# Potential outcomes & threats to validity

Fill out your reading report on iCollege!

**February 19, 2020**

PMAP 8521: Program Evaluation for Public Service
Andrew Young School of Policy Studies
Spring 2020

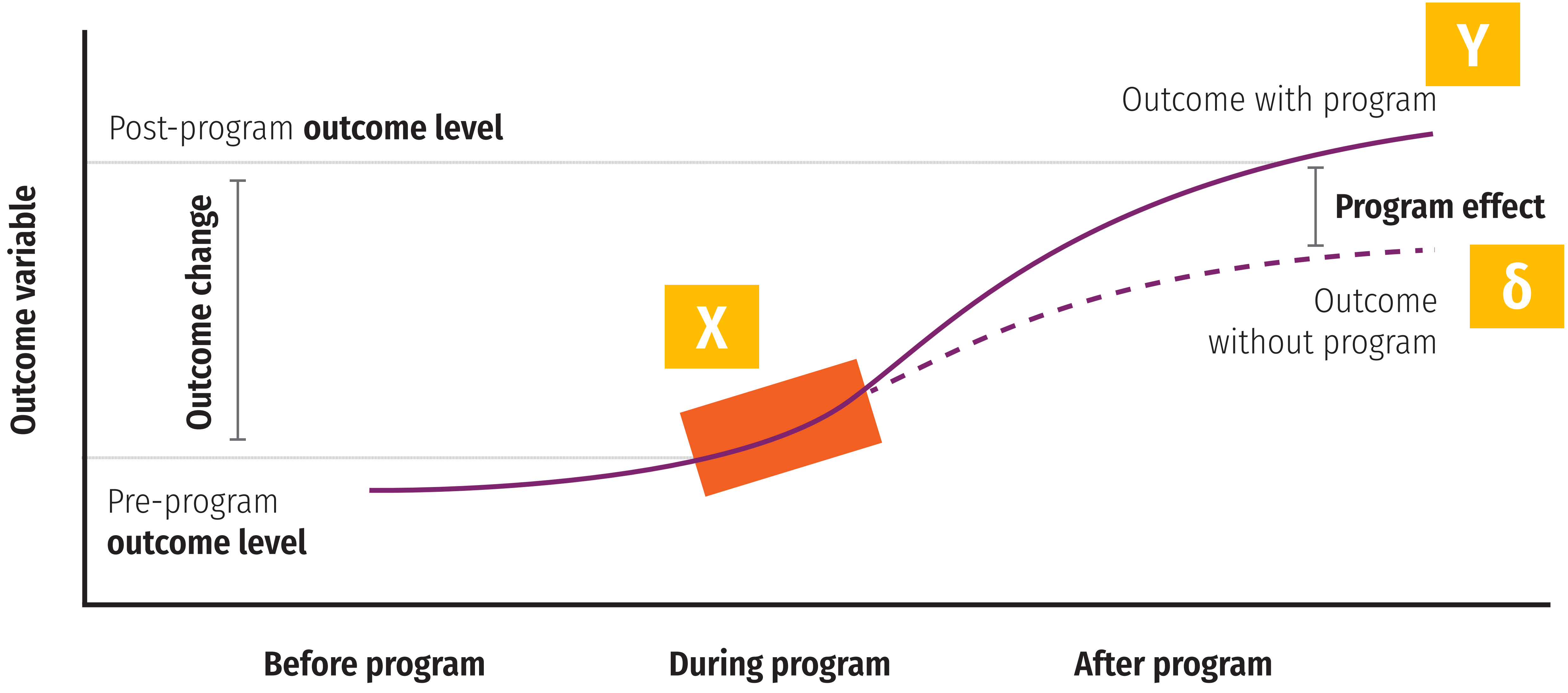
# Plan for today

Potential outcomes

The Four Horsemen of Validity

# Potential outcomes

# Program effect



# Some equation translations

P = probability distribution

$$\delta = P(Y|do(X))$$

E = expected value, or average

$$\delta = E(Y|do(X)) - E(Y|!do(X))$$

$$\delta = (Y|X = 1) - (Y|X = 0)$$

$$\delta = Y_1 - Y_0$$

# Fundamental problem of causal inference

$$\delta_i = \cancel{Y_i^1} - Y_i^0$$

**Individual-level effects are impossible to observe!**

**No individual counterfactuals!**

# Average treatment effect (ATE)

## Solution: Use averages instead

$$ATE = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$$

**Difference between average/expected value when program is on vs. expected value when program is off**

$$\delta = (\bar{Y} | P = 1) - (\bar{Y} | P = 0)$$

| Person | Sex | Treated? | Outcome with program | Outcome without program |
|---|---|---|---|---|
| 1 | M | TRUE | 80 | 60 |
| 2 | M | TRUE | 75 | 70 |
| 3 | M | TRUE | 85 | 80 |
| 4 | M | FALSE | 70 | 60 |
| 5 | F | TRUE | 75 | 70 |
| 6 | F | FALSE | 80 | 80 |
| 7 | F | FALSE | 90 | 100 |
| 8 | F | FALSE | 85 | 80 |

| Person | Sex | Treated? | Outcome with program | Outcome without program | Effect |
|---|---|---|---|---|---|
| 1 | M | TRUE | 80 | 60 | **20** |
| 2 | M | TRUE | 75 | 70 | **5** |
| 3 | M | TRUE | 85 | 80 | **5** |
| 4 | M | FALSE | 70 | 60 | **10** |
| 5 | F | TRUE | 75 | 70 | **5** |
| 6 | F | FALSE | 80 | 80 | **0** |
| 7 | F | FALSE | 90 | 100 | **−10** |
| 8 | F | FALSE | 85 | 80 | **5** |

