

Will this Intervention Work in this Population? Designing Randomized Trials for Generalization

Elizabeth Tipton

Associate Professor of Statistics

Co-Director, Statistics for Evidence-Based Policy and Practice (STEPP) Center

Faculty Fellow, Institute for Policy Research

Northwestern University

Randomized trials are now common

Over the past 20 years:

The **Institute for Education Sciences** has funded close to 500 Goal 3 studies and Goal 4 studies. *Results from these trials are published in the What Works Clearinghouse.*

“Nudge” experiments are becoming common in economics and psychology. *In 2015, President Obama issued an Executive Order calling for policy makers in government to apply findings from these trials to design better policies.*

The World Bank and JPAL have funded over 200 randomized studies of education, social welfare, and health interventions in **developing countries**. *Results from these studies are used to guide development policy.*

But they aren't necessarily ideal

- Cluster randomized, with a small # of clusters (< 50)
- Simple design:
 - 2-arm design (50/50 T/C) with business-as-usual control
- Led by teams of experts in interventions, not statistics
- Designed to meet WWC (or other) simple guidelines

But, most importantly for this talk:

They take place nearly entirely in samples purely of convenience.

Sample ATE \neq Population ATE

Clearly if treatment impacts vary and the sample differs in distribution on moderators underlying this variation:

$$\mathbf{PATE \neq SATE}$$

This is not simply academic:

- This bias can be on the same order as bias from non-random treatment assignment.¹
- There is increasing evidence that target populations and samples differ.²

1. Bell, Olsen, Orr & Stuart (2016) 2. Tipton (2015), Tipton et al (2016), Fellers (2017)

Post-hoc corrections

In some cases, we can adjust ATE estimates based upon population information, using :

- Propensity score post-stratification³⁴
- Propensity score inverse probability weighting⁵
- Maximum entropy weighting⁶
- Use of bounding approaches⁷

But there are limitations

The effectiveness of these methods is limited in practice because of **under-coverage**.

If the population included a subset of units not represented in the sample (i.e., probability of selection = 0), no amount of statistical adjustment will solve this.

For this reason, much of my research asks:

How can we design better trials so that results **do generalize to policy-relevant populations?**

What is the status quo?

Typical trials

Units (often sites – e.g., schools – and individuals) are targeted for recruitment by researchers. In education, these schools tend to be:

- In districts with many schools.
- Near research hubs and major cities.
- In prior relationships with researchers.

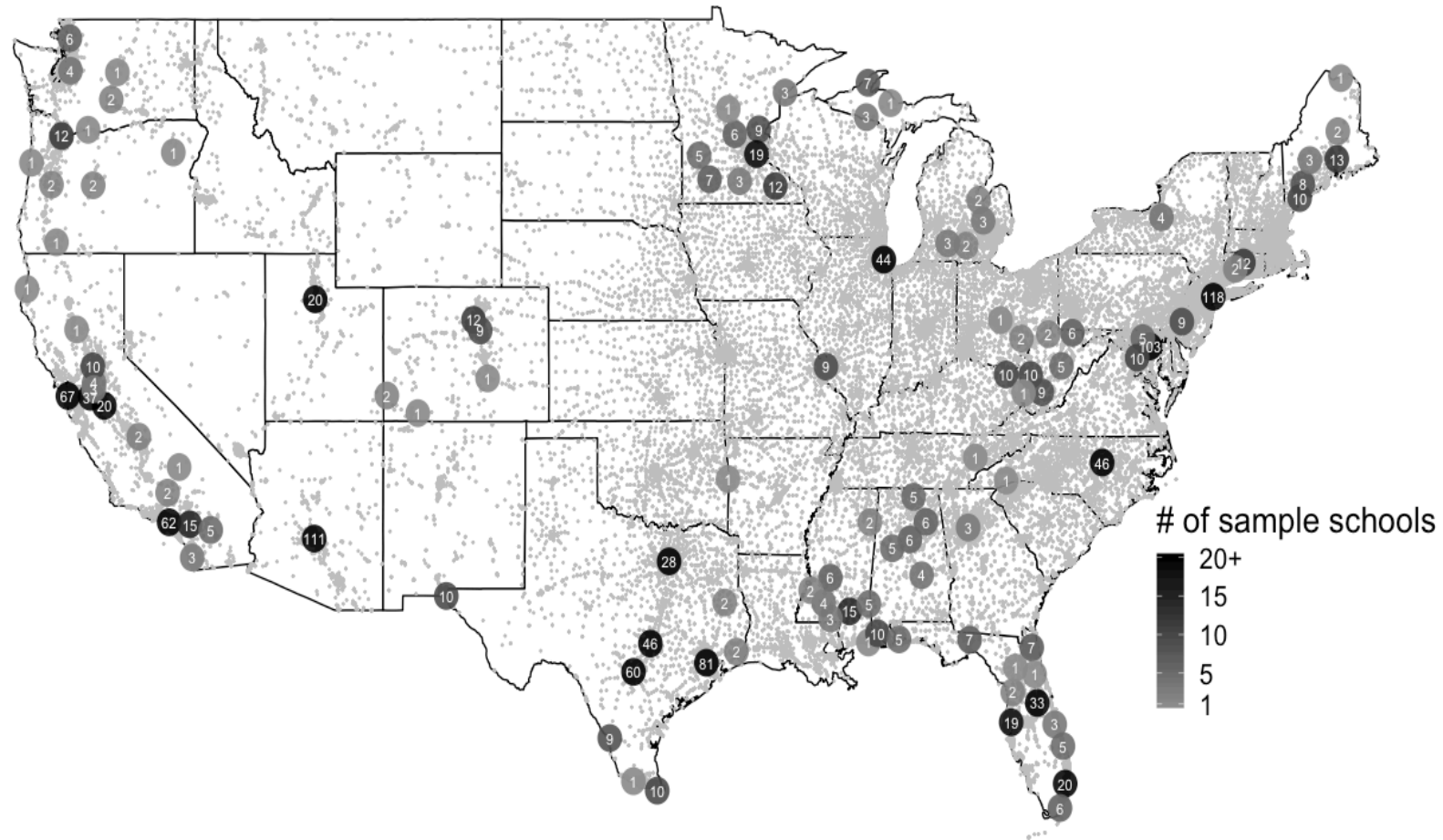
Once contacted, many units decline participation:

