A Course in Experimental Economics

John A. List

IF YOU WANT TO CITE OR GIVE ATTRIBUTION TO THIS SLIDE DECK, PLEASE CITE THE FORTHCOMING BOOK:

List, John A. "A Course in Experimental Economics," University of Chicago Press, forthcoming.

Some Tips for “Start-Ups” to do Better Field Experiments and Getting Your Work Published

J.A. List
U. Chicago, ANU, & NBER

~~Some Tips for Doing Better Field Experiments and~~

~~Getting Your Work Published~~

John List Unplugged

Some Tips for "Start-Ups" to do Better Field Experiments and Getting Your Work Published

J.A. List
U. Chicago, ANU, & NBER

S.R. Goal: Induce you to go into your AFE slides and your working papers and make changes to improve the work

L.R. Goal: Induce you to change the way you design, examine, interpret, and/or write-up your future studies

Human Capital Goals:
Teach you something that you didn't know that you wanted to know
Help you become a better field experimentalist

Today's Outline

- Internal Validity Sundries
  - A. Exclusion Restrictions in P.O. Framework
  - B. Power
    - Optimal Design
    - Inference
  - C. Within-Subject Thoughts
  - D. MHT

Today's Outline

- Inference

E. Introducing Option C: Policy-Based Evidence

F. It's More Than a Label to Me

G. Generalizability

  - Use Mediation and Moderation to help
  - SANS conditions

A. Potential Outcomes Framework

If you assume:

- SUTVA
- Observability
- Complete Compliance
- Statistical Independence

Treatment: $D_i = 1$	Control: $D_i = 0$	Treatment effect: τ_i
Observed: $Y_1(1)$	**Missing:** $Y_1(0)$	$\tau_1 = Y_1(1) - Y_1(0) = ?$
Missing: $Y_2(1)$	**Observed:** $Y_2(0)$	$\tau_2 = Y_2(1) - Y_2(0) = ?$

Then you can recover ATE:

$$\tau = E[Y_i(1)] - E[Y_i(0)]$$

B. Power

When I read papers I find statements such as:

*I can reject at the 5% level

*My effects are economically significant

And, when people present and I raise power issues early on, they often smile and say:

*I will be rejecting nulls so all is good!

One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects

Zacharias Maniadis, Fabio Tufano, John List

2014 AER

A simple Bayesian framework

$$PSP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

PSP: the probability that a declaration of a research finding, made upon reaching statistical significance, is true.