$$\delta = (\bar{Y}|P = 1) - (\bar{Y}|P = 0)$$

**ATE =** **5**

# Conditional ATE (CATE)

ATE in subgroups

**Is the program more effective for specific sexes?**

| Person | Sex | Treated? | Outcome with program | Outcome without program | Effect |
|---|---|---|---|---|---|
| 1 | M | TRUE | 80 | 60 | **20** |
| 2 | M | TRUE | 75 | 70 | **5** |
| 3 | M | TRUE | 85 | 80 | **5** |
| 4 | M | FALSE | 70 | 60 | **10** |
| 5 | F | TRUE | 75 | 70 | **5** |
| 6 | F | FALSE | 80 | 80 | **0** |
| 7 | F | FALSE | 90 | 100 | **−10** |
| 8 | F | FALSE | 85 | 80 | **5** |

$$\delta = (\bar{Y}_{\text{Male}} | P = 1) - (\bar{Y}_{\text{Male}} | P = 0)$$

**CATE$_{\text{Male}}$ = 10**

$$\delta = (\bar{Y}_{\text{Female}} | P = 1) - (\bar{Y}_{\text{Female}} | P = 0)$$

**CATE$_{\text{Female}}$ = 0**

# ATT & ATU

## Average treatment on the treated

**ATT / TOT**

**Effect for those with treatment**

## Average treatment on the untreated

**ATU / TUT**

**Effect for those with without treatment**

| Person | Sex | Treated? | Outcome with program | Outcome without program | Effect |
|---|---|---|---|---|---|
| 1 | M | TRUE | 80 | 60 | **20** |
| 2 | M | TRUE | 75 | 70 | **5** |
| 3 | M | TRUE | 85 | 80 | **5** |
| 4 | M | FALSE | 70 | 60 | **10** |
| 5 | F | TRUE | 75 | 70 | **5** |
| 6 | F | FALSE | 80 | 80 | **0** |
| 7 | F | FALSE | 90 | 100 | **−10** |
| 8 | F | FALSE | 85 | 80 | **5** |

$$\delta = (\bar{Y}_{\text{Treated}} | P = 1) - (\bar{Y}_{\text{Treated}} | P = 0)$$

**ATT = 8.75**

$$\delta = (\bar{Y}_{\text{Untreated}} | P = 1) - (\bar{Y}_{\text{Untreated}} | P = 0)$$

**ATU = 1.25**

# ATE, ATT, & ATU

**The ATE is the weighted average of ATT and ATU**

$$(8.75 \times 4/8) + (1.25 \times 4/8)$$

$$4.375 + 0.625$$

$$5$$

# Selection bias

ATE and ATT aren't always the same

ATE = ATT + Selection bias

$5 = 8.75 + x$

$x = -3.75$

Randomization fixes this, makes $x = 0$

# Actual data

| Person | Sex | Treated? | Actual outcome |
|---|---|---|---|
| 1 | M | TRUE | 80 |
| 2 | M | TRUE | 75 |
| 3 | M | TRUE | 85 |
| 4 | M | FALSE | 60 |
| 5 | F | TRUE | 75 |
| 6 | F | FALSE | 80 |
| 7 | F | FALSE | 100 |
| 8 | F | FALSE | 80 |

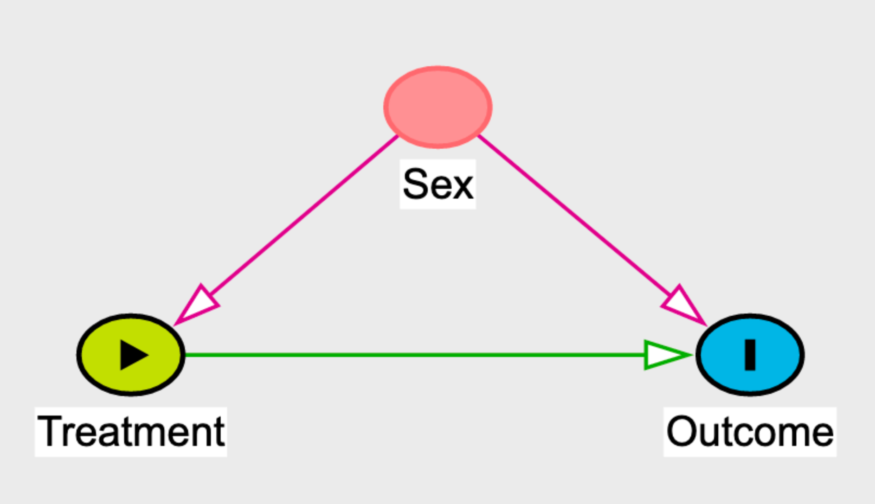
**Treatment not randomly assigned**

**We can't see unit-level causal effects**

# Actual data

| Person | Sex | Treated? | Actual outcome |
|---|---|---|---|
| 1 | M | TRUE | 80 |
| 2 | M | TRUE | 75 |
| 3 | M | TRUE | 85 |
| 4 | M | FALSE | 60 |
| 5 | F | TRUE | 75 |
| 6 | F | FALSE | 80 |
| 7 | F | FALSE | 100 |
| 8 | F | FALSE | 80 |

**Treatment seems to be correlated with sex**



# Actual data

| Person | Sex | Treated? | Actual outcome |
|---|---|---|---|
| 1 | M | TRUE | 80 |
| 2 | M | TRUE | 75 |
| 3 | M | TRUE | 85 |
| 4 | M | FALSE | 60 |
| 5 | F | TRUE | 75 |
| 6 | F | FALSE | 80 |
| 7 | F | FALSE | 100 |
| 8 | F | FALSE | 80 |

