

# The Four Horsemen of Validity

## Session 6

PMAP 8521: Program evaluation  
Andrew Young School of Policy Studies

# Plan for today

**Construct validity**

**Statistical conclusion validity**

**Internal validity**

**External validity**

# Construct validity

**A new program hopes to  
improve student commitment to school**

**Participants score 200 points higher on the  
SAT and have a 0.3 higher GPA, on average**

**Success!**

**Success?**

# The Streetlight Effect



# Construct validity

**Are you measuring what you want to measure?**

**Do test scores measure commitment to school?  
Teacher performance? Principal skill?**

**Test scores measure how good kids are at taking tests**

**This is why we spend so much time  
on outcome measurement construction!**

# Statistical conclusion validity

# Statistical conclusion validity

**Are your statistics correct?**

**Statistical power**

**Violated assumptions of statistical tests**

**Fishing and p-hacking**

**Spurious statistical significance**



# Power

A training program causes incomes to rise by \$40

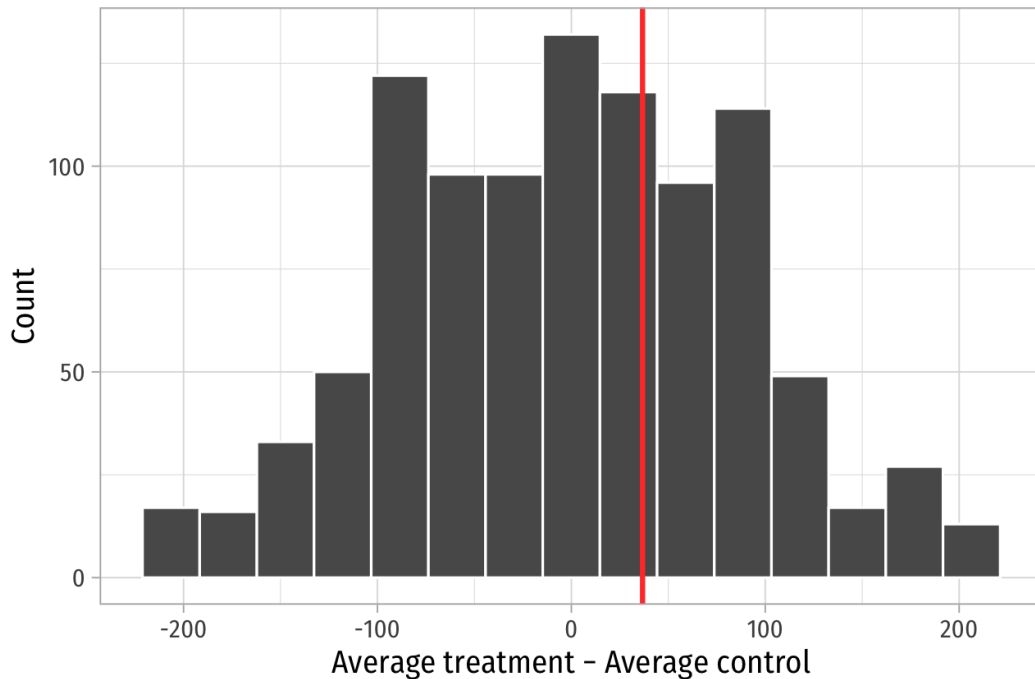
<b>Person</b>	<b>Group</b>	<b>Before</b>	<b>After</b>	<b>Difference</b>
295	Control	122.09	229.04	106.95
126	Treatment	205.60	199.84	-5.76
400	Control	133.25	130.40	-2.85
94	Treatment	270.11	206.56	-63.54
250	Control	344.37	222.89	-121.49
59	Treatment	312.41	268.06	-44.35