- α: Level of statistical significance
- 1 - β: Level of power

π: can think of this as the prior

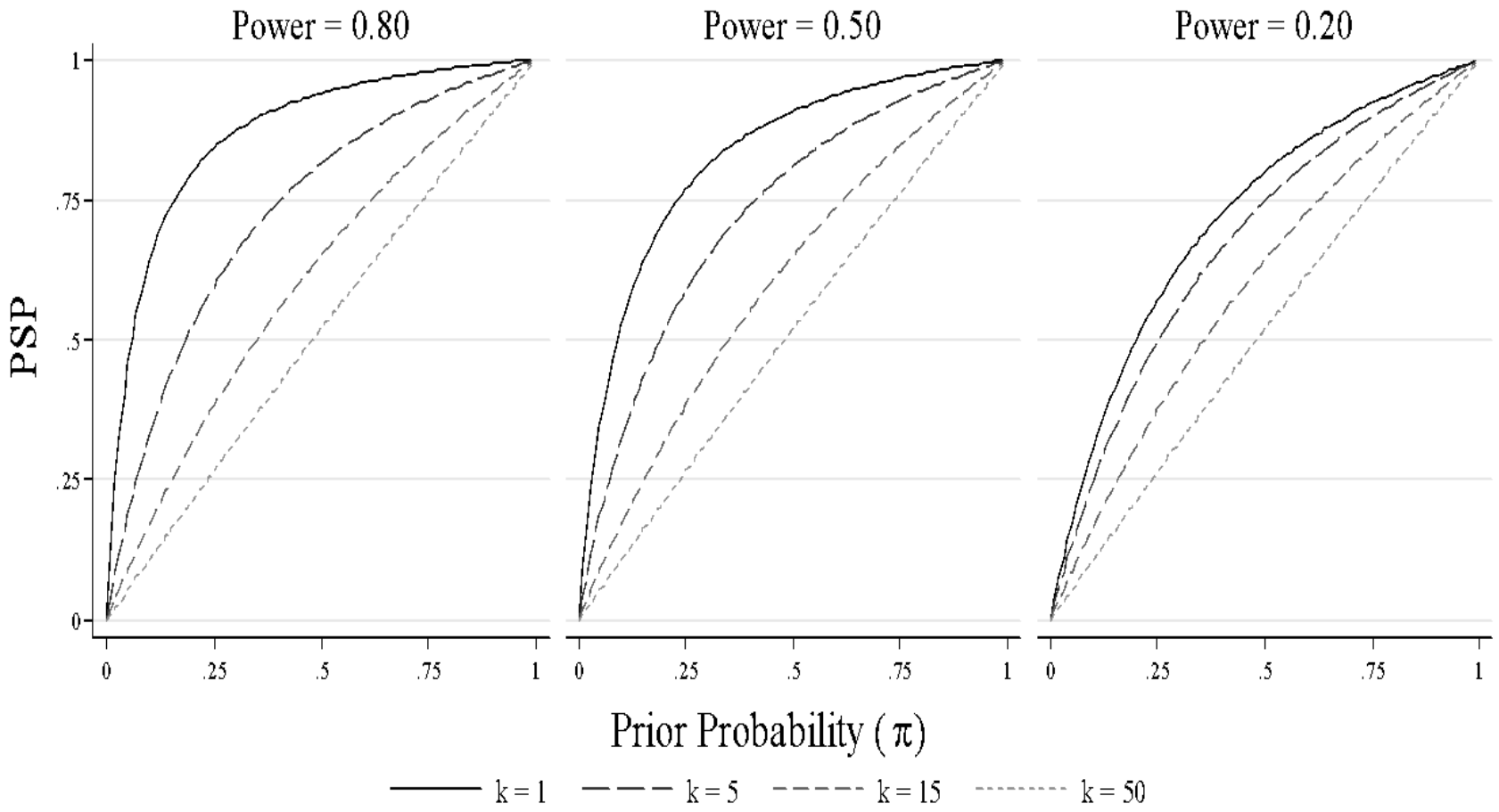
Considering importance of power

- Take the derivative of the PSP with respect to power, 1 - β, and you see that:

$$\frac{\P PSP}{\P(1- b)} > 0$$

- That is, holding α constant, the more power, the higher the PSP for a given result

PSP & Power


Power = 0.80
Power = 0.50
Power = 0.20
PSP
1
.75
.5
.25
0
0
.25
.5
.75
1
Prior Probability (π)
k = 1
k = 5
k = 15
k = 50

Upshot for Individual Scholars?

- Not only p-values, but power and priors matter too.

- We can take a step in the right direction if we present our results by showing how PSP changes due to our findings over a broad range of priors:

"If the prior was .3 then the PSP is .85 because of our research"

Provocative Point #1

Research should largely be evaluated based on how much it moves our priors

Original Study

When prior probability π = .01 and power (1 - β) = .8, the **PSP = ??? (with one study).**

Post-Study Probability = $\frac{(1-\beta)\pi}{(1-\beta)\pi+\alpha(1-\pi)}$

π = fraction of associations true

1 - β = power

α = significance level

Maniadis, Tufano, List (2014, AER)

Original Study

When prior probability π = .01 and power (1 - β) = .8, the **PSP = 0.02 (with one study).**

Post-Study Probability = $\frac{(1-\beta)\pi}{(1-\beta)\pi+\alpha(1-\pi)}$

π = fraction of associations true

1 - β = power

α = significance level

Maniadis, Tufano, List (2014, AER)

1st Replication

When prior probability π = .01 and power (1 - β) = .8, the

PSP = 0.10 (with one replication).

Post-Study Probability = $\frac{(1-\beta)\pi}{(1-\beta)\pi+\alpha(1-\pi)}$

π = fraction of associations true

1 - β = power

α = significance level

2nd Replication

When prior probability π = .01 and power (1 - β) = .8, the

PSP = 0.47 (with two replications).

Post-Study Probability = $\frac{(1-\beta)\pi}{(1-\beta)\pi+\alpha(1-\pi)}$

π = fraction of associations true

1 - β = power

α = significance level

3rd Replication

When prior probability π = .01 and power (1 - β) = .8, the

PSP = 0.91 (with three replications).

Post-Study Probability = $\frac{(1-\beta)\pi}{(1-\beta)\pi+\alpha(1-\pi)}$

π = fraction of associations true

1 - β = power

α = significance level

C. Power: Optimal Design

- Everyone knows that you should have a sample size that allows you to make proper inference.
- Fewer people follow simple rules of power tests to choose sample sizes efficiently before they conduct the experiment.

Power

A. Our usual approach stems from the standard regression model: under a true null what is the probability of observing the coefficient that we observed?

B. Power calculations are quite different, exploring if the alternative hypothesis is true, then what is the probability that the estimated coefficient lies outside the 95% CI defined under the null.

A Quick Design Example

(List, Sadoff, Wagner, 2011, *Simple Rules of Thumb, Experimental Economics*)

A. 0/1 treatment, equal outcome variances

B. Treatment Intensity—continuous treatment

C. Clusters

Example: Treatment Intensity

- Consider our recent field experiment exploring the education production fct:
  - We could use various incentive levels—from $1 to $9, but school wanted at most 5 values.
  - Assume that the outcome variance is equal across various cells.
- How should we allocate the sample over these 5 values if we have 1000 subjects?

Reconsider what we are doing:

$$Y = XB + e$$

- One goal in this case is to derive the most precise estimate of *B* by using exogenous variation in X.

- Recall that the standard error of *B* is

$$\frac{var(e)}{n * var(x)}$$

Rules of Thumb

- Linear: ½ sample at X=1 and ½ at X=9.
- Quadratic: ¼, ½, ¼ at X=1, X=5, and X=9
- Intuitively: the test for a quadratic effect compares the mean of the outcomes at the extremes to the mean of the outcome at the midpoint

Provocative Point #2

Theory/experiment links should go beyond merely nature of treatments to include which cells are actually populated

If your model is linear why are you populating more than 2 cells?

C. Within-Subject Designs

- Each unit i receives multiple treatments at different times t.
- So observe, e.g. $Y_{i,t=1}(1)$ and $Y_{i,t=0}(0)$
- And can estimate $\tau_i = Y_{i,t=1}(1) - Y_{i,t=0}(0)$
- But now need extra assumptions for Internal Validity:
  - No Anticipation
  - Temporal Stability
  - Causal Transience

Within Design

- Within design provides the full joint dbn, even with heterogeneity: we recover outcomes in both states directly
- We can now estimate the entire treatment effect dbn! Median effects, fraction helped, inequality, etc.
- Bonus: more power for a given n

Stealth Within-Subject Designs

List (2004; QJE; Nature and Extent of Discrimination in the Marketplace)

