THREATS TO VALIDITY

PMAP 8521: Program Evaluation for Public Service

October 7, 2019

Fill out your reading report on iCollege!

PLAN FOR TODAY

Potential outcomes

The Four Horsemen of Validity

Questions!

POTENTIAL OUTCOMES

POTENTIAL OUTCOMES

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

δ = Causal impact of program

P = Program

Y = Outcome

$$\delta = Y_1 - Y_0$$

Fundamental problem of causal inference

$$\delta_i = Y_i^1 - Y_i^0$$

Individual-level effects are impossible to observe

Average treatment effect

Difference between expected value when program is on vs. expected value when program is off

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

Average treatment effect

Can be found for a whole population, on average

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

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 average treatment effect

CATE

Effect 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

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

$CATE_{Male}$ = 10

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

$CATE_{Female}$ = 0

Average treatment on the treated

ATT / TOT

Effect for those with treatment

Average treatment on the untreated

ATU / TUT

Effect for those 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 = weighted average of ATT and ATU

$$(8.75 \times 0.5) + (1.25 \times 0.5)$$

$$4.375 + .625$$

$$5$$

Selection bias

ATT and ATE aren’t always the same

$$\text{ATE} = \text{ATT} + \text{Selection bias}$$

$$5 = 8.75 - x$$

$$x = 3.75$$

Randomization fixes this, makes x = 0

THE FOUR HORSEMEN OF VALIDITY

CAN ANTS COUNT?!

https://www.youtube.com/watch?v=7DDF8WZFnoU

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

TABLE 2

Background Characteristics of Students in Treatment and Control Groups (Total numbers of cases in parentheses)

Characteristic	*All students in the study*			*All students with scores three or four years after application*		
	Choice students	*Control students*	*p value*[a]	*Choice students*	*Control students*	*p value*[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

SELECTION

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

How to fix

Shift time around

ELSEVIER

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?

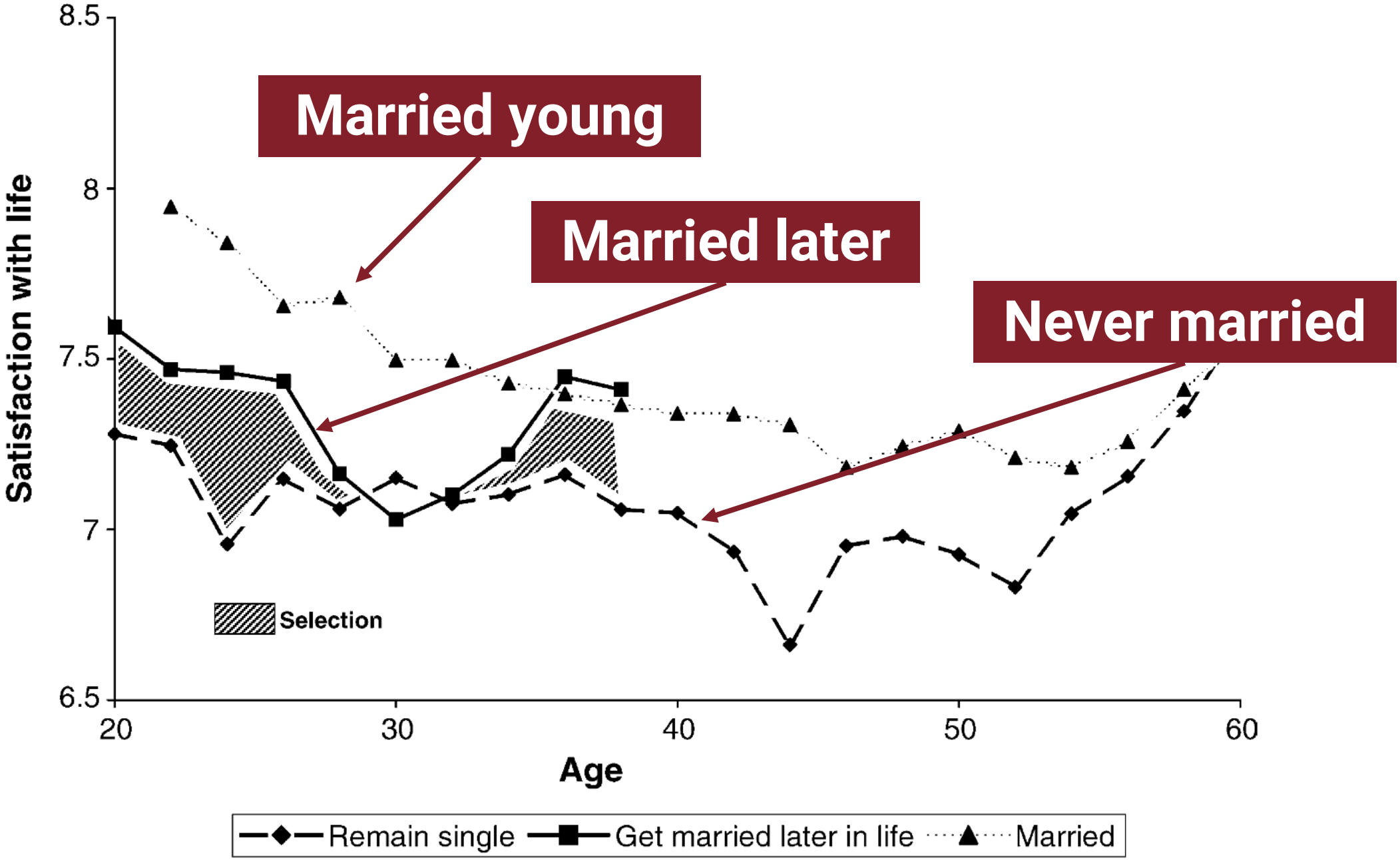
Alois Stutzer *,[1], Bruno S. Frey [1]