- Perhaps another program in place.
- Don't have capacity – e.g., leadership change.
- Never respond to phone calls.

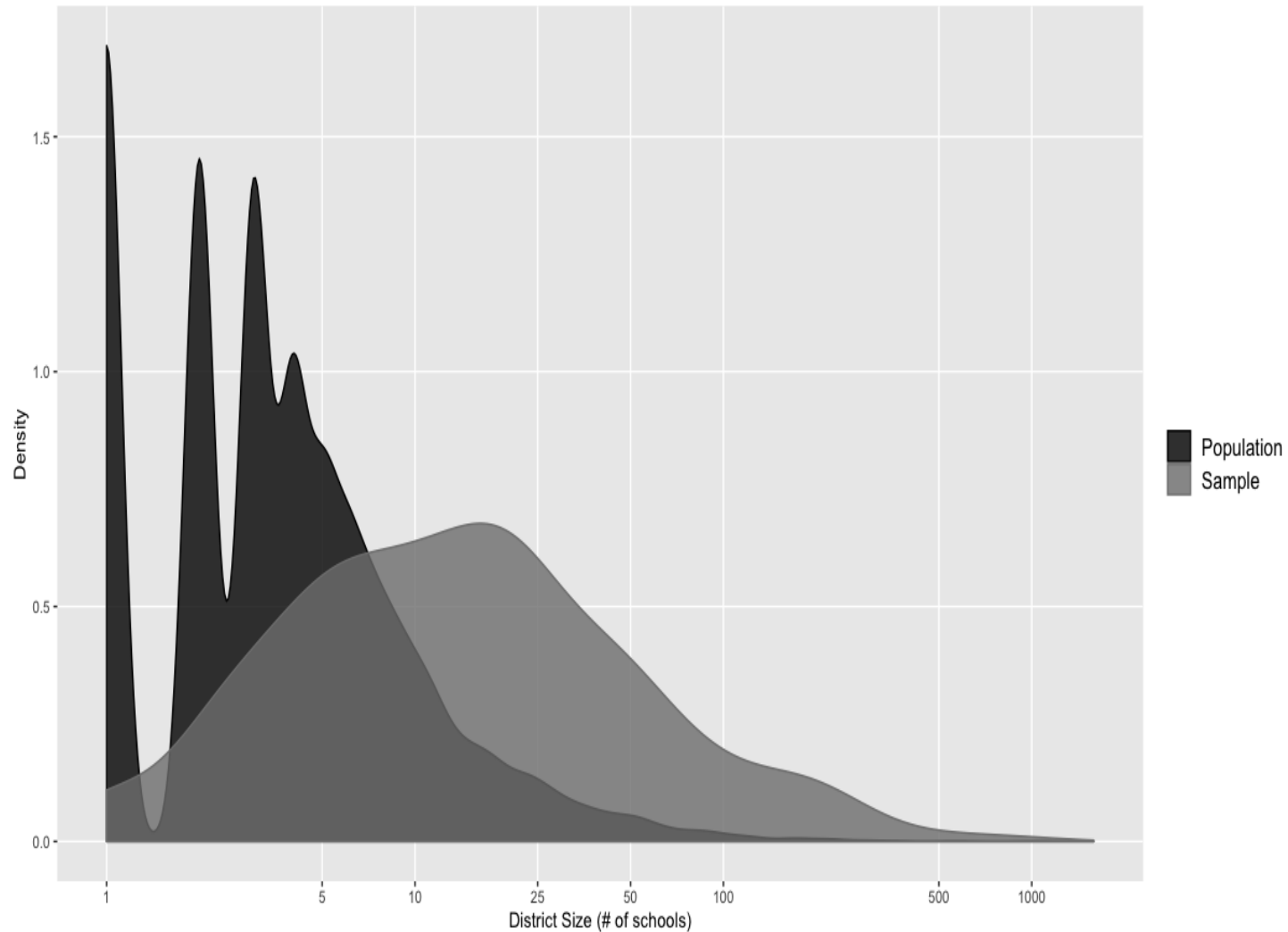
Most importantly, **very little information on this process is recorded.**

