treatment group go untreated
  - Can reduce power significantly
    - JTPA: only about 65% of those assigned to treatment actually enrolled in training (compliance was almost perfect in the control group)
- Attrition
  - Can destroy validity if observed potential outcomes are not representative of all potential outcomes even with randomization
  - E.g. control group subjects are more likely to drop out of a study
    - JTPA: only 3 percent dropped out
- Spillovers
  - Should be dealt with in the design

Most Common Threats to External Validity

- Non-representative sample
  - E.g. laboratory versus field experimentation
  - Subjects are not the same population that will be subject to the policy, known as “randomization bias”

Most Common Threats to External Validity

- Non-representative sample
  - E.g. laboratory versus field experimentation
  - Subjects are not the same population that will be subject to the policy, known as “randomization bias”
- Non-representative program
  - The treatment differs in actual implementations
  - Scale effects
  - Actual implementations are not randomized (nor full scale)

External Validity? Experimental Sites versus all Sites

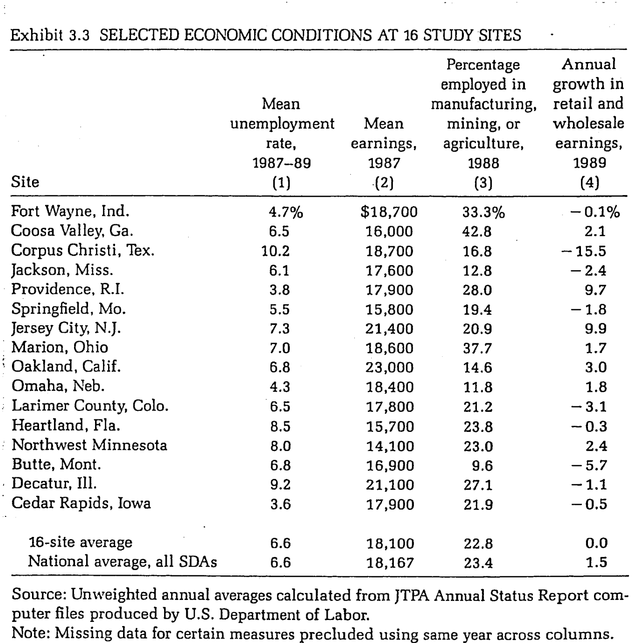


Exhibit 3.3 SELECTED ECONOMIC CONDITIONS AT 16 STUDY SITES

Site	Mean unemployment rate, 1987–89 (1)	Mean earnings, 1987 (2)	Percentage employed in manufacturing, mining, or agriculture, 1988 (3)	Annual growth in retail and wholesale earnings, 1989 (4)
Fort Wayne, Ind.	4.7%	$18,700	33.3%	−0.1%
Coosa Valley, Ga.	6.5	16,000	42.8	2.1
Corpus Christi, Tex.	10.2	18,700	16.8	−15.5
Jackson, Miss.	6.1	17,600	12.8	−2.4
Providence, R.I.	3.8	17,900	28.0	9.7
Springfield, Mo.	5.5	15,800	19.4	−1.8
Jersey City, N.J.	7.3	21,400	20.9	9.9
Marion, Ohio	7.0	18,600	37.7	1.7
Oakland, Calif.	6.8	23,000	14.6	3.0
Omaha, Neb.	4.3	18,400	11.8	1.8
Larimer County, Colo.	6.5	17,800	21.2	−3.1
Heartland, Fla.	8.5	15,700	23.8	−0.3
Northwest Minnesota	8.0	14,100	23.0	2.4
Butte, Mont.	6.8	16,900	9.6	−5.7
Decatur, Ill.	9.2	21,100	27.1	−1.1
Cedar Rapids, Iowa	3.6	17,900	21.9	−0.5
16-site average	6.6	18,100	22.8	0.0
National average, all SDAs	6.6	18,167	23.4	1.5

Source: Unweighted annual averages calculated from JTPA Annual Status Report computer files produced by U.S. Department of Labor.

Note: Missing data for certain measures precluded using same year across columns.

Internal vs. External Validity

Which one is more important?

> One common view is that internal validity comes first. If you do not know the effects of the treatment on the units in your study, you are not well-positioned to infer the effects on units you did not study who live in circumstances you did not study. (Rosenbaum 2010, p. 56)

Internal vs. External Validity

Which one is more important?

> One common view is that internal validity comes first. If you do not know the effects of the treatment on the units in your study, you are not well-positioned to infer the effects on units you did not study who live in circumstances you did not study. (Rosenbaum 2010, p. 56)

Randomization addresses internal validity. External validity is often addressed by comparing the results of several internally valid studies conducted in different circumstances and at different times.

The same issues apply in observational studies.

Hardwork is in the Design and Implementation

- Statistics are often easy; the implementation and design are often hard.