# Power

## Survey 10 participants

Simulated world with no difference

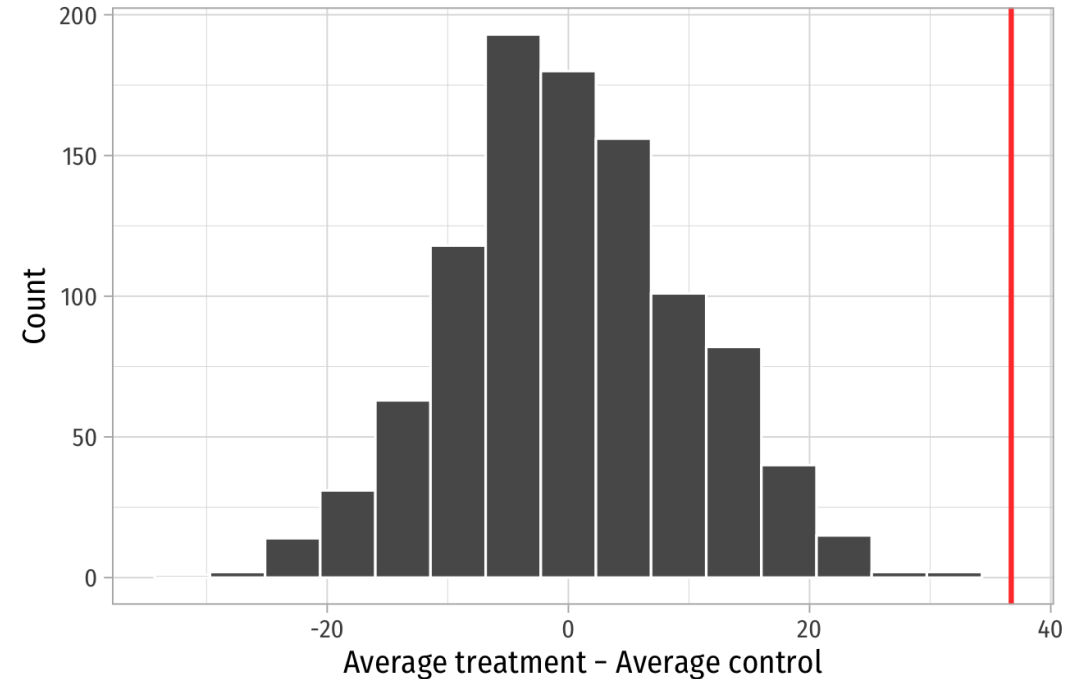
N = 10; p = 0.896



## Survey 200 participants

Simulated world with no difference

N = 200; p = <0.001



# What's the right sample size?

Use a statistical power calculator to make sure you can potentially detect an effect

statistical power calculator



# Test assumptions

Every statistical test has certain assumptions

For instance, for OLS:

Linearity

Homoscedasticity

Independence

Normality

Make sure you're doing the stats correctly

# Fishing and p-hacking

Wouldn't it be awesome to run thousands of models with different combinations of variables until you find coefficients that are statistically significant?

**Don't!**

## Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

---

**1 CHOOSE A POLITICAL PARTY**

Republicans Democrats

---

**2 DEFINE TERMS**

Which politicians do you want to include?