- Natural field experiment of List (2004) exploring discrimination in the sportscard market. Confederates shop at unsuspecting dealers.
- I used a 'stealth design' by having the same dealer bargain people of different races and genders. Can observe both Y_{i1} and Y_{i0} for each unit
- This can go beyond measuring average treatment effects to explore the **entire** joint distribution of outcomes.
- Findings suggest there is a strong tendency for minorities to receive initial and final offers that are inferior to those received by majorities, and that the observed discrimination is not due to animus, but represents statistical discrimination

Provocative Point #3

Plausible internal validity for nearly every within-subject design demands a stealth approach

→ *Majority of NFEs measuring discrimination have this capability but do not design in this manner or analyze the data accordingly*

D. MHT

Multiple hypothesis testing refers to any instance in which a family of hypotheses are carried out at once and one has to decide which hypotheses to reject

A common, yet potentially critical, flaw that leads to an excess of false positives

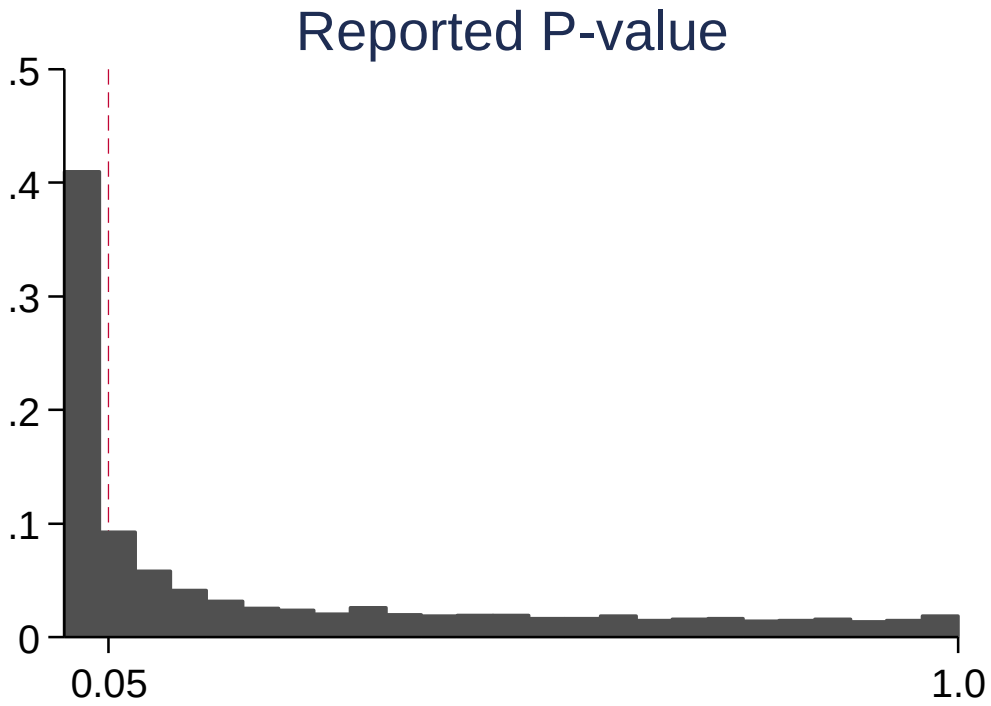
Just because you have your plan pre-registered does not mean you can ignore

Three Major Cases of MHT

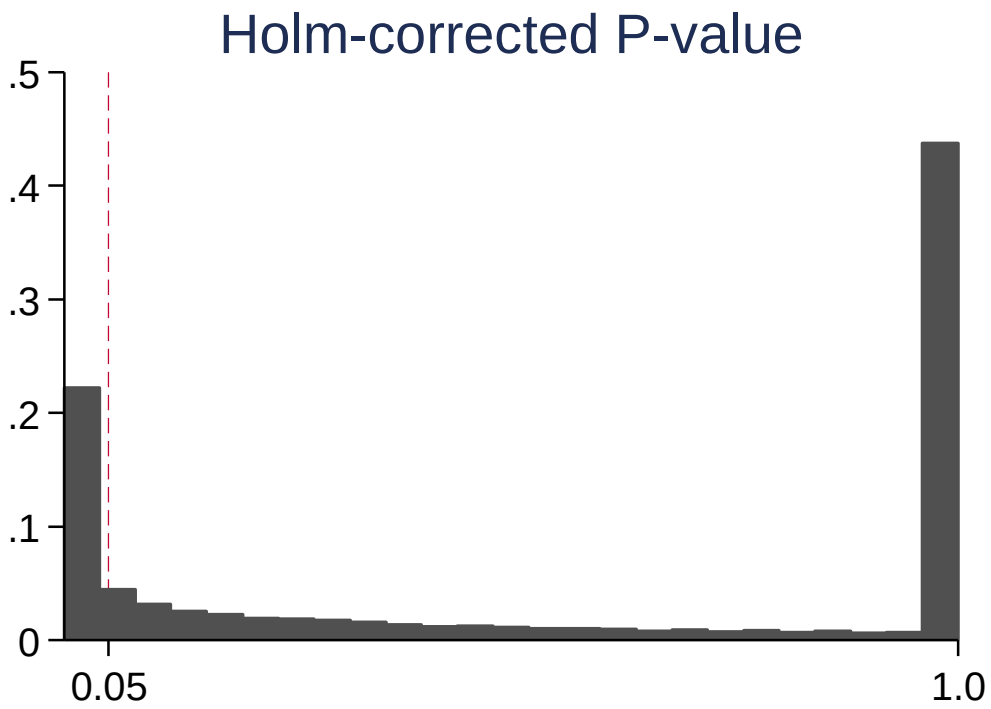
A. More than one outcome
B. Heterogeneity (sub-group analysis)
C. More than one treatment

I will wager a day's wage that 99%+ of studies in the '22 AFE have one or more of these issues

How Big of a Problem is it?