**We can estimate ATE by finding weighted average of sex-based CATEs**

**As long as we assume/pretend treatment was randomly assigned within each sex = unconfoundedness**

$$\widehat{\text{ATE}} = \pi_{\text{Male}} \widehat{\text{CATE}}_{\text{Male}} + \pi_{\text{Female}} \widehat{\text{CATE}}_{\text{Female}}$$

# Actual data

| Person | Sex | Treated? | Actual outcome |
|---|---|---|---|
| 1 | M | TRUE | 80 |
| 2 | M | TRUE | 75 |
| 3 | M | TRUE | 85 |
| 4 | M | FALSE | 60 |
| 5 | F | TRUE | 75 |
| 6 | F | FALSE | 80 |
| 7 | F | FALSE | 100 |
| 8 | F | FALSE | 80 |

$\text{CATE}_{\text{Male}}$ = 20

$\text{CATE}_{\text{Female}}$ = −11.67

ATE = 4.16

$$\widehat{\text{ATE}} = \pi_{\text{Male}} \widehat{\text{CATE}}_{\text{Male}} + \pi_{\text{Female}} \widehat{\text{CATE}}_{\text{Female}}$$

# DON'T DO THIS

| Person | Sex | Treated? | Actual outcome |
|---|---|---|---|
| 1 | M | TRUE | 80 |
| 2 | M | TRUE | 75 |
| 3 | M | TRUE | 85 |
| 4 | M | FALSE | 60 |
| 5 | F | TRUE | 75 |
| 6 | F | FALSE | 80 |
| 7 | F | FALSE | 100 |
| 8 | F | FALSE | 80 |

$\text{CATE}_{\text{Treated}}$ = 78.75

$\text{CATE}_{\text{Untreated}}$ = 80

ATE = −1.25

**Only do this if treatment is random!**

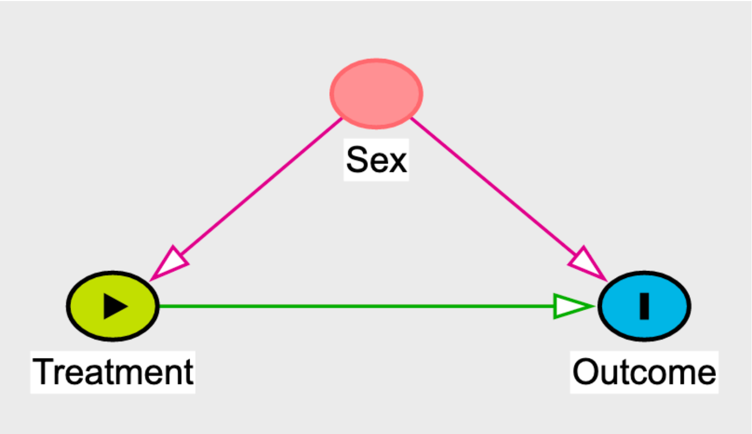
$$\widehat{\text{ATE}} = \widehat{\text{CATE}}_{\text{Treated}} - \widehat{\text{CATE}}_{\text{Untreated}}$$

# Matching and ATEs

$$\widehat{\text{ATE}} = \pi_{\text{Male}}\widehat{\text{CATE}}_{\text{Male}} + \pi_{\text{Female}}\widehat{\text{CATE}}_{\text{Female}}$$

**We chose sex here because it correlates with (and confounds) the outcome**

And we assumed unfoundedness;
that treatment is randomly assigned within the groups



**Does attending a private university cause an increase in earnings?**

TABLE 2.1
The college matching matrix

| | | Private | | | Public | | | |
|---|---|---|---|---|---|---|---|---|
| Applicant group | Student | Ivy | Leafy | Smart | All State | Tall State | Altered State | 1996 earnings |
| A | 1 | | Reject | Admit | | Admit | | 110,000 |
| | 2 | | Reject | Admit | | Admit | | 100,000 |
| | 3 | | Reject | Admit | | Admit | | 110,000 |
| B | 4 | Admit | | | Admit | | Admit | 60,000 |
| | 5 | Admit | | | Admit | | Admit | 30,000 |
| C | 6 | | Admit | | | | | 115,000 |
| | 7 | | Admit | | | | | 75,000 |
| D | 8 | Reject | | | Admit | Admit | | 90,000 |
| | 9 | Reject | | | Admit | Admit | | 60,000 |

*Note:* Enrollment decisions are highlighted in gray.

| Applicant group | Student | Private: Ivy | Private: Leafy | Private: Smart | Public: All State | Public: Tall State | Public: Altered State | 1996 earnings |
|---|---|---|---|---|---|---|---|---|
| A | 1 | | Reject | Admit | | Admit | | 110,000 |
| | 2 | | Reject | Admit | | Admit | | 100,000 |
| | 3 | | Reject | Admit | | Admit | | 110,000 |
| B | 4 | Admit | | | Admit | | Admit | 60,000 |
| | 5 | Admit | | | Admit | | Admit | 30,000 |
| C | 6 | | Admit | | | | | 115,000 |
| | 7 | | Admit | | | | | 75,000 |
| D | 8 | Reject | | | Admit | Admit | | 90,000 |
| | 9 | Reject | | | Admit | Admit | | 60,000 |

*Note:* Enrollment decisions are highlighted in gray.

