Analysis with Experiments

Dr. Richard W. Evans

November 5, 7, 2018

Chicago is a great place for experiments

- Economics: Richard Thaler, John List, and Steven Levitt
  - Richard Thaler, 2017 Nobel Prize in Economics
  - Best behavioral economists: Levitt (14), List (15), Thaler (24)
  - Levitt, *Freakonomics*
  - John list research methodology page
- Psychology: Leslie Kay, Marc Berman
  - Berman, FMRI/EEG work, "Fractal Brain and Cognitive Effort"
  - Kay, rats, "How the Questions We Ask Affect the Answers We Get: A Lesson from Asking Rats How They Smell"

Experiments vs. observational and surveys

- Observational and survey data: doe not intentionally, systematically change the world
- Experiments: precisely intervene in world to see have data change
- Experiments: ideal for answering cause and effect questions

Experiments

- **Weak experiment:** intervene in world and measure outcomes
  - "Perturb and observe"
  - Problem: No baseline or control to compare against
- Randomized controlled experiment
  - Randomly select treatment group to receive treatment
  - Compare against control group that does not receive treatment

Importance of control group

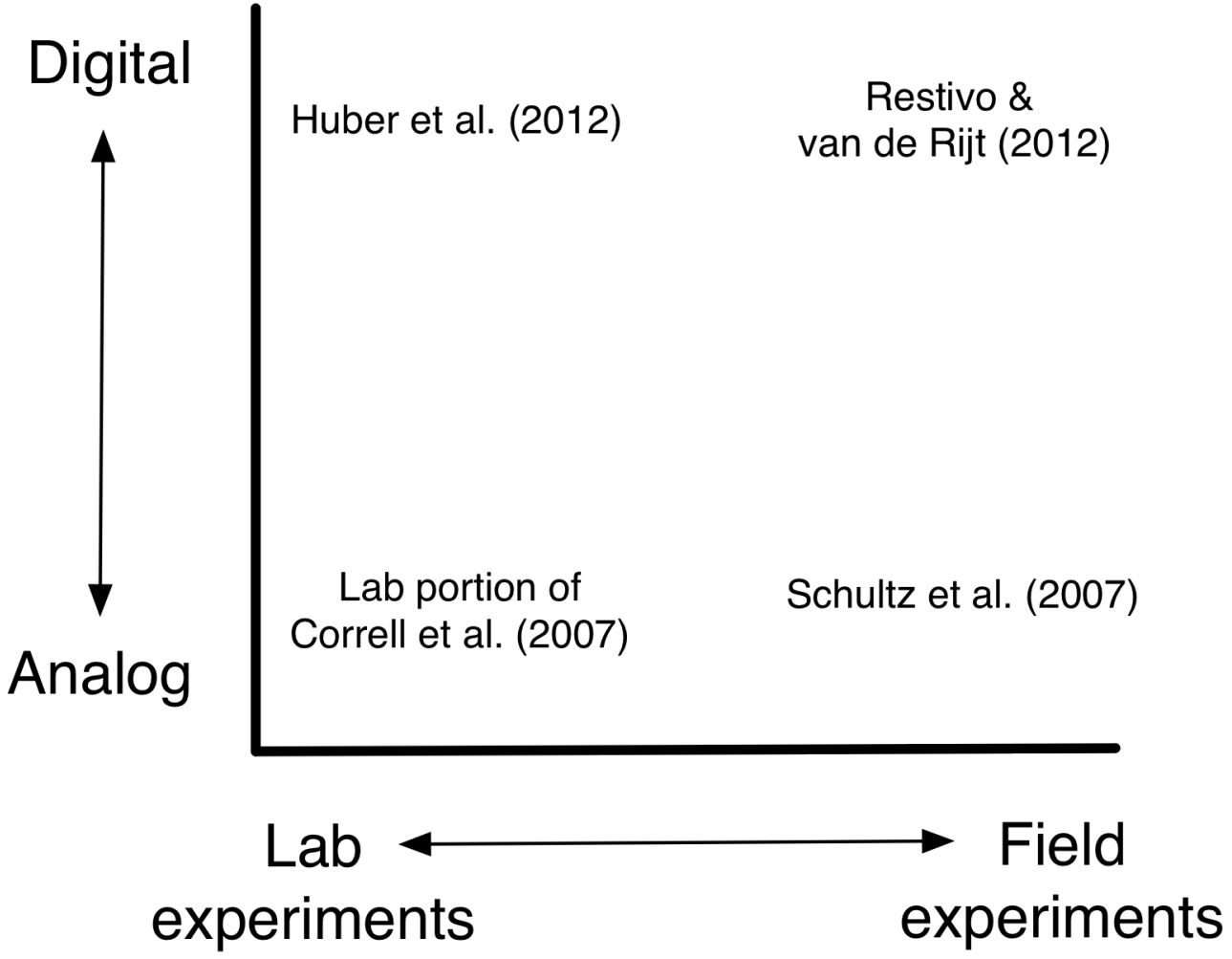
- Random selection of treatment ensures that the only thing changing, on average, among two groups is the treatment
- Restivo and van de Rijt (2012)
  - Effect of informal peer rewards
  - Randomly give stars to Wikipedia contributors
  - Surprise: Found that recipients made fewer contributions afterward
  - With control: Control group had fewer contributions. Stars had positive effect despite negative levels.
- Andrew Gelman on Caesar's Casino CEO: 5 ways to get fired from Caesar's
  - (1) theft, (2) sexual harassment, (3) running an experiment without a control group
  - (4) keeping a gambling addict away from the casino
  - (5) chapter 11 bankruptcy proceedings