Corrected P-values



One Approach to MHT Corrections

Multiple Hypothesis Testing in Experimental Economics; List, Shaikh, and Xu (2019, EE)

A. More than one outcome
B. Heterogeneity (sub-group analysis)
C. More than one treatment

https://github.com/seidelj/mht

MHT In Regression Framework

Multiple Testing with Covariate Adjustment in Experimental Economics; List, Shaikh, and Vayalinkal (2021, NBER)

A. More than one outcome
B. Heterogeneity (sub-group analysis)
C. More than one treatment

https://github.com/vayalinkal/mhtexp2

Provocative Prognostication #1

The chief reason for false positives in experimental economics is not p-hacking, file drawer issues, or lack of replication, it is not adjusting for multiple testing

Inference

E. Introducing Option C

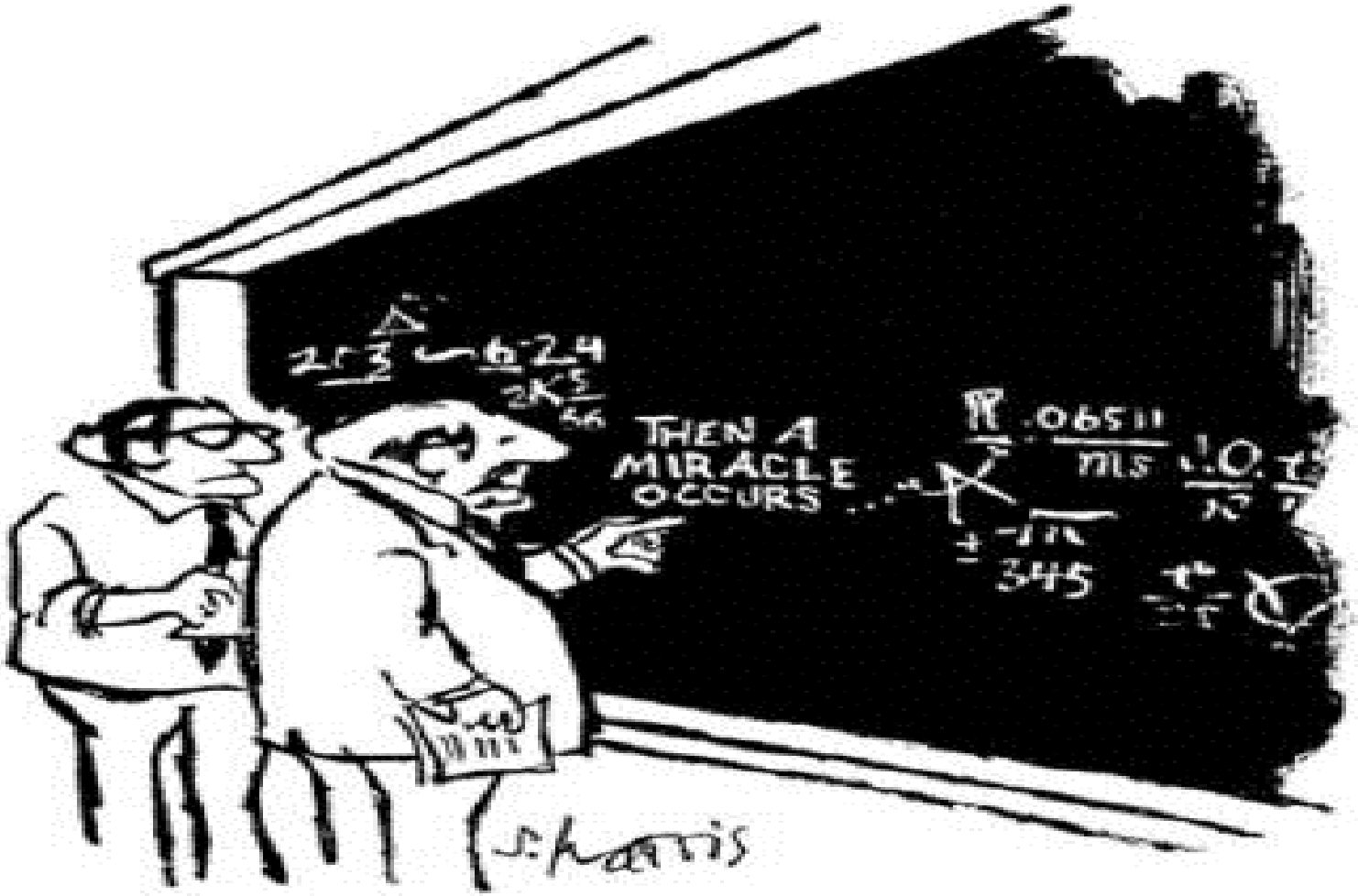
Where we are as a field

- As empiricists we have focused almost exclusively on ***how best to generate data to show intervention effects along with mechanisms***

- The next frontier: can this idea or policy scale? What is the science behind that expectation?