**Average private − Average public**

(110,000 + 100,000 + 60,000 + 115,000 + 75,000) / 5 = $92,000

(110,000 + 30,000 + 90,000 + 60,000) / 4 = $72,500

($92,500 × 5/9) − ($72,500 × 4/9) = $19,166.67

**This is wrong!**

$$\widehat{\text{ATE}} = \pi_{\text{Private}} \widehat{\text{CATE}}_{\text{Private}} - \pi_{\text{Public}} \widehat{\text{CATE}}_{\text{Public}}$$

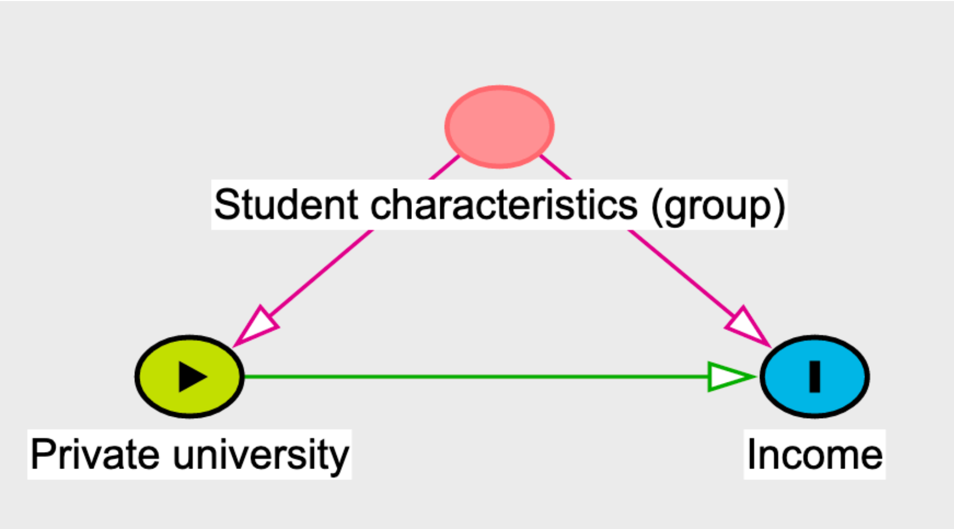
# Grouping and matching

| | | Private | | | Public | | | |
|---|---|---|---|---|---|---|---|---|
| Applicant group | Student | Ivy | Leafy | Smart | All State | Tall State | Altered State | 1996 earnings |
| A | 1 | | Reject | Admit | | Admit | | 110,000 |
| | 2 | | Reject | Admit | | Admit | | 100,000 |
| | 3 | | Reject | Admit | | Admit | | 110,000 |
| B | 4 | Admit | | | Admit | | Admit | 60,000 |
| | 5 | Admit | | | Admit | | Admit | 30,000 |
| C | 6 | | Admit | | | | | 115,000 |
| | 7 | | Admit | | | | | 75,000 |
| D | 8 | Reject | | | Admit | Admit | | 90,000 |
| | 9 | Reject | | | Admit | Admit | | 60,000 |

*Note:* Enrollment decisions are highlighted in gray.

**These groups look like they have similar characteristics**

(Unconfoundedness?)



| Applicant group | Student | Private: Ivy | Private: Leafy | Private: Smart | Public: All State | Public: Tall State | Public: Altered State | 1996 earnings |
|---|---|---|---|---|---|---|---|---|
| A | 1 | | Reject | Admit | | Admit | | 110,000 |
| | 2 | | Reject | Admit | | −$5,000 | | 100,000 |
| | 3 | | Reject | Admit | | Admit | | 110,000 |
| B | 4 | Admit | | | | $30,000 | | 60,000 |
| | 5 | Admit | | | | | | 30,000 |
| C | 6 | | Admit | | | ??? | | 115,000 |
| | 7 | | Admit | | | | | 75,000 |
| D | 8 | Reject | | | | ??? | | 90,000 |
| | 9 | Reject | | | | | | 60,000 |

*Note:* Enrollment decisions are highlighted in gray.

(−$5,000 × 3/5) + ($30,000 × 2/5) = $9,000

**This is less wrong!**

$$\widehat{\text{ATE}} = \pi_{\text{Group A}} \widehat{\text{CATE}}_{\text{Group A}} + \pi_{\text{Group B}} \widehat{\text{CATE}}_{\text{Group B}}$$

# Matching with regression

$$\text{earnings} = \alpha + \beta_1 \text{Private} + \beta_2 \text{Group A} + \epsilon$$

```
model_earnings <- lm(Earnings ~ Private + Group A, data = schools)
```

| term | estimate | std_error | statistic | p_value |
|---|---|---|---|---|
| Intercept | 40000 | 11952.29 | 3.3467 | 0.08 |
| Private | **10000** | 13093.07 | 0.7638 | 0.52 |
| Group A | 60000 | 13093.07 | 4.5826 | 0.04 |

$B_1$ = $10,000

This is less wrong!

Significance details!

# The Four Horsemen of Validity

# Threats to validity

Internal validity

External validity

Construct validity

Statistical conclusion validity

# Internal validity

## Omitted variable bias

Selection

Attrition

## Trends

Maturation

Secular trends

Seasonality

Testing

Regression

## Study calibration

Measurement error

Time frame of study

## Contamination

Hawthorne

John Henry

Spillovers

Intervening events

# Selection

If people can choose to enroll in a program, those that enroll will be different than those that do not

**How to fix**

Randomization into treatment and control groups

# Selection

If people can choose *when* to enroll in a program, time might influence the result