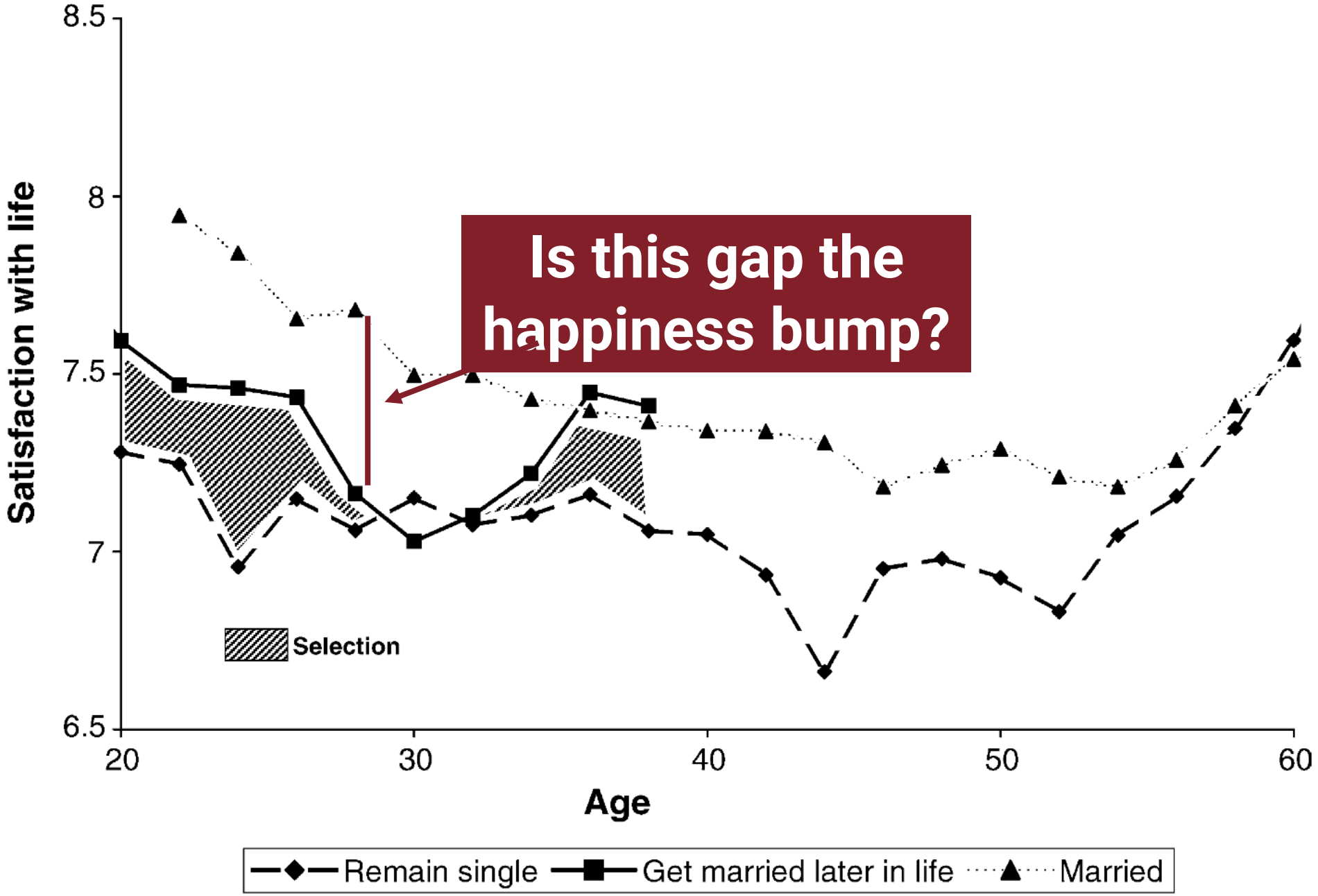
University of Zurich, Switzerland