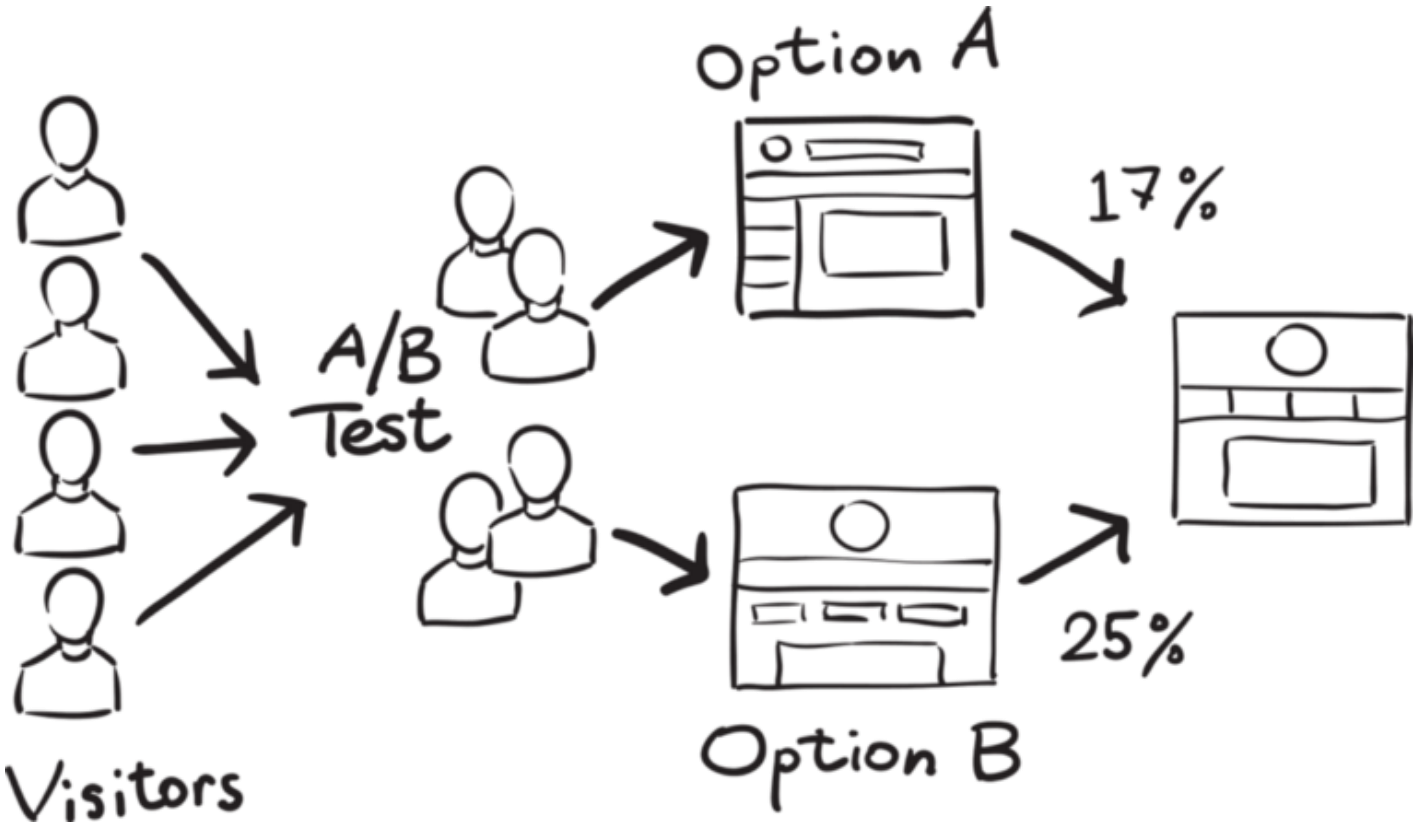
- Should our initial research approach change if we want our program to scale?

Scaling has been a bit like this…



"I think you should be more explicit here in step two."

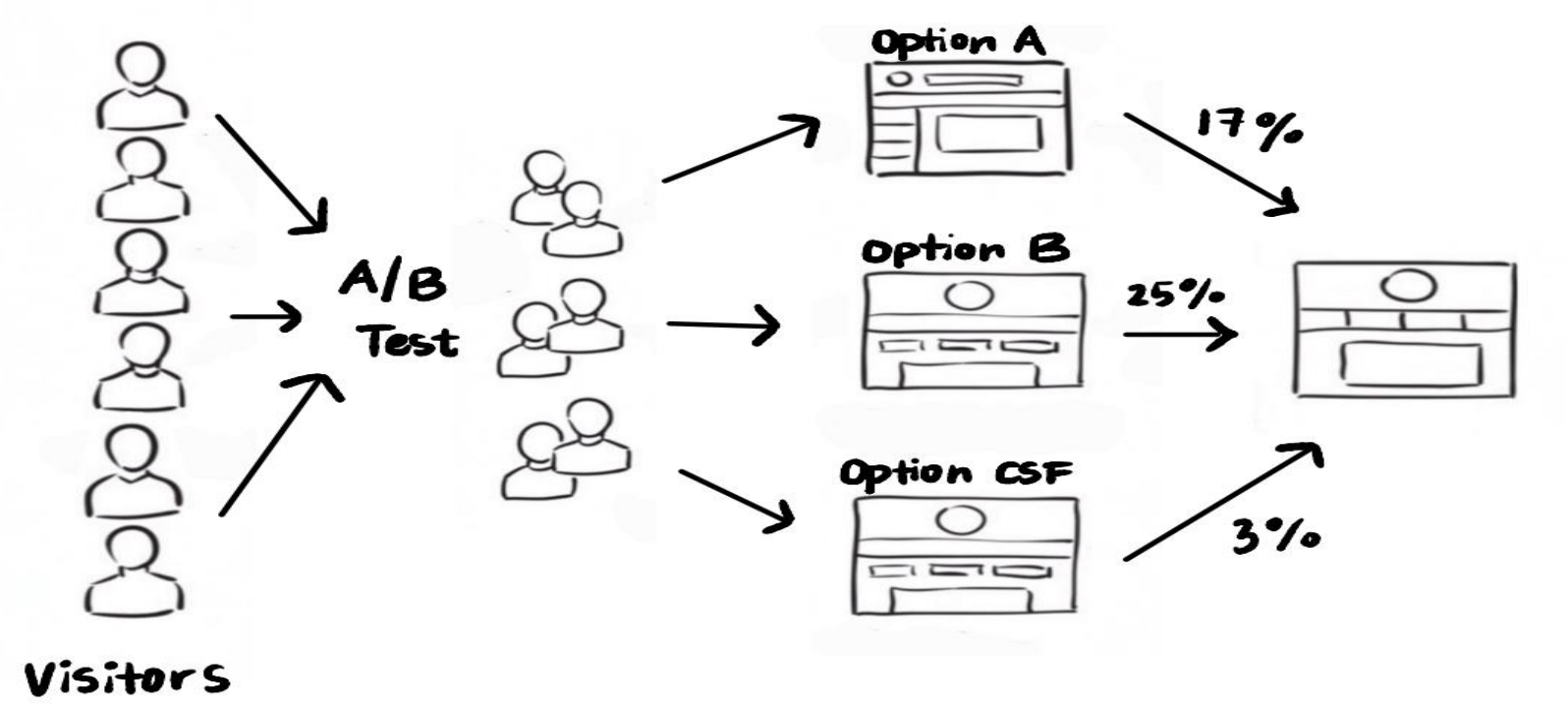
Today's Approach: do A/B with a dash of respondent heterogeneity