**How to fix**

Shift time around

The Journal of Socio-Economics 35 (2006) 326–347

The Journal of Socio-Economics

www.elsevier.com/locate/econbase

# Does marriage make people happy, or do happy people get married?

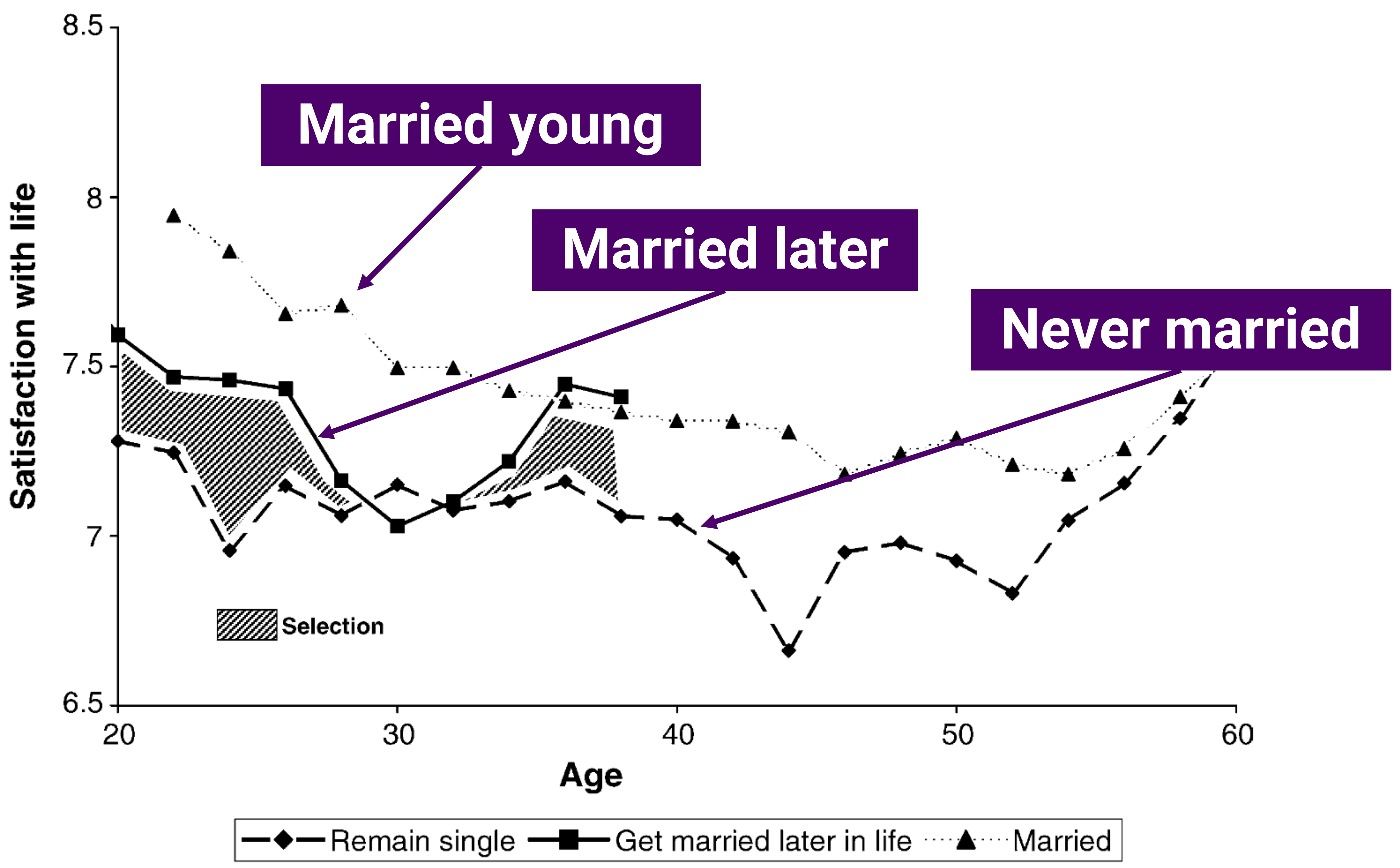
Alois Stutzer *,[1], Bruno S. Frey [1]

*University of Zurich, Switzerland*

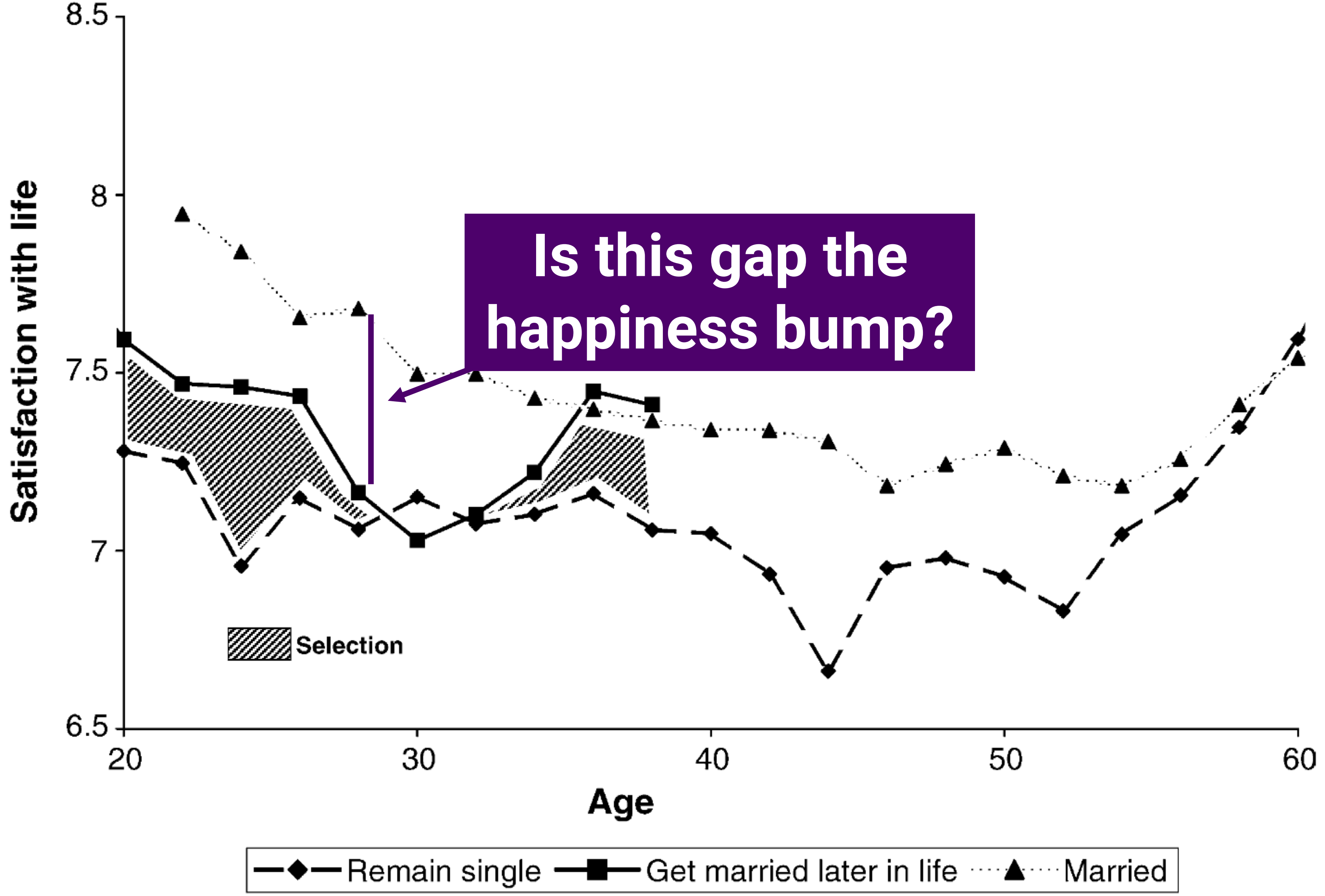
Received 4 June 2003; accepted 12 October 2004