Example: Patience and rationality experiment

One Now Versus **Two Later**

- How many of you would prefer **one** Snickers bar **today** versus **two** Snickers bars **tomorrow**?
- How many of you would prefer **one** Snickers bar in **30 days** versus **two** Snickers bars in **31 days**?

Two dimensions of experiments



Two dimensions of experiments

- Lab versus field
  - Lab experiments: participants enter lab setting and perform behavioral activities. Often undergrads paid small amounts
  - Lab experiments offer near total control of environment
  - Field experiments: take place in more realistic, native setting, more representative groups.
  - Field experiments have less control
- Analog versus digital
  - Digital: make use of digital infrastructure to recruit participants, randomize/deliver treatments, and measure outcomes.
  - Partially digital: Use devices in physical world to deliver treatments, measure outcomes
  - Analog experiments have fewer participants
  - Digital experiments can have millions of participants
  - Digital experiments happen over longer time scale

Experiment weaknesses

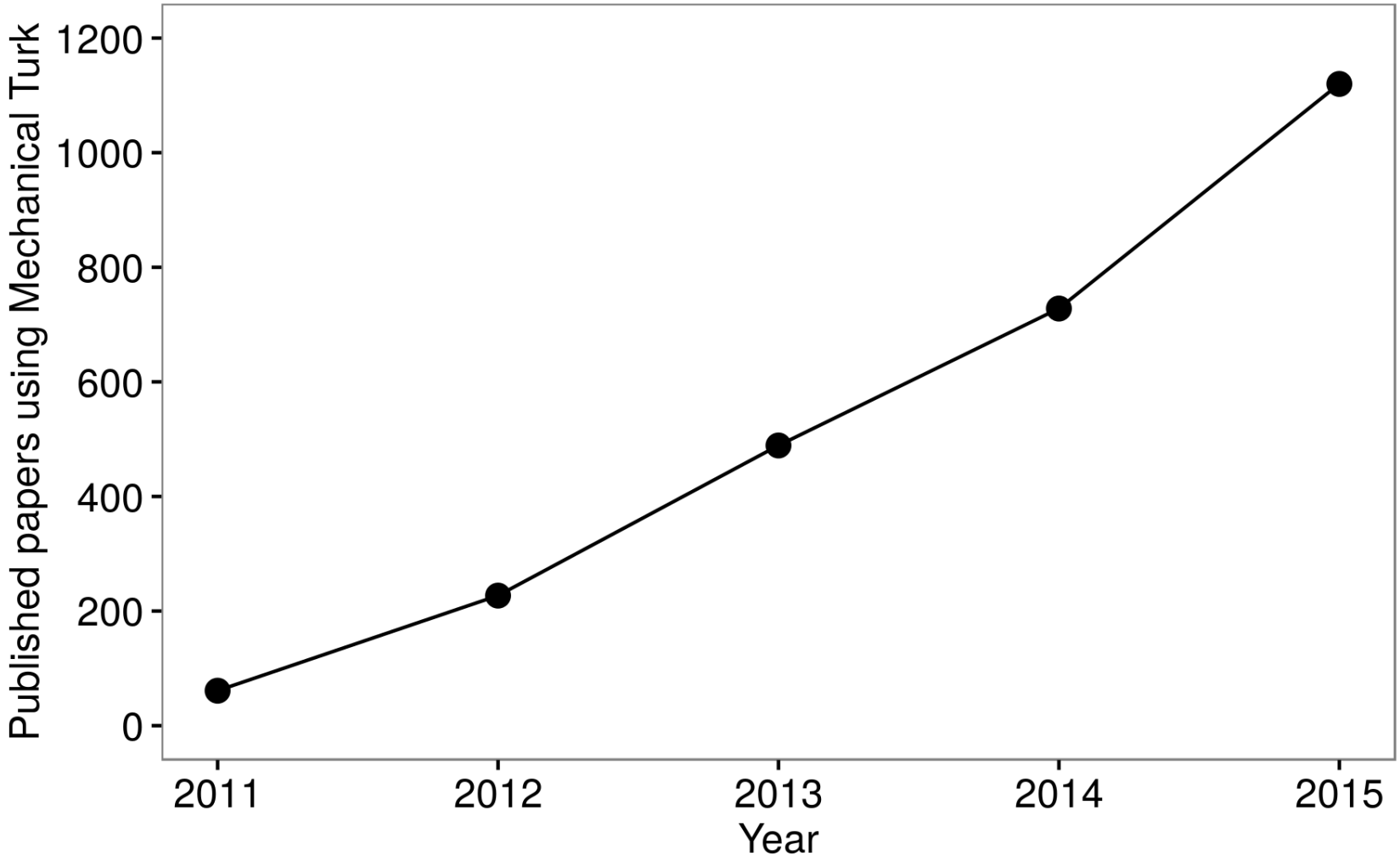
1. Cannot be used to study past
2. Environment dependence, compliance problems, equilibrium effects
3. Increased ethical concerns