**3 IS THERE A RELATIONSHIP?**

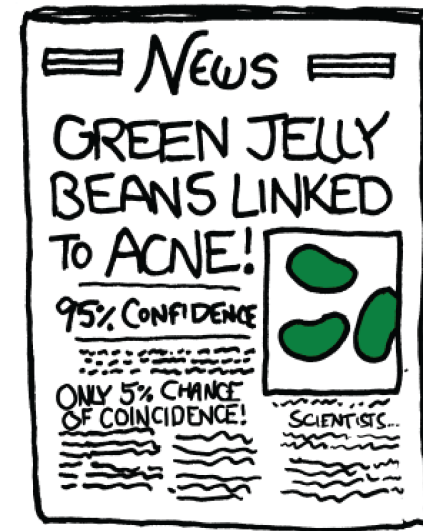
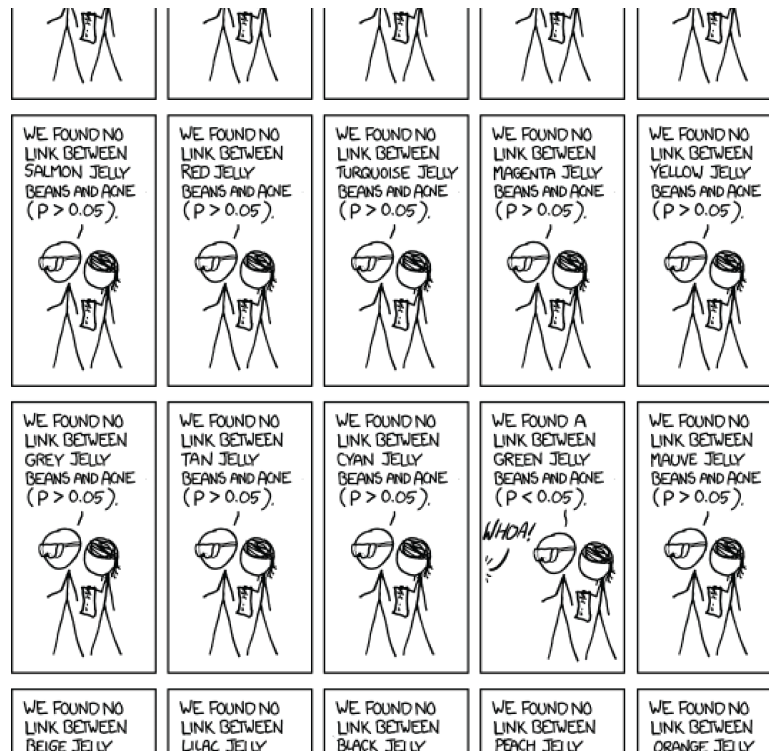
Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot represents one month of data.

**4 IS YOUR RESULT SIGNIFICANT?**

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That

# Spurious statistical significance

If  $p$  threshold is 0.05 and you measure 20 outcomes, 1 will likely show correlation by chance



# Internal validity

# Internal validity

**Omitted variable bias**

Selection

Attrition

**Trends**

Maturation

Secular trends

Seasonality

Testing

Regression

**Study calibration**

Measurement error

Time frame

**Contamination**

Hawthorne

John Henry

Spillovers

Intervening events



# Selection

**If people can choose to enroll in a program, those who enroll will be different from those who 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**



ELSEVIER

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

The Journal of  
**Socio-  
Economics**

[www.elsevier.com/locate/econbase](http://www.elsevier.com/locate/econbase)