**Abstract**

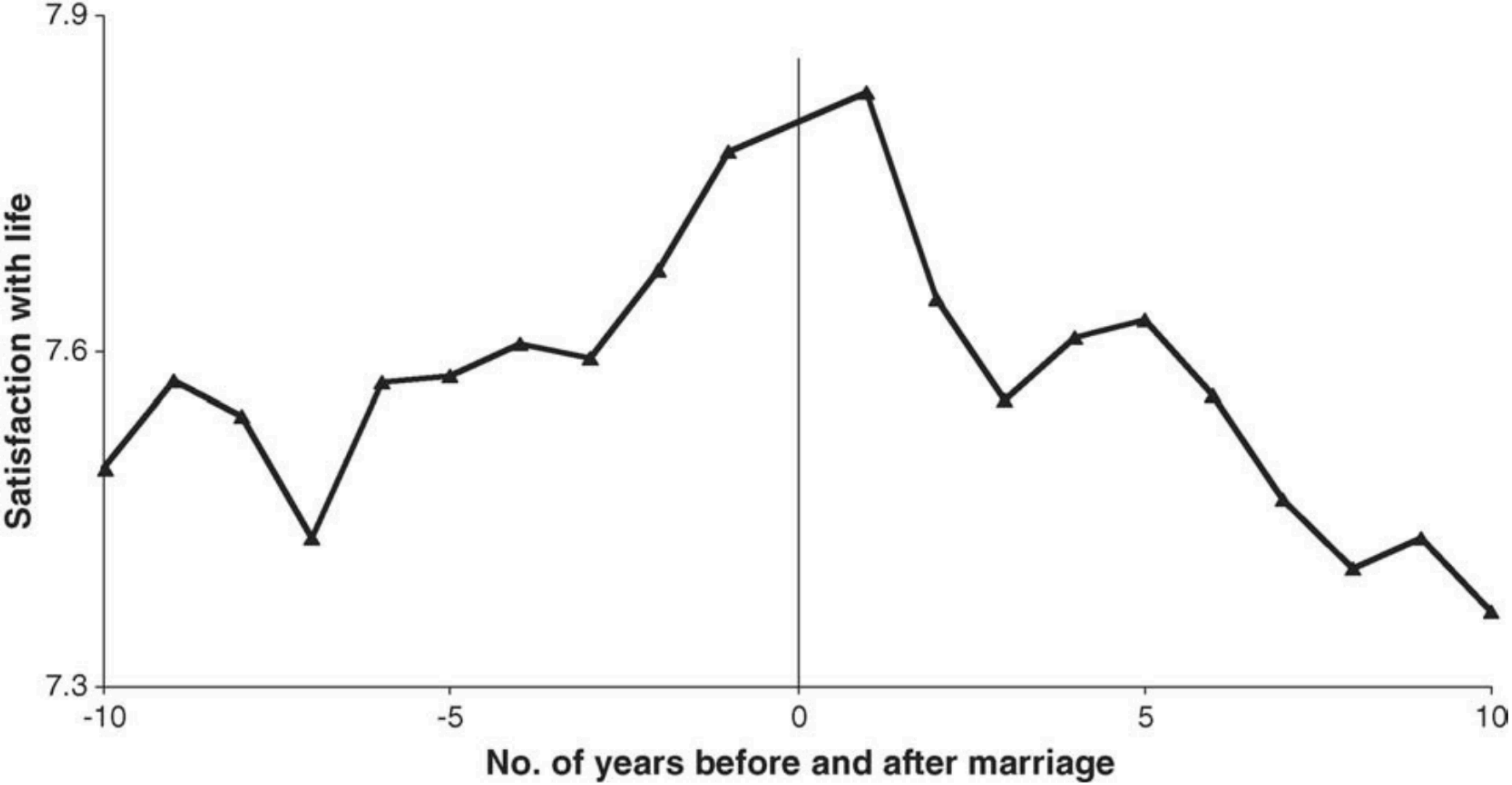
This paper analyzes the causal relationships between marriage and subjective well-being in a longitudinal data set spanning 17 years. We find evidence that happier singles opt more likely for marriage and that there are large differences in the benefits from marriage between couples. Potential, as well as actual, division of labor seems to contribute to spouses' well-being, especially for women and when there is a young family to raise. In contrast, large differences in the partners' educational level have a negative effect on experienced life satisfaction.