Efficacy tests are the wrong approach if you want to change the world at scale

- How do we tackle what is necessary?

Add Option C: Critical Scale Features



Option C = Policy-Based Evidence

Tomorrow's Goal: block on key situational features just like we block on individual characteristics today

➔Policy-Based Evidence

Provocative Prognostication #2

Experimentalists will soon be **banned** *from saying "this has import for policymakers" unless they have policy-based evidence*

F. Labels, Labels, Labels

Data Generation

Harrison and List (2004 JEL)

_____AFE_____FFE_____NFE_____

Lab **[field experiments]** **DID, PSM, IV, STR, etc.**

- conventional lab experiment (Lab)
  - employs a standard subject pool of students, an abstract framing, and an imposed set of rules
- artefactual field experiment (AFE)
  - same as a conventional lab experiment but with a non-standard subject pool
- framed field experiment (FFE)
  - same as an artefactual field experiment but with field context in the commodity, task, information, stakes, time frame, etc.
- natural field experiment (NFE)
  - same as a framed field experiment but where the environment is the one that the subjects naturally undertake these tasks, such that the subjects do not know that they are in an experiment

External Validity & Potential Outcomes

- Estimating $\tau = E[Y_i(1) - Y_i(0)|P_i = 1, E, A, X_i]$

- Everything is CATE. But conditional on what?
  - Environment E
  - Awareness of experiment A
  - Heterogeneity Xi
  - Individuals out/in the study $P \in \{0,1\}$

Randomization over p=1 group

- This yields a treatment effect estimate of the mean outcome differences between participants and non-participants from the $p = 1$ group.

Lab

Select University

AFE + FFE

Select Market

Lab

Select University

Participate in Experiment (P = 1)

AFE + FFE

Select Market

Participate in Experiment (P = 1)

Lab

Select University

P = 1

What about P = 0 People?

AFE + FFE

Select Market

P = 1

What about P = 0 People?

Generalizing to P = 0 Types

- How can we generalize lab, AFEs, and FFEs to the non-participating ($p = 0$) population?
- Some scholars use structural assumptions

Treatment-Specific Selection Bias

- If a subject has a trait that is correlated with participation status and with the outcome variable via treatment, then generalization is frustrated through *treatment-specific selection bias*

Treatment-Specific Selection Bias - Example

- Parents who believe *Head Start* will have a significant effect on their child may be more likely to enroll, and if those beliefs are true, then the subjects who participate in the experiment differ from those who do not participate (and generally so will τ)

- Typical selection bias refers to differences in outcomes for participants and non-participants in the non-treated state. Now, we are referring to differences in the treated state

Generalizing to P = 0 Types

- We can go beyond structural assumptions
- Use actual data from NFEs to tackle this challenge

Lab

Select University

Participate (P = 1)

What about P = 0?

AFE + FFE

Select Market

Participate (P = 1)

What about P = 0?

NFE

Lab

Select University

Participate (P = 1)

What about P = 0?

AFE + FFE

Select Market

Participate (P = 1)

What about P = 0?

NFE

Select Market

Lab

Select University

Participate (P = 1)

What about P = 0?

AFE + FFE

Select Market

Participate (P = 1)

What about P = 0?

NFE

Select Market

Randomly Chosen From

Lab

Select University

Participate (P = 1)

What about P = 0?

AFE + FFE

Select Market

Participate (P = 1)

What about P = 0?

NFE

Select Market

Randomly Chosen From

P=1 & P=0 People

Lab

Select University

Participate (P = 1)

What about P = 0?

AFE + FFE

Select Market

Participate (P = 1)

What about P = 0?