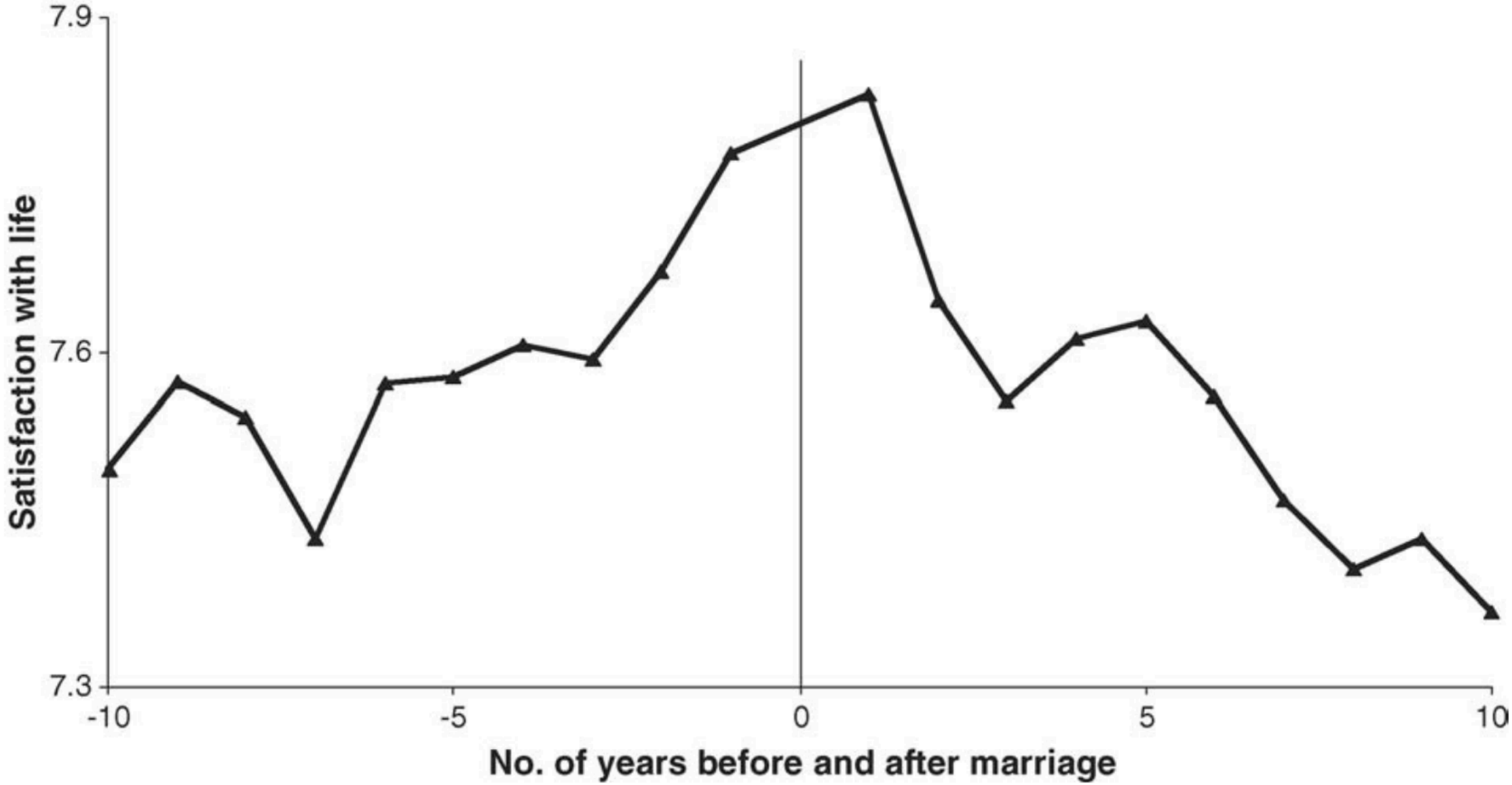
Received 4 June 2003; accepted 12 October 2004