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

Alois Stutzer<sup>\*,1</sup>, Bruno S. Frey<sup>1</sup>

*University of Zurich, Switzerland*

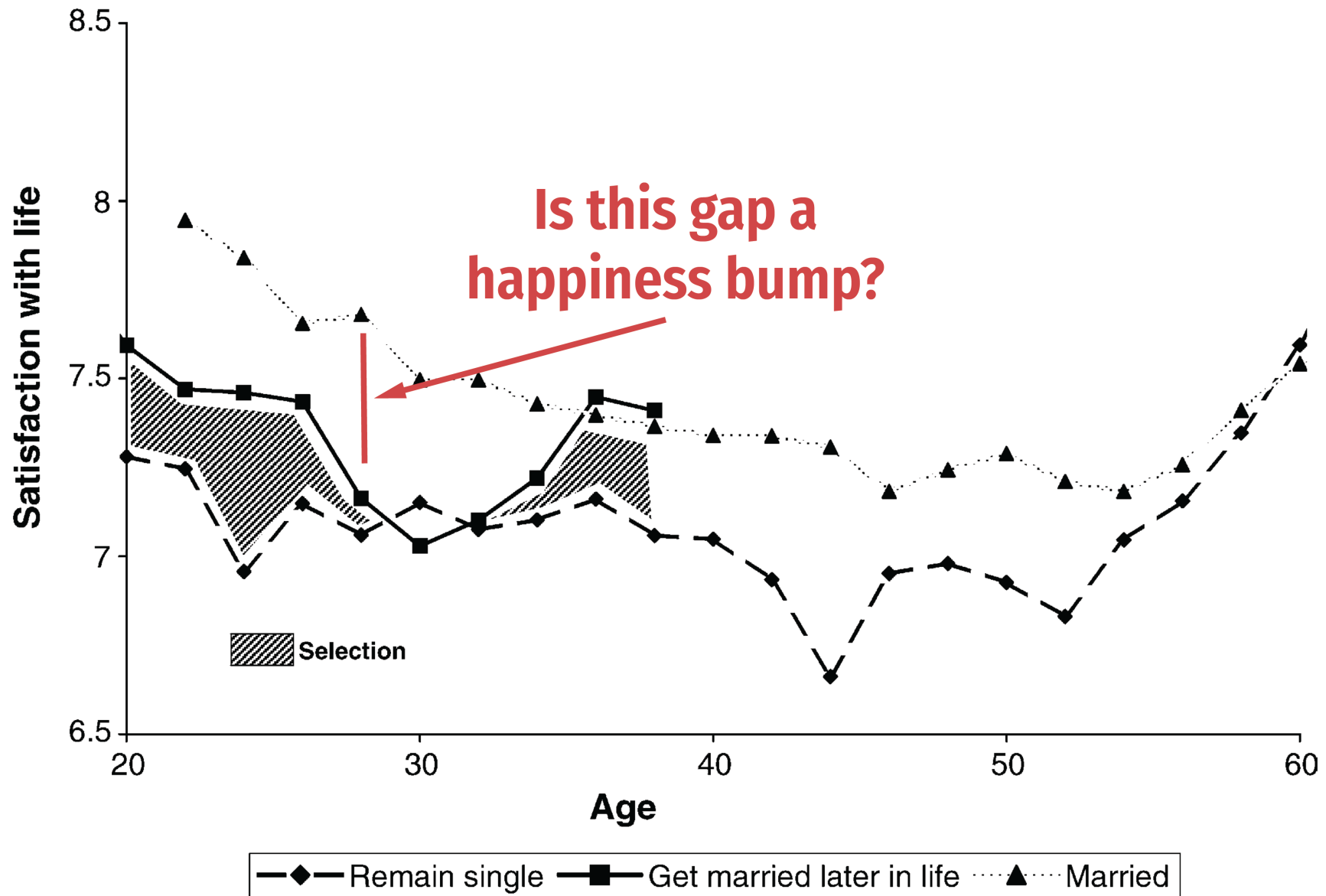
Received 4 June 2003; accepted 12 October 2004