Married young
Married later
Never married
Satisfaction with life
8.5
8
7.5
7
6.5
20
30
40
50
60
Age
Selection
Remain single
Get married later in life
Married



Is this gap the happiness bump?
Satisfaction with life
8.5
8
7.5
7
6.5
20
30
40
50
60
Age
Selection
Remain single
Get married later in life
Married


7.9
7.6
7.3
Satisfaction with life
-10
-5
0
5
10
No. of years before and after marriage


Green space and mental health
Dr Ian Alcock
(Epidemiologist)
UNIVERSITY OF EXETER | MEDICAL SCHOOL
https://vimeo.com/83228781

# Attrition

If the people who leave a program or study are different than those that stay, the effects will be biased

**How to fix**

Check characteristics of those that stay and those that leave

# Fake microfinance program results

| ID | Increase in income | Remained in program |
|---|---|---|
| 1 | $3.00 | Yes |
| 2 | $3.50 | Yes |
| 3 | $2.00 | Yes |
| 4 | $1.50 | No |
| 5 | $1.00 | No |

**ATE with attriters = $2.20**

**ATE without attriters = $2.83**

# Maturation

Growth is expected naturally, like checking if a program helps child cognitive ability (Sesame Street)

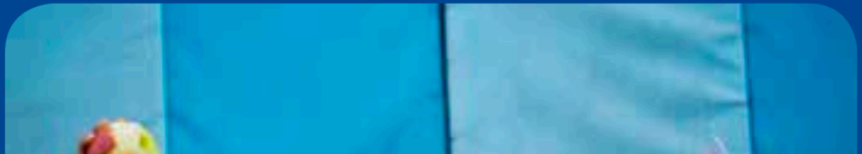
**How to fix**

Use a comparison group to remove the trend

# New Study Finds Sesame Street Improves School Readiness

Research coauthored by Wellesley College economist **Phillip B. Levine** and University of Maryland economist **Melissa Kearney**, finds that greater access to Sesame Street in the show's early days helped children do better in school.

**When Sesame Street first aired in 1969, five million children watched a typical episode. That's the preschool equivalent of a Super Bowl every day.**



# Secular trends

Trends in data are happening because of larger global processes

Recessions

Cultural shifts

Marriage equality

**How to fix**

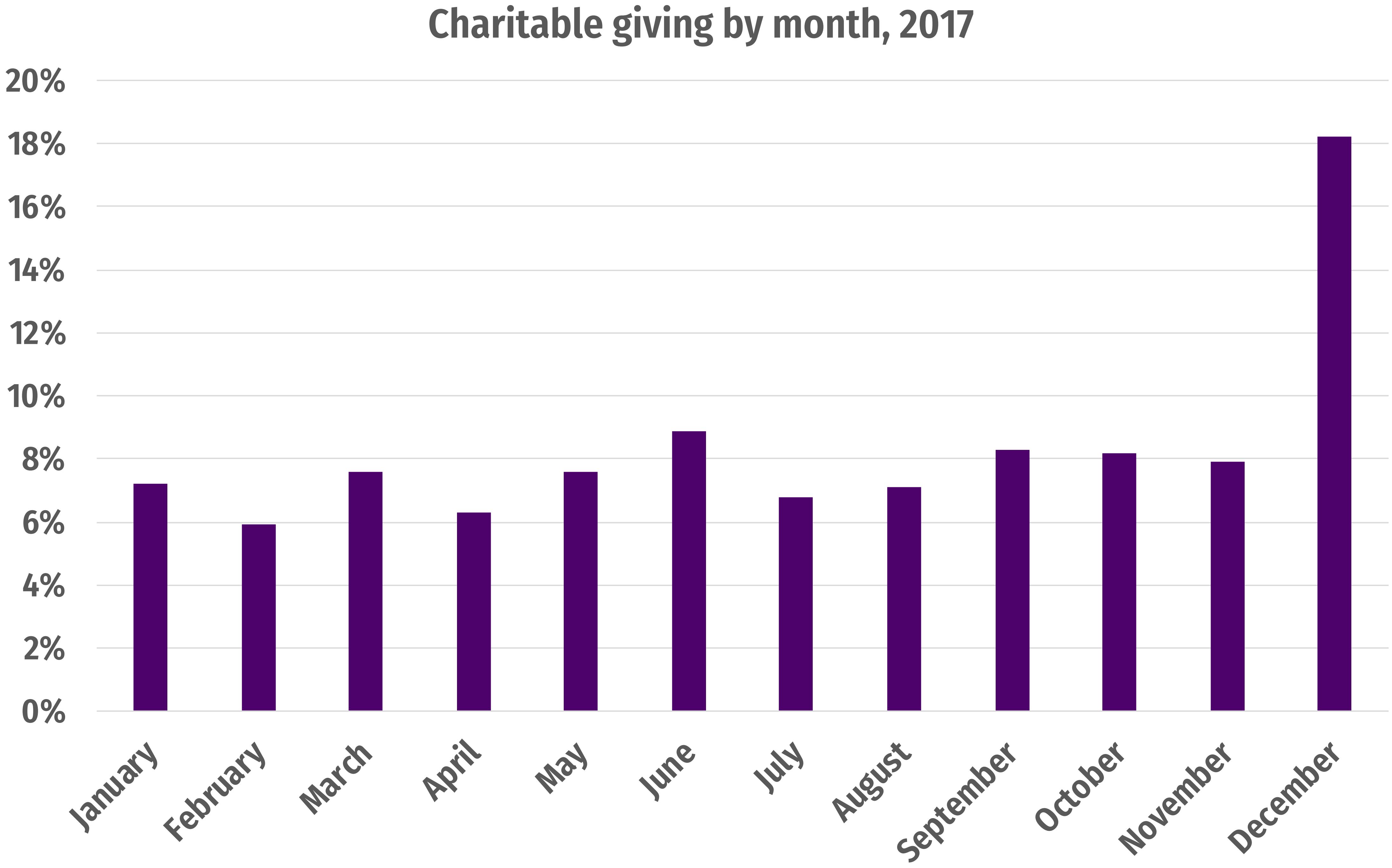
Use a comparison group to remove the trend