Digital experiments: Amazon Mechanical Turk

- What is Amazon Mechanical Turk and why is it called that?
  - Wikipedia The Turk [Mechanical Turk]
  - Assignment 5
  - Amazon Mechanical Turk

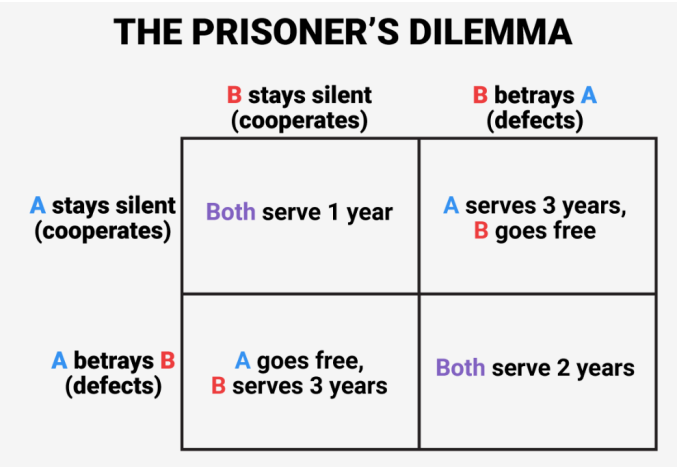
Mechanical Turk

Mechanical Turk



Classic Lab Experiment: Prisoner's dilemma

List, John A., "Friend or Foe? A Natural Experiment of the Prisoner's Dilemma," *Review of Economics and Statistics*, 88:3 (August 2006), pp. 463-471.



- Joker did a Prisoner's dilemma variant with the two boats at the end of *The Dark Knight*
- List (2006) controlled for partner similarities in Prisoner's Dilemma game show
- List found that age discrimination in cooperation

Richer experimental designs

Simple experiments

- Most experiments are simple experiments
- Narrowly focused
- Does this treatment work?