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)

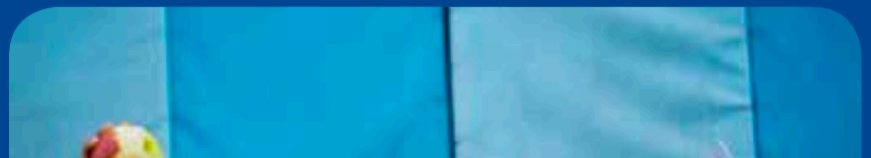
How to fix

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

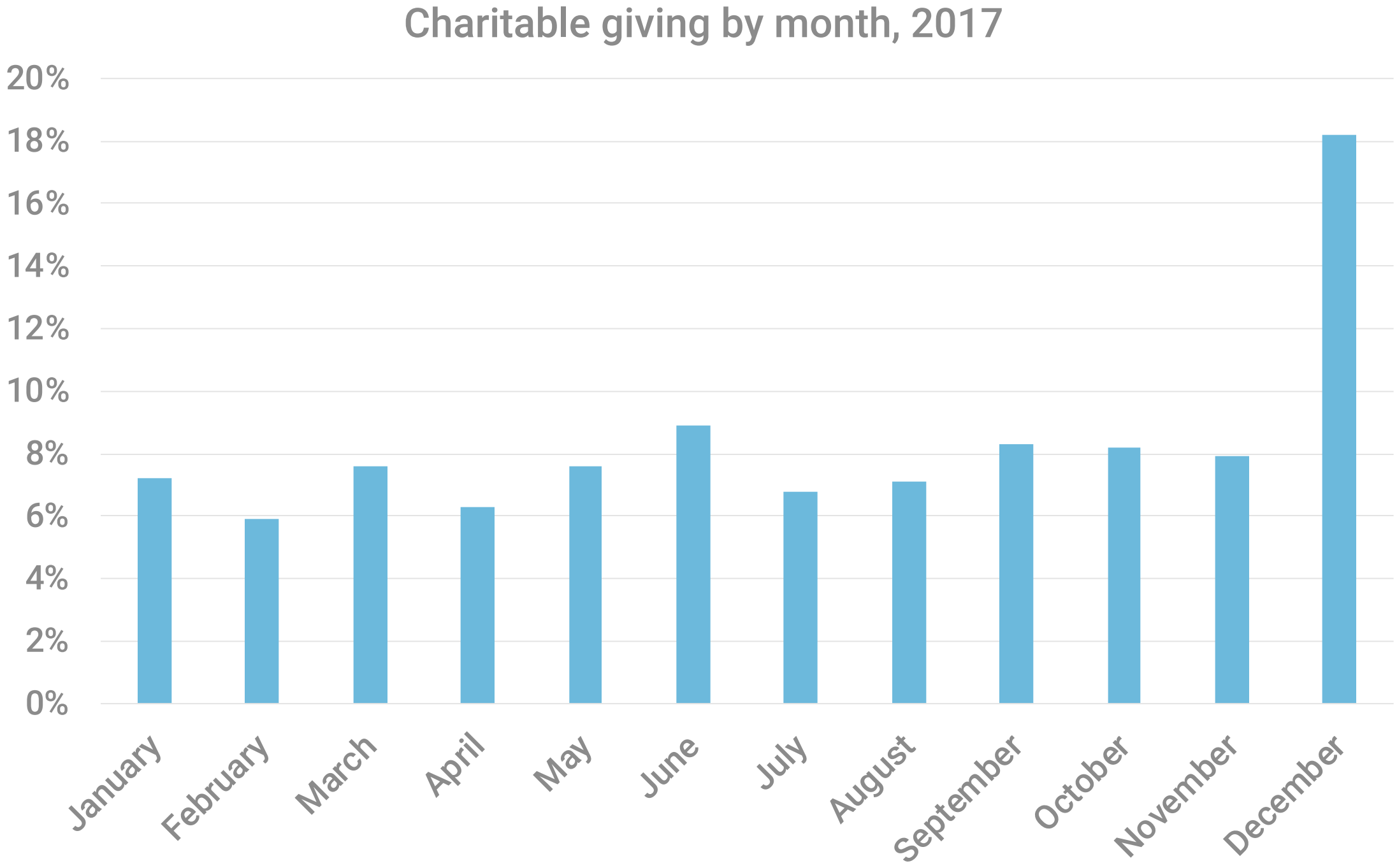
Use a comparison group to remove the trend

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

Your turn!

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

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

QUESTIONS