# Seasonal trends

Trends in data are happening because of regular time-based trends

**How to fix**

Compare observations from same time period or use yearly/monthly averages


Charitable giving by month, 2017
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
January
February
March
April
May
June
July
August
September
October
November
December

# Testing

Repeated exposure to questions or tasks will make people improve

**How to fix**

Change tests, don’t offer pre-tests maybe, use a control group that receives the test

# Regression to the mean

People in the extreme have a tendency to become less extreme over time

Luck

Crime and terrorism

Hot hand effect

## How to fix

Don’t select super high or super low performers

# Measurement error

Measuring the outcome incorrectly will mess with effect

How to fix

Measure the outcome well

# Time frame

If the study is too short, the effect might not be detectable yet; if the study is too long, attrition becomes a problem

## How to fix

Use prior knowledge about the thing you're studying to choose the right length

# Hawthorne effect

Observing people makes them behave differently

How to fix

Hide? Use completely unobserved control groups

# John Henry effect

Control group works hard to prove they're as good as the treatment group

## How to fix

Keep two groups separate

# Spillover effect

Control groups naturally pick up what the treatment group is getting

Externalities

Social interaction

Equilibrium effects

## How to fix

Keep two groups separate, use distant control groups

# Reducing Intimate Partner Violence through Informal Social Control: A mass media experiment in rural Uganda

## Research Method

Blocked and clustered field experiment with 6,449 respondents in 112 villages.

## Country

Uganda

## Co-Authors

Donald Green, Anna Wilke

## Partners

Innovations for Poverty Action (IPA Uganda), Peripheral Vision International (PVI)

## Research Question

Can mass media shore up informal channels for reducing intimate partner violence?

## Abstract

We assess a mass media campaign designed to reduce intimate partner violence (IPV). A placebo-controlled experiment conducted in 2016 exposed over 10,000 Ugandans in 112 rural villages to a sequence of three short video dramatizations of IPV. A seemingly unrelated opinion survey conducted eight months later indicates that villages in which IPV videos were aired experienced substantially less IPV in the preceding six months than villages that were shown videos on other topics. A closer look at mechanisms reveals that the IPV videos had little effect on attitudes about the legitimacy of IPV. Nor did the videos increase empathy with IPV victims or change perceptions about whether domestic violence must be stopped before it escalates. The most plausible causal channel appears to be a change in norms: women in the treatment group became less likely to believe that they would be criticized for meddling in the affairs of others if they were to report IPV to local leaders, and their personal willingness to intervene increased substantially. These results suggest that education-entertainment has the potential to markedly reduce the incidence of IPV in an enduring and cost-effective manner.

## Paper

See here for latest working paper.

## Replication Archive

Replication by JPAL underway, data forthcoming.

# Intervening events

Something happens that affects one of the groups and not the other

**How to fix**

¯_(ツ)_/¯

# Internal validity

## Omitted variable bias

Selection

Attrition

## Trends

Maturation

Secular trends

Seasonality

Testing

Regression

## Study calibration

Measurement error

Time frame of study

## Contamination

Hawthorne

John Henry

Spillovers

Intervening events

# Fixing internal validity

## Randomization fixes a host of big issues

Selection | Maturation | Regression to the mean

## Randomization doesn’t fix everything!

Attrition | Contamination | Measurement

# External validity

## Findings are generalizable to the entire universe or population

### Hospital lights increase risk of dying in patients with heart disease

Sunday, September 01, 2019 by: Melissa Smith
*Tags: brain inflammation, Cardiac Arrest, cardiovascular disease, death, dim light, heart disease, heart health, hospital lights, hospital rooms, Hospitals, lighting, lights, mortality, research, white light*

**5,900**
VIEWS

**justsaysinmice**
@justsaysinmice

IN MICE

**Hospital lights increase risk of dying in patients with heart disease**
Hospitals may want to consider changing the lights they use in their rooms, especially for patients who suffered a cardiac arrest. A study published in the...
naturalnews.com

10:36 AM · Sep 4, 2019 · Twitter Web App

# External validity

Laboratory conditions vs. real world

Study volunteers are weird

(**W**estern, **e**ducated, from **i**ndustrialized, **r**ich, and **d**emocratic countries)

Not everyone takes surveys

Online surveys

Amazon Mechanical Turk

Random digit dialing

# External validity

## Different circumstances in general

Does a study in one state apply to other states?

Does a mosquito net trial in Eritrea transfer to Bolivia?

# Construct validity

## The Streetlight Effect

# Construct validity

You're measuring the thing you want to measure

Do test scores work for school evaluation?

Test scores measure how good kids are at taking tests

This is why we spent so much time on outcome measurement construction

# Statistical conclusion validity

Are your stats correct?

Statistical power

Violated assumptions of statistical tests

Fishing and p-hacking and error rate problem

If p = 0.05, and you measure 20 outcomes, 1 of those will likely show correlation

# Threats to validity

## Internal validity

Omitted variable bias

Trends

Study calibration

Contamination

## External validity

## Construct validity

## Statistical conclusion validity

# Internal validity

## Omitted variable bias

Selection

Attrition

## Trends

Maturation

Secular trends

Seasonality

Testing

Regression

## Study calibration

Measurement error

Time frame of study

## Contamination

Hawthorne

John Henry

Spillovers

Intervening events