Richer experimental designs

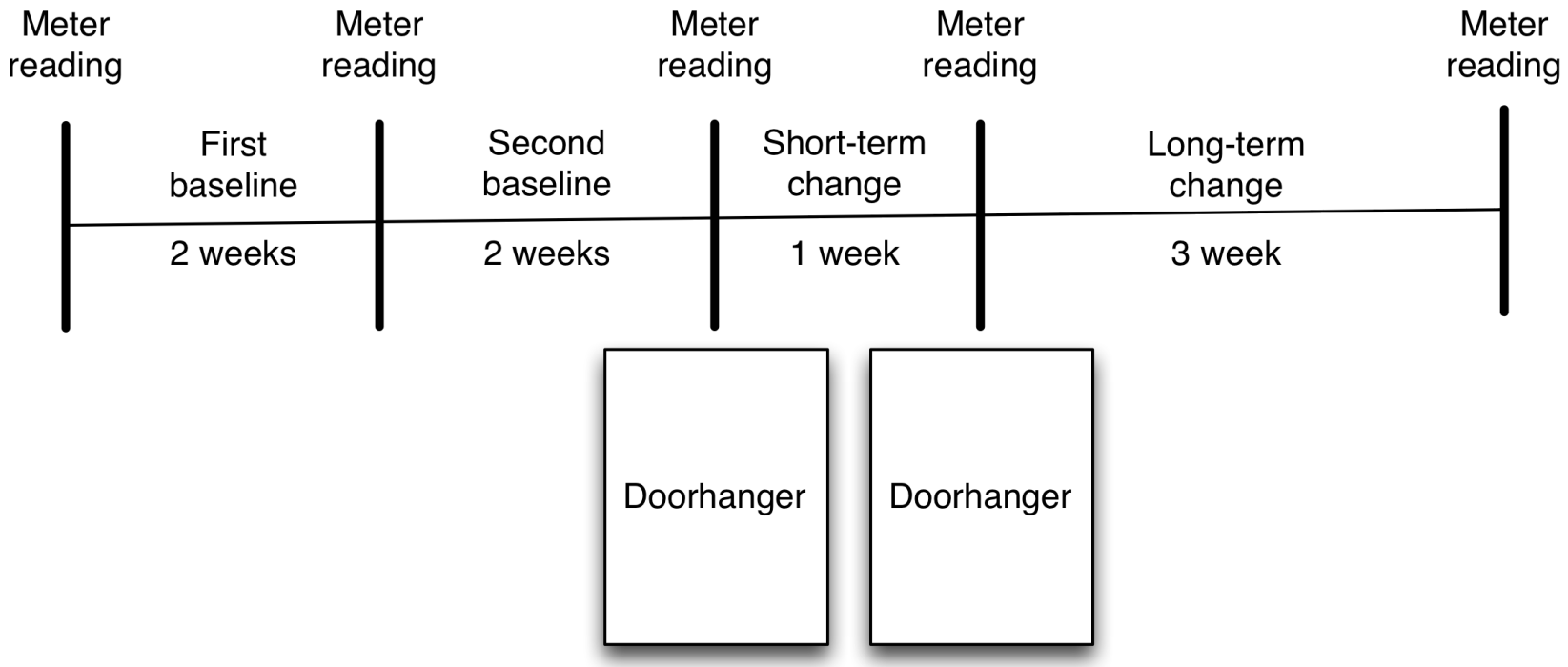
- Validity (outside applicability, generalizability)
- Heterogeneity of treatment effects (many bins)
- Mechanisms (requires theory)

Schultz, et al (2007): Electricity bills

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius, "The Constructive, Destructive, and Reconstructive Power of Social Norms," *Psychological Science*, 18:5 (May 2007), pp. 429-434.

- What is effect of normative messages?
- Results mixed in laboratory experiments.
- Is there a *boomerang effect*?
    - Mean reversion. Those above the norm go down, those below the norm go up.

Schultz, et al (2007): Design



- Energy saving tips (use fans instead of AC)
- **(Descriptive norm)** Own energy usage compared with average energy usage
- > 2nd Exp: **(Injunctive norm)** Emoticon for usage ☺ ☹