NFE

Select Market

Randomly Chosen From

P=1 & P=0 People

Strong Complements in Parameters Estimated Across Spectrum

An NFE provides *t* for the entire population:

$\tau = y_1 - y_0$

Lab, AFEs, and FFEs provide *t* for p = 1 group

Beyond sorting/selection effects, there might be interesting behavioral effects across the empirical spectrum, as they estimate different parameters (see Al-Ubaydli and List, On The Generalizability of Experimental Results in Economics, 2012)

F. Labels, Labels, Labels

	Laboratory	AFE	FFE	NFE
Population	College Students S=0	Population of Interest S=1	Population of Interest S=1	Population of Interest S=1
Environment	Artificial E=0	Artificial E=0	Natural E=1	Natural E=1
Selection into experiment / Awareness	Overt A=1	Overt A=1	Overt A=1	Covert A=0
Who do we observe?	Those Sorting into Experiment P=1	Those Sorting into Experiment P=1	Those Sorting into Experiment P=1	All in Market P = 0 and P=1

Provocative Prognostication #2

Field experimentalists will be **banned** *from using the term "**lab-in-the-field**" because it is a confusing term that combines AFEs and FFEs in a way that does not permit the reader to understand the relevant parameter estimated from the empirical exercise*

G. Generalizability

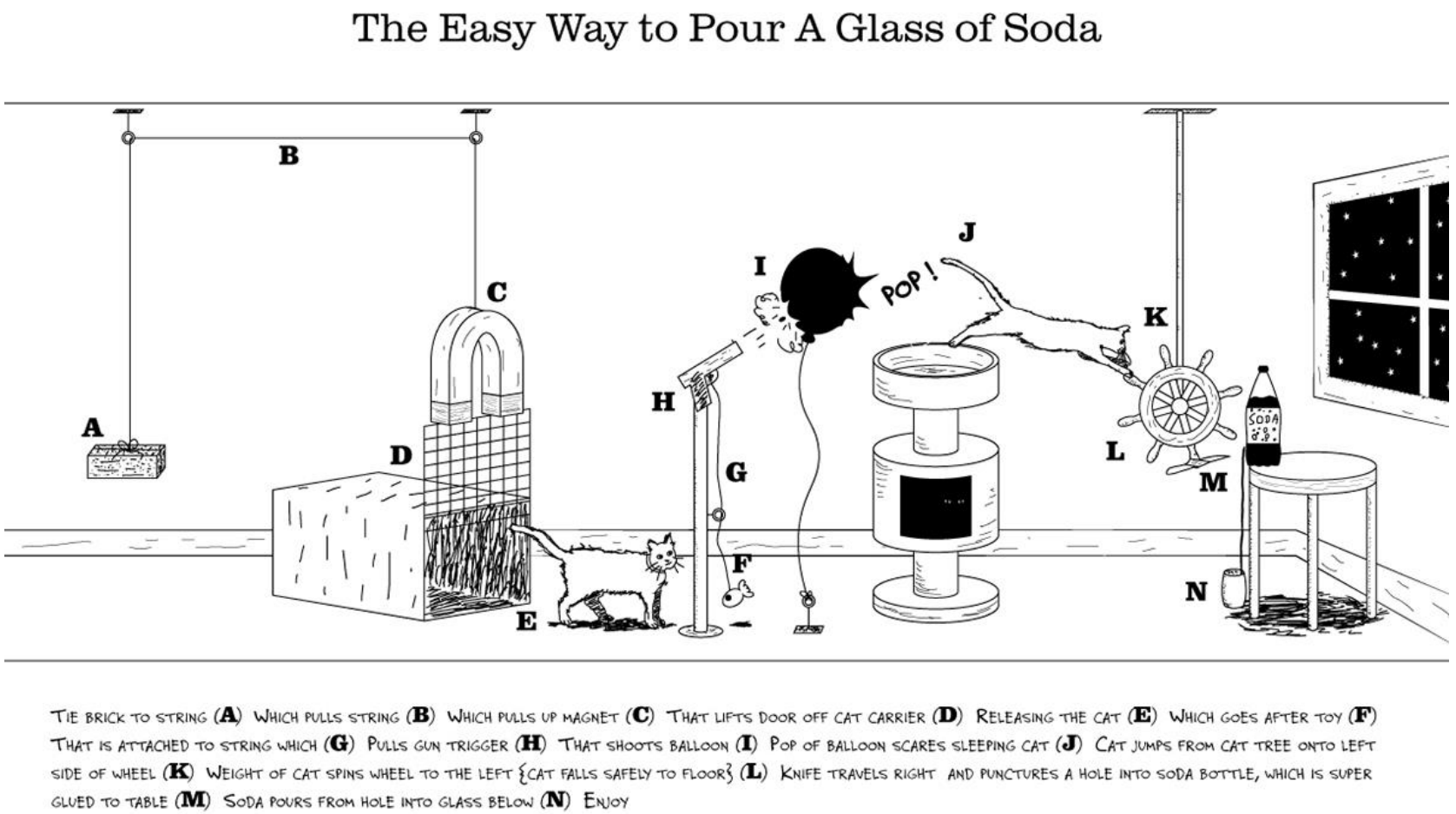
- Understanding mediators and moderators can help

but…..always keep in mind that…..

Every empirical result is externally valid to some setting, no empirical result is externally valid to all settings