Hardwork is in the Design and Implementation

- Statistics are often easy; the implementation and design are often hard.
- Find partners, manage relationships, identify learning opportunities.

Hardwork is in the Design and Implementation

- Statistics are often easy; the implementation and design are often hard.
- Find partners, manage relationships, identify learning opportunities.
- Designing experiments so that they are incentive-compatible:
    - Free “consulting”
    - Allocating limited resources (e.g. excessively large target groups)
    - Phased randomization as a way to mitigate ethical concerns with denial of treatment
    - Encouragement designs
    - Monitoring

Hardwork is in the Design and Implementation

- Statistics are often easy; the implementation and design are often hard.
- Find partners, manage relationships, identify learning opportunities.
- Designing experiments so that they are incentive-compatible:
    - Free "consulting"
    - Allocating limited resources (e.g. excessively large target groups)
    - Phased randomization as a way to mitigate ethical concerns with denial of treatment
    - Encouragement designs
    - Monitoring
- Potentially high costs.

Hardwork is in the Design and Implementation

- Statistics are often easy; the implementation and design are often hard.
- Find partners, manage relationships, identify learning opportunities.
- Designing experiments so that they are incentive-compatible:
  - Free “consulting”
  - Allocating limited resources (e.g. excessively large target groups)
  - Phased randomization as a way to mitigate ethical concerns with denial of treatment
  - Encouragement designs
  - Monitoring
- Potentially high costs.
- Many things can go wrong with complex and large scale experiments.
- Keep it simple in the field!

Ethics and Experimentation

- Fearon, Humphreys, and Weinstein (2009) used a field experiment to examine if community-driven reconstruction programs foster social reconciliation in post-conflict Liberian villages.
- Outcome: funding raised for collective projects in public goods game played with 24 villagers. Total payout to village is publicly announced.

Ethics and Experimentation

- Fearon, Humphreys, and Weinstein (2009) used a field experiment to examine if community-driven reconstruction programs foster social reconciliation in post-conflict Liberian villages.
- Outcome: funding raised for collective projects in public goods game played with 24 villagers. Total payout to village is publicly announced.

We received a report that leaders in one community had gathered villagers together after we left and asked people to report how much they had contributed. We moved quickly to prevent any retribution in that village, but also decided to alter the protocol for subsequent games to ensure greater protection for game participants.

Ethics and Experimentation

- Fearon, Humphreys, and Weinstein (2009) used a field experiment to examine if community-driven reconstruction programs foster social reconciliation in post-conflict Liberian villages.
- Outcome: funding raised for collective projects in public goods game played with 24 villagers. Total payout to village is publicly announced.

> We received a report that leaders in one community had gathered villagers together after we left and asked people to report how much they had contributed. We moved quickly to prevent any retribution in that village, but also decided to alter the protocol for subsequent games to ensure greater protection for game participants.
>
> These changes included stronger language about the importance of protecting anonymity, random audits of community behavior, facilitation of anonymous reporting of violations of game protocol by participants, and a new opportunity to receive supplemental funds in a postproject lottery if no reports of harassment were received.

Ethics and Experimentation

- **Respect for persons**: Participants in most circumstances must give informed consent.
    - Informed consent often done as part of the baseline survey.
    - If risks are minimal and consent will undermine the study, then informed consent rules can be waived.

Ethics and Experimentation

- **Respect for persons**: Participants in most circumstances must give informed consent.
    - Informed consent often done as part of the baseline survey.
    - If risks are minimal and consent will undermine the study, then informed consent rules can be waived.
- **Beneficence**: Avoid knowingly doing harm. Does not mean that all risk can be eliminated, but possible risks must be balanced against overall benefits to society of the research.
    - Note that the existence of a control group might be construed as denying access to some benefit.
    - But without a control group, generating reliable knowledge about the efficacy of the intervention may be impossible.

Ethics and Experimentation

- **Respect for persons**: Participants in most circumstances must give informed consent.
  - Informed consent often done as part of the baseline survey.
  - If risks are minimal and consent will undermine the study, then informed consent rules can be waived.
- **Beneficence**: Avoid knowingly doing harm. Does not mean that all risk can be eliminated, but possible risks must be balanced against overall benefits to society of the research.
  - Note that the existence of a control group might be construed as denying access to some benefit.
  - But without a control group, generating reliable knowledge about the efficacy of the intervention may be impossible.
- **Justice**: Important to avoid situations where one group disproportionately bears the risks and another stands to received all the benefits.
  - Evaluate interventions that are relevant to the subject population

Ethics and Experimentation

- IRB approval is required in almost all circumstances.
- If running an experiment in another country, you need to follow the local regulations on experimental research.
  - Often poorly adapted to social science.
  - Or legally murky whether or not approval is required.
- Still many unanswered questions and lack of consensus on the ethics of field experimentation within the social sciences!
  - Be prepared to confront wildly varying opinions on these issues.