Schultz, et al (2007): Results

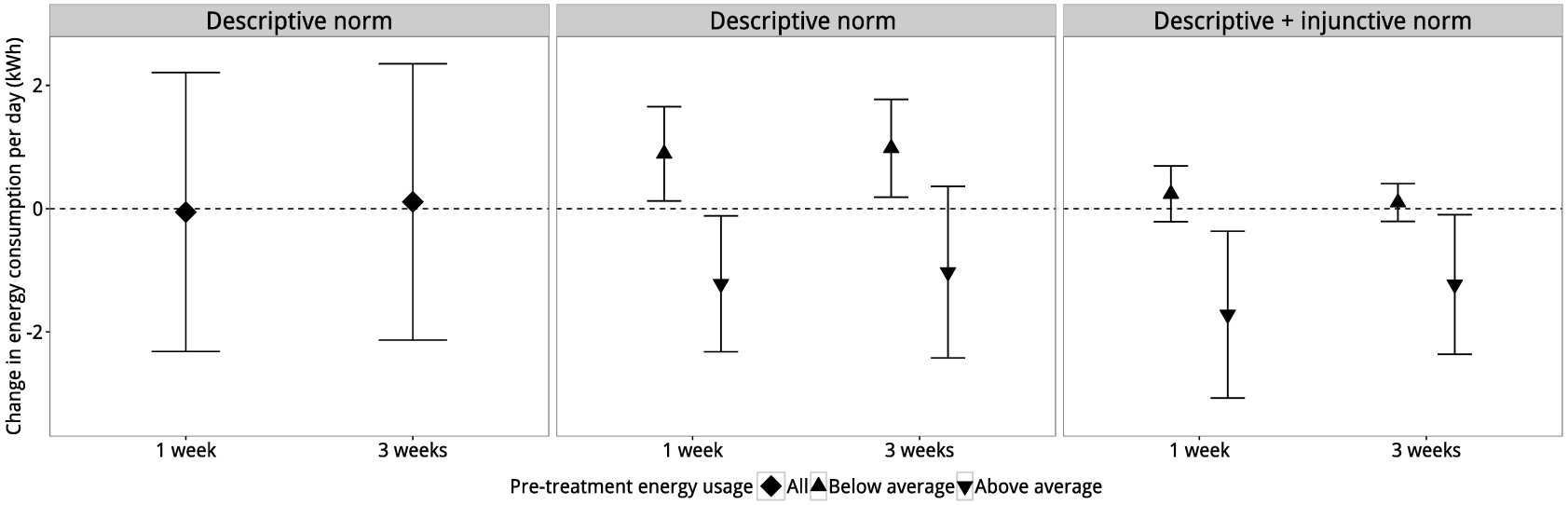
Simple result

- Treatment had no average effect

Richer results

- People above the mean usage decreased consumption
- People below the mean increased consumption (boomerang)
- With emoticons: People above mean reduced usage more
- With emoticons: people below the mean increased usage less