Tying a string to a brick causes the glass to be filled with soda

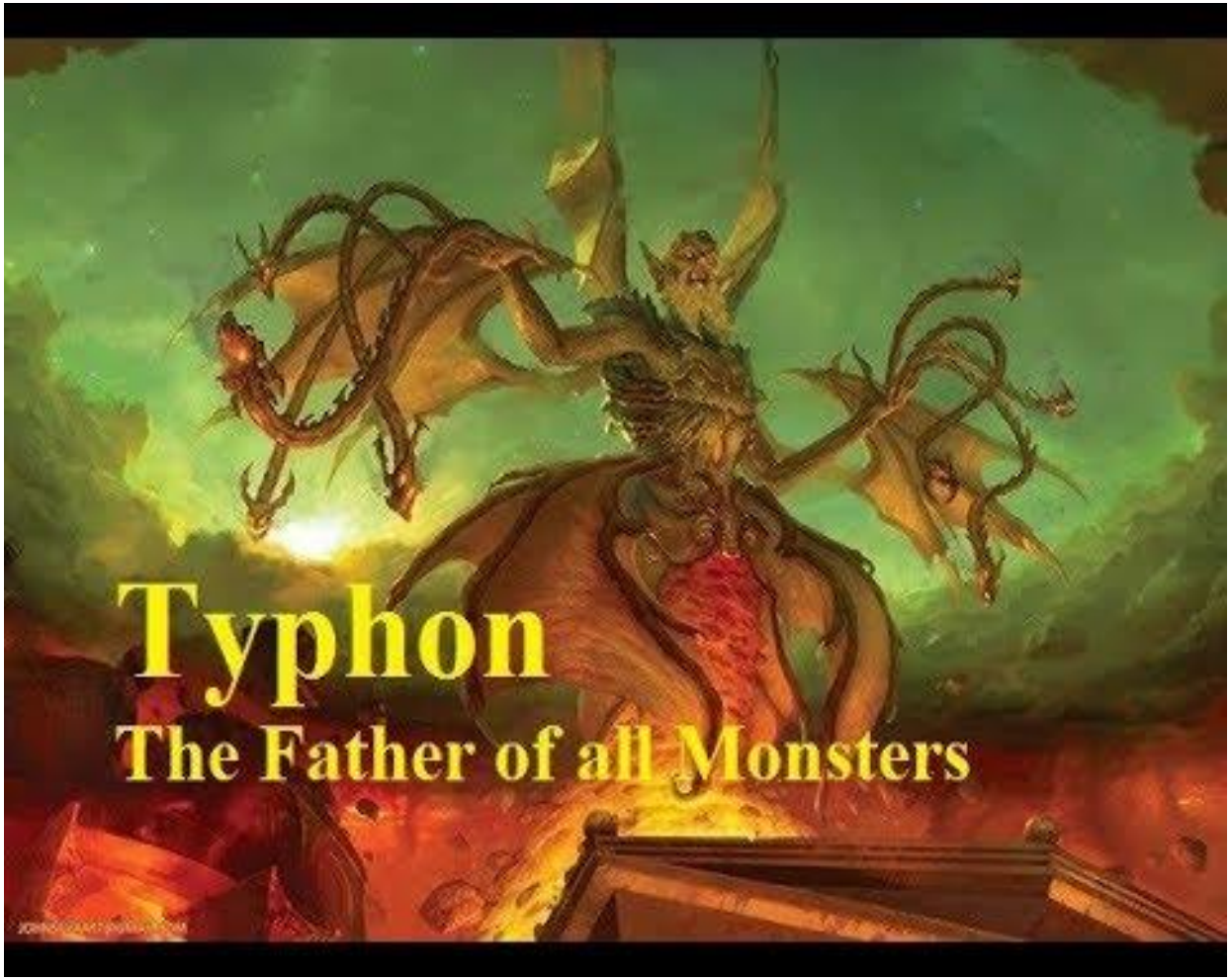
Understanding Mediation Helps in Generalizability



Understanding Moderation Helps in Generalizability

CHECC

Chicago Heights Early Childhood Center


Typhon
The Father of all Monsters

What is Typhon to Experimentalists Trying to Publish These Days?

EXTERNAL VALIDITY

External Validity

- More than 9 of 10 experimental studies that I handle at the JPE receive this comment in some form
- If we really want to stop all empiricism in top journals we should maintain our absent-minded complaining about this issue.
  - The truth is that if you really want, EVERY empirical study can be rejected based on this complaint (time, situation, and/or space will do you in every time)
  - Experimenters are especially susceptible because we have identification figured out (with naturally-occurring data, identification tends to be the showdown)

What to do?
Tackling it Head-On

- A satirical take:

List, John A. (2020) "Non est Disputandum de Generalizability? A Glimpse into The External Validity Trial"

Author Onus Probandi

- Much like we have features that give confidence of internal validity in our work, we need the same with EV
- Economic approach to EV provides 4 reporting areas that give insight into when preferences, beliefs, or individual constraints might vary importantly across populations of people, settings, situations, and time
- Our papers should report on these 4 areas just as we do with IV

An Example: $100 Million Nudge Study

Concerning generalizability of our empirical results, we follow the List (2020) 4 SANS conditions in our reporting**. First, in terms of selection, our sample is a subset of the firms and self-employed taxpaying entities in the Dominican Republic**. In particular, the IRSDR sent messages to a sample of more than 84,000 self-employed and firms that reported earning more revenue than the average firm in the Dominican Republic in the previous year. **In terms of attrition, our compliance rates are 100%, as we have records of the a**mount of taxes paid for everyone in our sample. **Considering naturalness** of the choice task, setting, and time frame, we use a natural field experiment (see Harrison and List (2004)), thus our setting is one in which subjects are engaged in a natural task and are not placed on an artificial margin.

Finally, in **terms of scaling our insights**, the results suggest that the benefit cost profile should shrink slightly as the program is extended to the remaining population of the Dominican Republic. This is because we have slightly larger firms in our experiment compared to the overall firm population. **Since we view the firm size results as a WAVE1 insight, in the nomenclature of List** (2020), replications need to be completed to understand if the size result can be applied to other tax paying populations as well as large players in other markets.

A Thought Experiment on Uniqueness

- EV Litmus Test: can you confidently say that it is difficult/impossible to find a better setting than the one studied in your paper to test convincingly the relevance of your conjectures?

- Many critics view unique settings as a negative distraction. This is the exact opposite way to think of the issue: if the unique setting itself allows you to do the relevant test at hand and no other setting can, then you have found the *perfect* domain for your study.

Provocative Prognostication #3

The SANS conditions will end the EV lunacy once and for all

If I have offended anyone, I apologize!