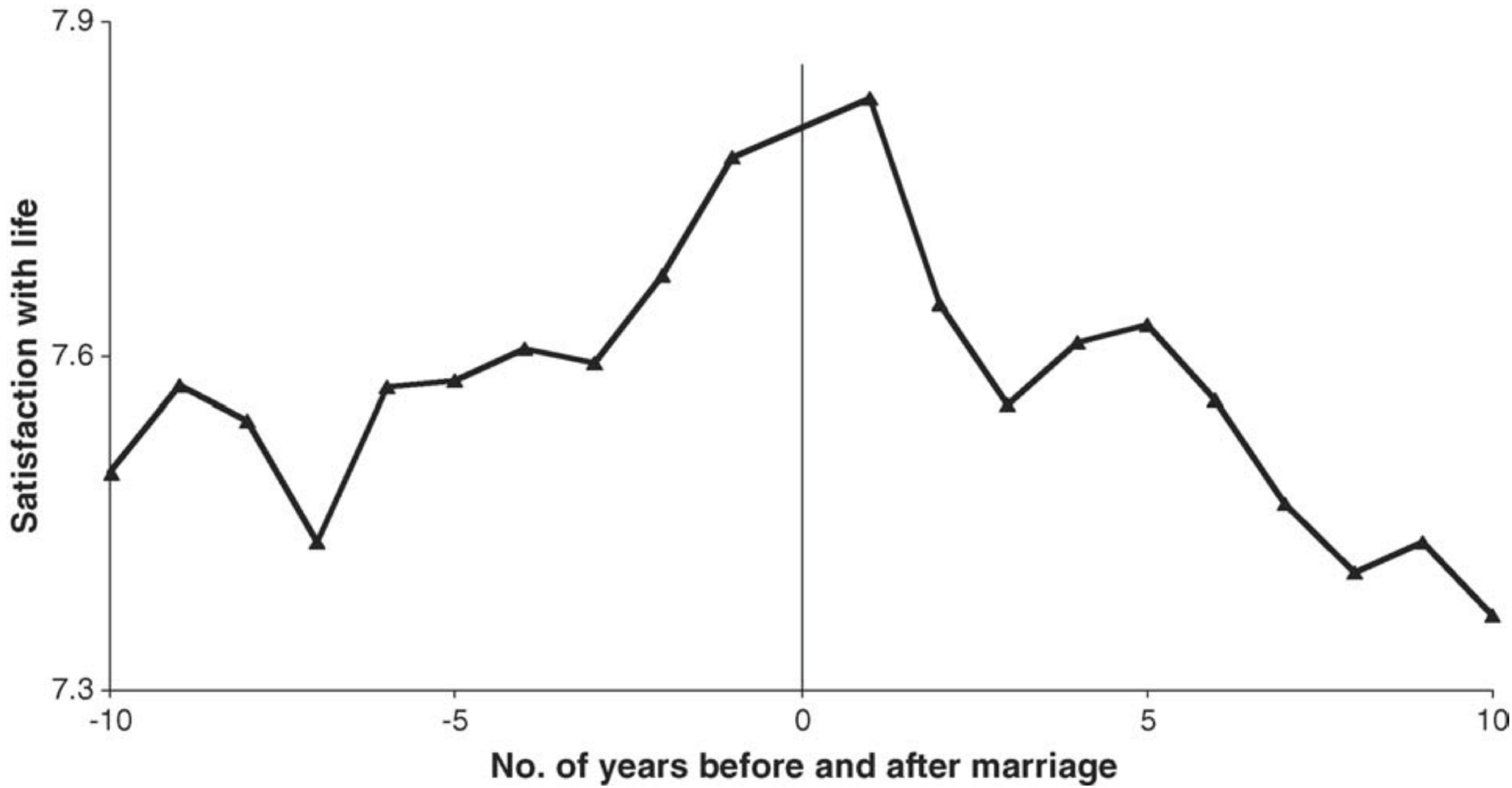
---

## 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.







# Attrition

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

**How to fix**

**Check characteristics of those who stay and those who 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**

**e.g. programs targeted at childhood development contend with the fact that children develop on their own too**

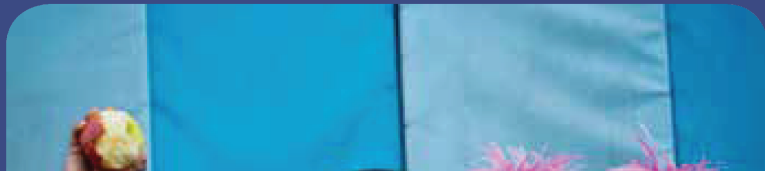
**How to fix**

**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

Patterns in data happen  
because of larger global processes

Recessions

Cultural shifts

Marriage equality

How to fix

Use a comparison group to remove the trend

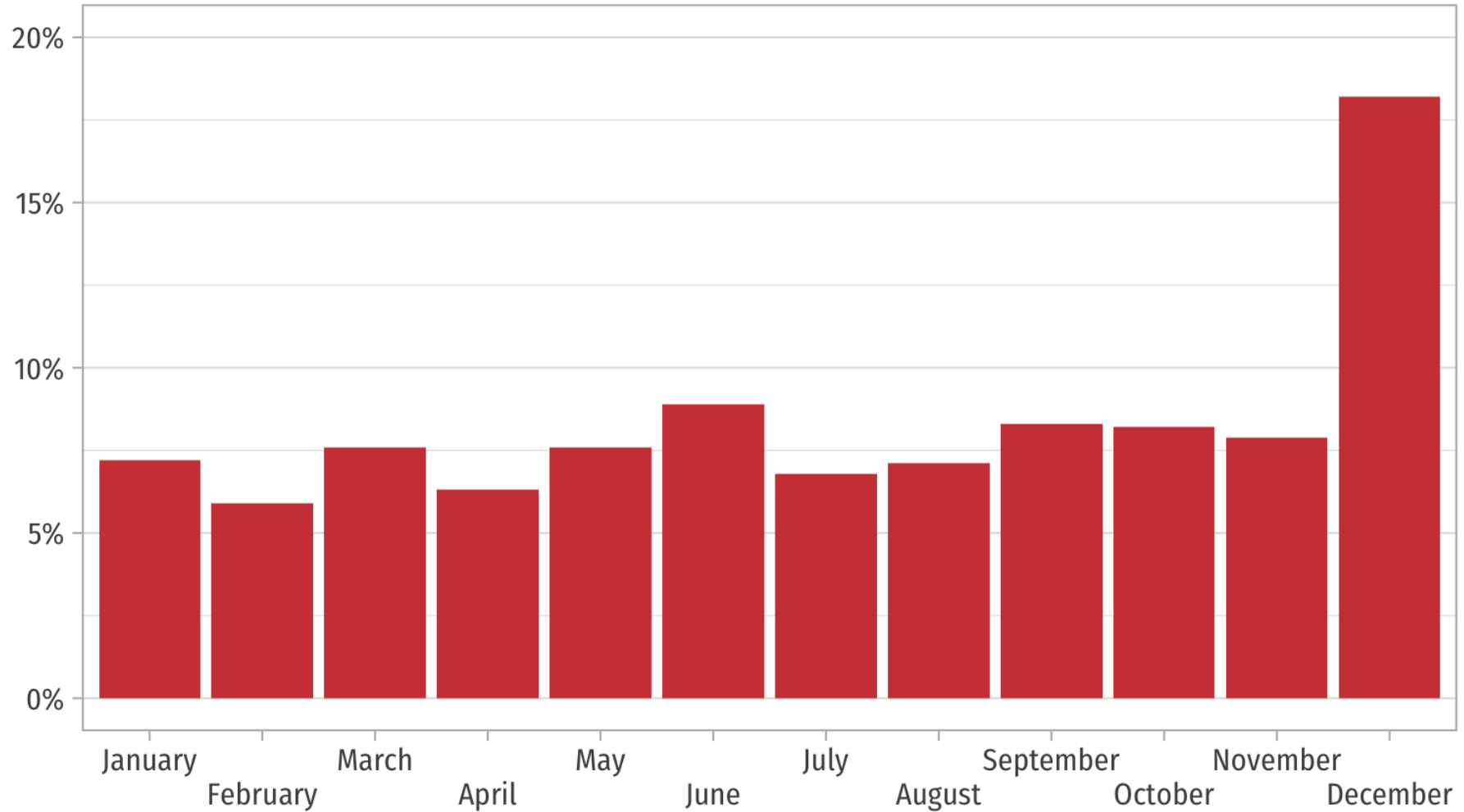
# Seasonal trends

**Patterns in data happen 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



# Testing

**Repeated exposure to questions or tasks will make people improve naturally**

**How to fix**

**Change tests, maybe don't offer pre-tests, 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 bias the 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

# Intervening events

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

**How to fix**



# Internal validity

**Omitted variable bias**

Selection

Attrition

**Trends**

Maturation

Secular trends

Seasonality

Testing

Regression

**Study calibration**

Measurement error

Time frame

**Contamination**

Hawthorne

John Henry

Spillovers

Intervening events

# Fixing internal validity

**Randomization fixes a host of issues**

**Selection**

**Maturation**

**Regression to the mean**

**Randomization doesn't fix everything!**

**Attrition**

**Contamination**

**Measurement**

# External validity



# Generalizability


Are your findings generalizable to the whole 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

# Lab conditions vs. real world

Study volunteers are weird

Western, educated, from industrialized, rich, and democratic countries

Not everyone takes surveys

Online surveys

Amazon Mechanical Turk

Random digit dialing

# Different settings and circumstances

**Does a study in one state  
apply to other states?**

**Does the effect from a mosquito net trial  
in Eritrea transfer to Bolivia?**