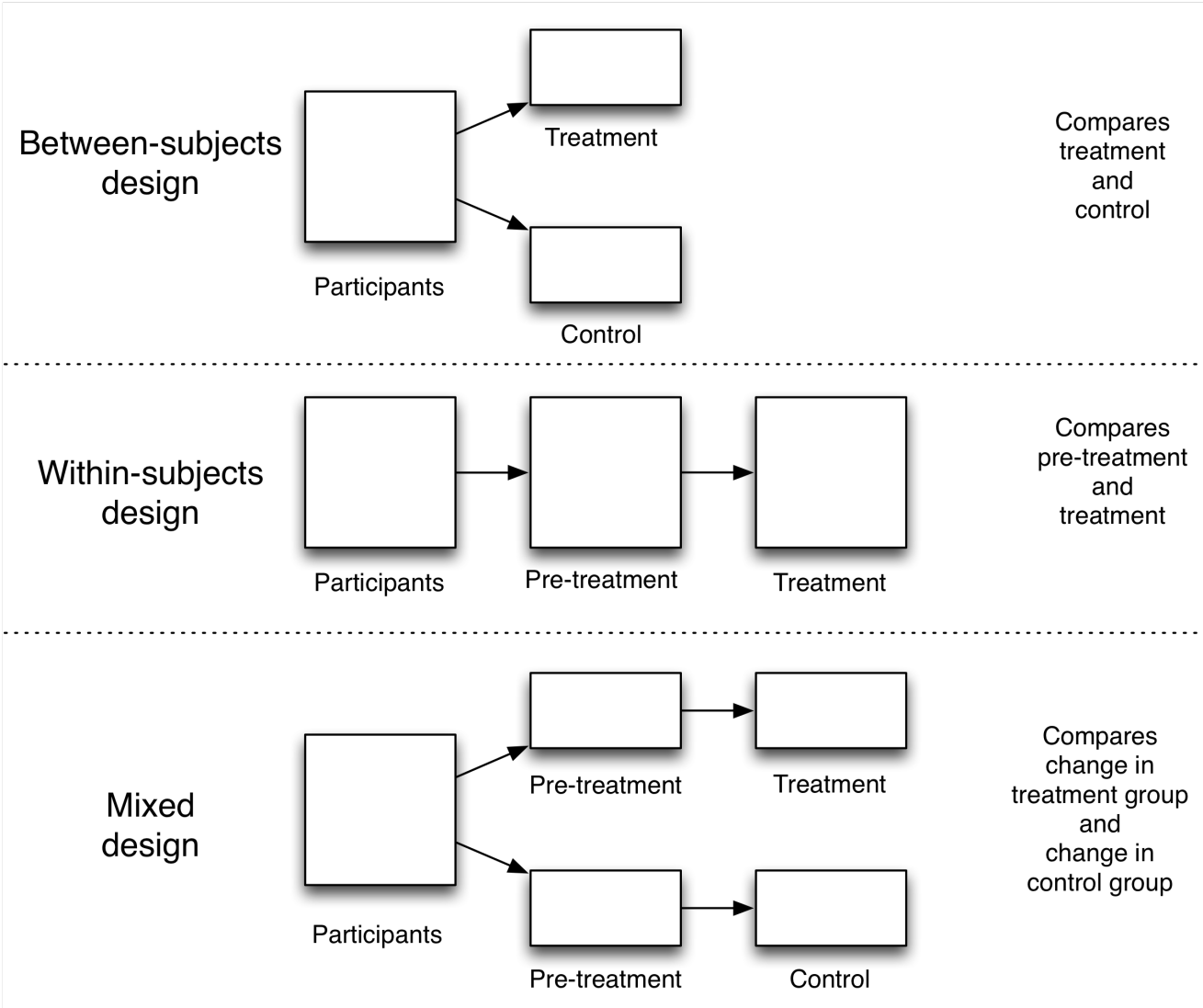
Schultz, et al (2007): Results



Schultz, et al (2007): Questions

1. What is the control group in Schultz, et al (2007)?
   - Between-subject vs. within subject designs
   - Would this study be better if they had a control group? What would that look like?
2. How did Schultz, et al (2007) deal with heterogeneous treatment effects?
   - Are there any other heterogeneous treatment effects they should have covered?
   - (Answer is always yes): Alcott and Rogers (2014), Costa and Kahn (2013)
     - Alcott and Rogers (2014): short-run vs. long run
     - Costa and Kahn (2013): Environmentalist ideology

Richer experimental designs



Validity

1. Statistical validity
   - Are statistics done right?
   - *Damned Lies and Statistics*
   - Equally likely in digital vs. analogue
2. Internal validity
   - Were experimental procedures done correctly?
   - Wiki: Ambiguous temporal precedence, confounding, selection bias, maturation, repeated testing, instrument change, attrition, and more
   - Easier to ensure in digital, treatment more often gets to subject
3. Construct validity
   - Match between data and theoretical constructs
   - Are doorhangers with tags really descriptive and injunctive norms?
   - Bigger concern in digital experiments
4. External validity

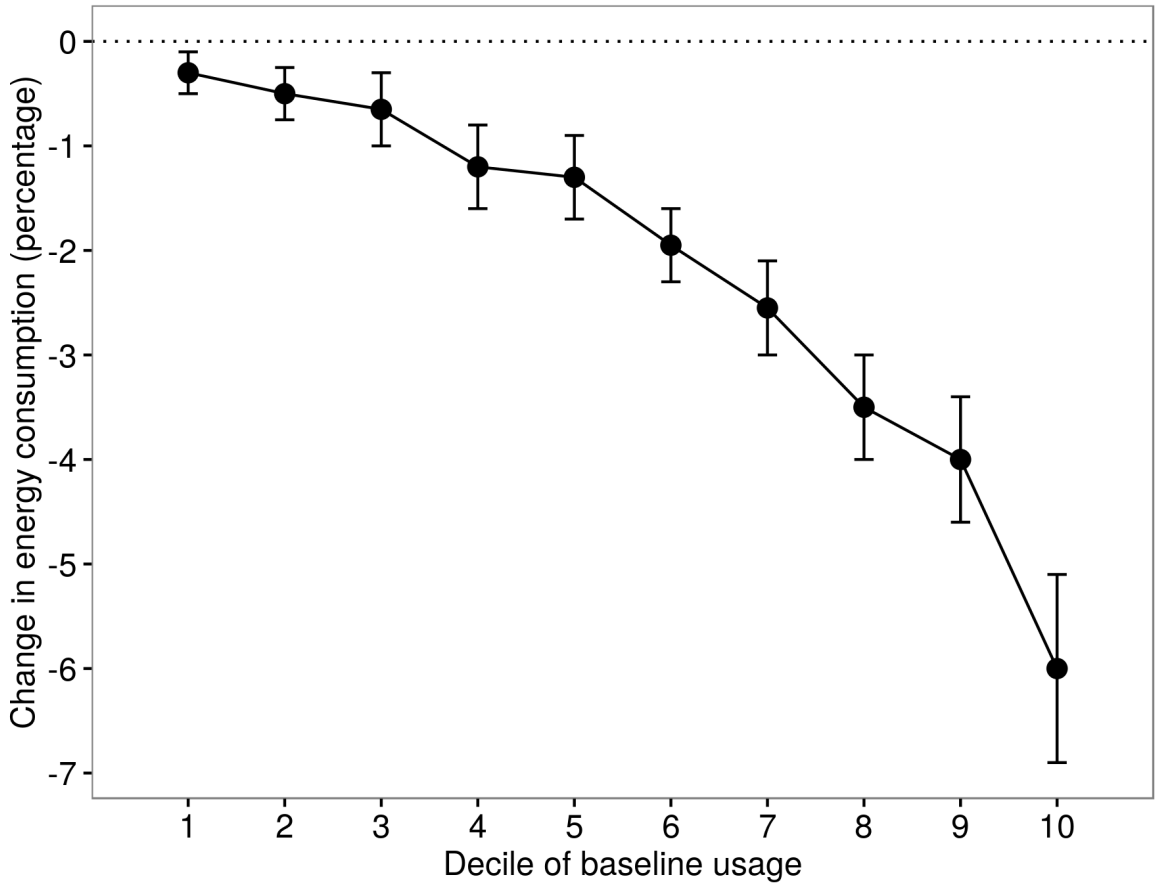
Validity

1. Statistical validity
2. Internal validity
   - Were experimental procedures done correctly?
   - Wiki: Ambiguous temporal precedence, confounding, selection bias, maturation, repeated testing, instrument change, attrition, and more
   - Easier to ensure in digital, treatment more often gets to subject
3. Construct validity
   - Match between data and theoretical constructs
   - Are doorhangers with tags really descriptive and injunctive norms?
   - Bigger concern in digital experiments
4. External validity
   - Can results be generalized?
   - Often requires replication in other settings
   - Easier with digital because of larger scale

Heterogeneity of Treatment Effect

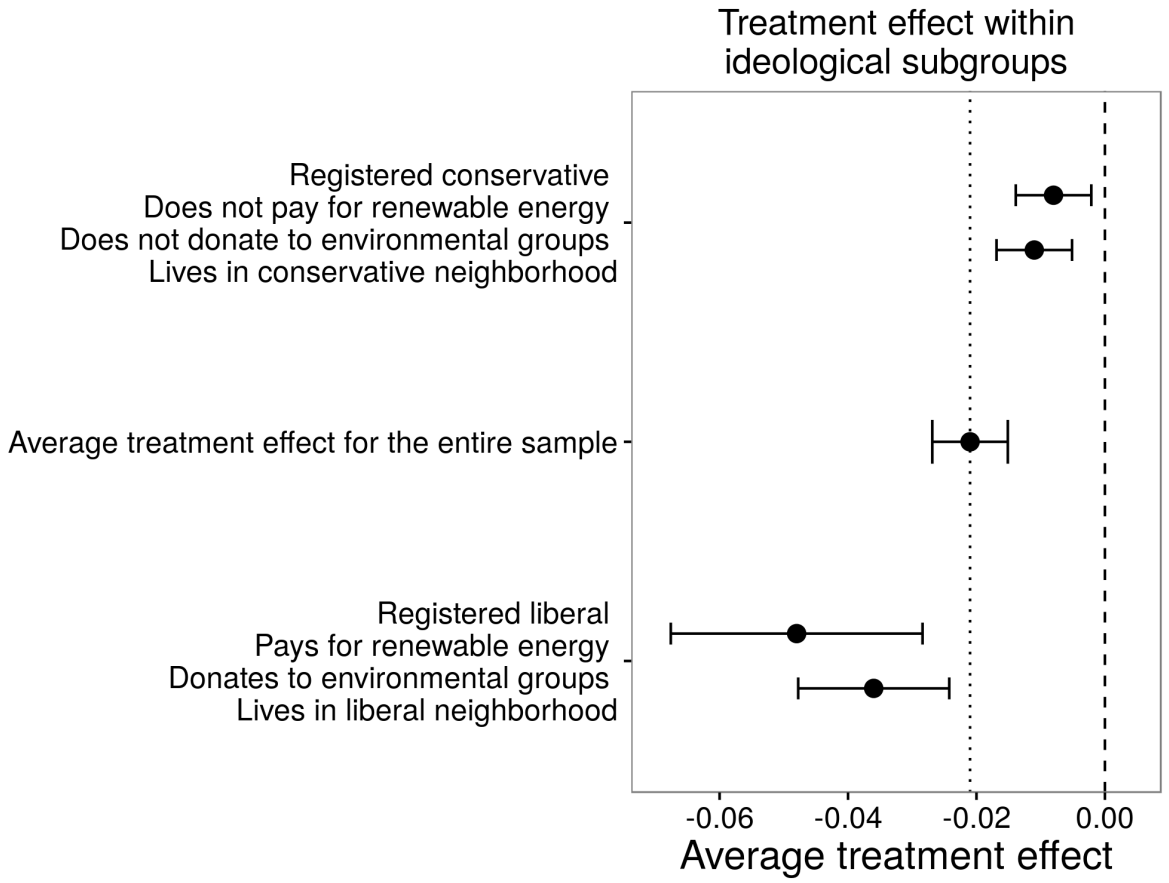
- Confounders, boomerang effects
- Divide up your groups until you isolate uniform effects
- Can this be used perniciously?
  - Gerrymandering, data mining
  - How could you avoid the appearance of this?
    - File a pre-experimental design
    - Show that any more divisions create no change
    - Sensitivity analysis
- How many papers are written by just this extension?

Alcott (2011), behavior change by usage level



Change in energy use was different for deciles of energy usage

Costa and Kahn (2013), behavior change by ideology



Change in energy use was different for user ideology

Mechanisms

- Experiments measure what happened
- Mechanisms explain why and how it happened
- Mechanism is another word for model

Digital experiments help uncover mechanisms

- Enable collecting and processing more data
- Enable testing many treatments

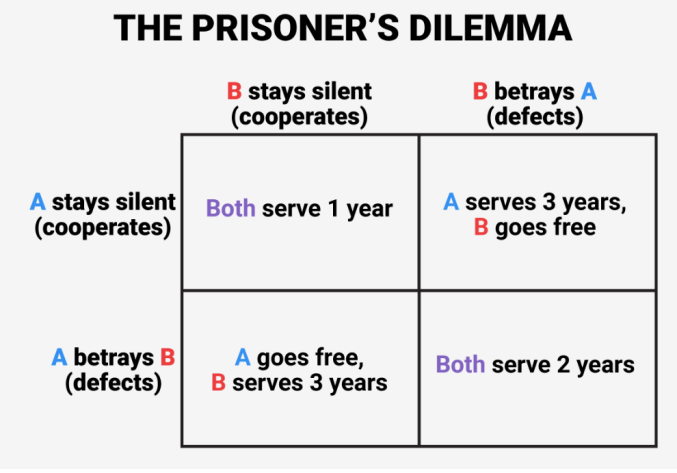
But the best experiments/research...

The best research combines theory and empirical work: mechanisms and experiments.

Mechanisms: Atheoretical approach

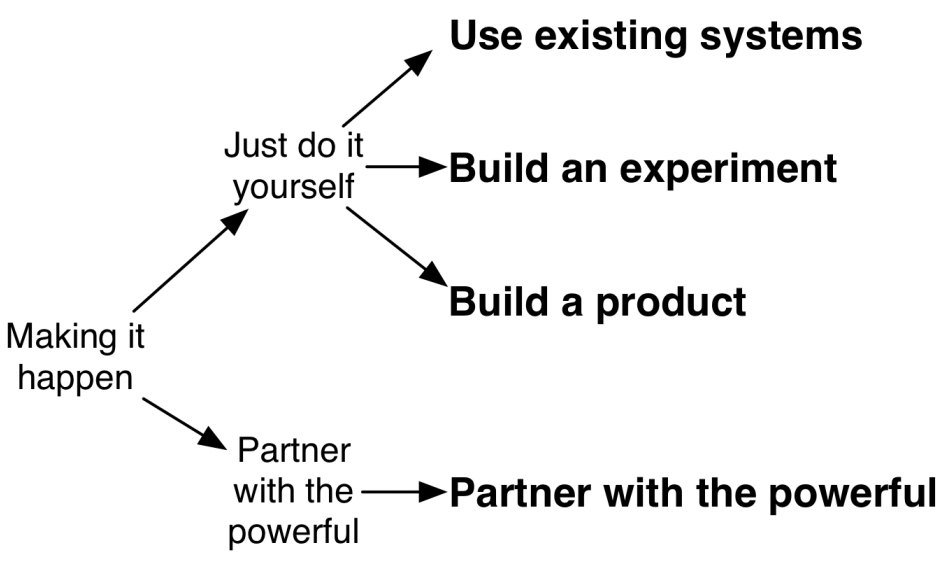
- Test everything
  - Hard to do
  - Take myriad data
- Full factorial design (2^k factorial design)
  - If k potential treatments, 2^k groups to test
- Write down 3 electrical options: tips, appeal, peer info.
  - Should be 8, including control

Mechanisms: Theoretical approach



- Theory suggests how people should behave
- Test the theory on people: field, lab, data
- Adjust the theory based on evidence
- Examples: Epstein-Zin preferences, habit persistence, hyperbolic discounting, adaptive expectations, investment and price frictions.

Tradeoffs of experiment platforms



	Cost	Control	Realism	Ethics
Use existing systems	Low	Low	High	Potentially complex
Build an experiment	Medium	High	Medium	Relatively easy
Build a product	High	High	High	Relatively easy
Partner with the powerful	Low	Medium	High	Potentially complex