The study “begins” once units are recruited. **“Statistics” begins at random assignment.**

Schools in 34 IES funded studies (2010-2015)



School district size matters



What would statisticians like?

If the goal is to estimate a PATE, researchers would define clearly:

- A target population;
- Eligibility criteria for their study;
- Estimands of interest (and priorities);
- Resource constraints and recruitment strategies;
- Discussion of possible sources of treatment effect heterogeneity.

And then develop a sample selection strategy to estimate the PATE.

Don't we already know this?

While this problem seems very similar to designing a sample survey, there are some unique issues:

- The sample size is **typically small** (< 60 schools);
- The estimand is a **difference** – the ATE – and thus variation in treatment effects, not outcomes, matters;
- **Refusal rates** are high and *recruitment has typically been ad hoc*;

And, most importantly:

*Those planning these experiments are **not used** to thinking about generalization.*

What do we need?

1. Empirical work that convinces those planning (and funding) trials that **generalization is a real problem.**
2. Understand the **constraints** affecting this research:
 - **Data:** Target population data is messy
 - **Knowledge:** No training in populations, demography, or sampling
 - **Time:** Proposals require sample (before funding)
 - **Money:** Recruitment is expensive
3. Develop **simple methods and tools** that can improve this research.

The Generalizer

Sampling ignorability: The Goal

We want a sample selection strategy such that:

- The sample that results is closer to being a “miniature” of the population
- On the set of covariates the explain variation in treatment impacts.

This last condition is the **Sampling Ignorability** condition in generalization:

$$\Delta = [Y(1) - Y(0)] \perp Z | \mathbf{X}$$

Target population

What data can we use?

- In **education**, there are the Common Core of Data, as well as Census data (district level), and state longitudinal data systems.⁹
- In **social welfare**, there are Census data, the American Community Survey, data from the Bureau of Labor Statistics, etc.¹⁰

Population definitions can be **broad or narrow**.

9. Tipton (2014); 10. Tipton & Peck (2017)

The magic of stratification

In both design-based and model-based sampling, stratification is an important tool for reducing variance / increasing similarity.

But now we seek such balance on **a large set of covariates**, those that may explain variation in treatment impacts:

- We aren't sure in advance which variables explain variation in treatment impacts;
- We'd like good balance between the sample and population, even with small samples.

Example

Category	Covariates
Student	% students ELL
	% students F/RL
	Race/ethnicity of district
	% White
	% Hispanic
	% Black/African American
	% other
Community	Educational attainment
	% Grade 8 or lower
	% <HS grad
	% HS grad
	% Postsecondary
	% 5- to 17-year-olds in poverty
	% labor force

Census area financials
District

Median income (overall)

Urbanicity of districts

% Urban

% Suburban

% Town or rural

Geographic location

% Northeast

% Midwest

% South

% West

District revenue (thousands)

Number of students

in district*

Number of schools

in district*

Constraints

The total number of strata (H) is limited by three factors:

1. Sample sizes are small (typically 20 - 60).
2. Each additional stratum creates an **additional constraint** for those recruiting sites into the study. This creates additional costs in terms of both time, personnel, and money.
3. Recruitment has largely been an **ad hoc process**. This requires formalization of this process, which creates additional costs.

k -means cluster analysis

We have p covariates, but we'd like H strata ($H \ll p$). The goal is for these strata to be close to homogenous on these covariates.

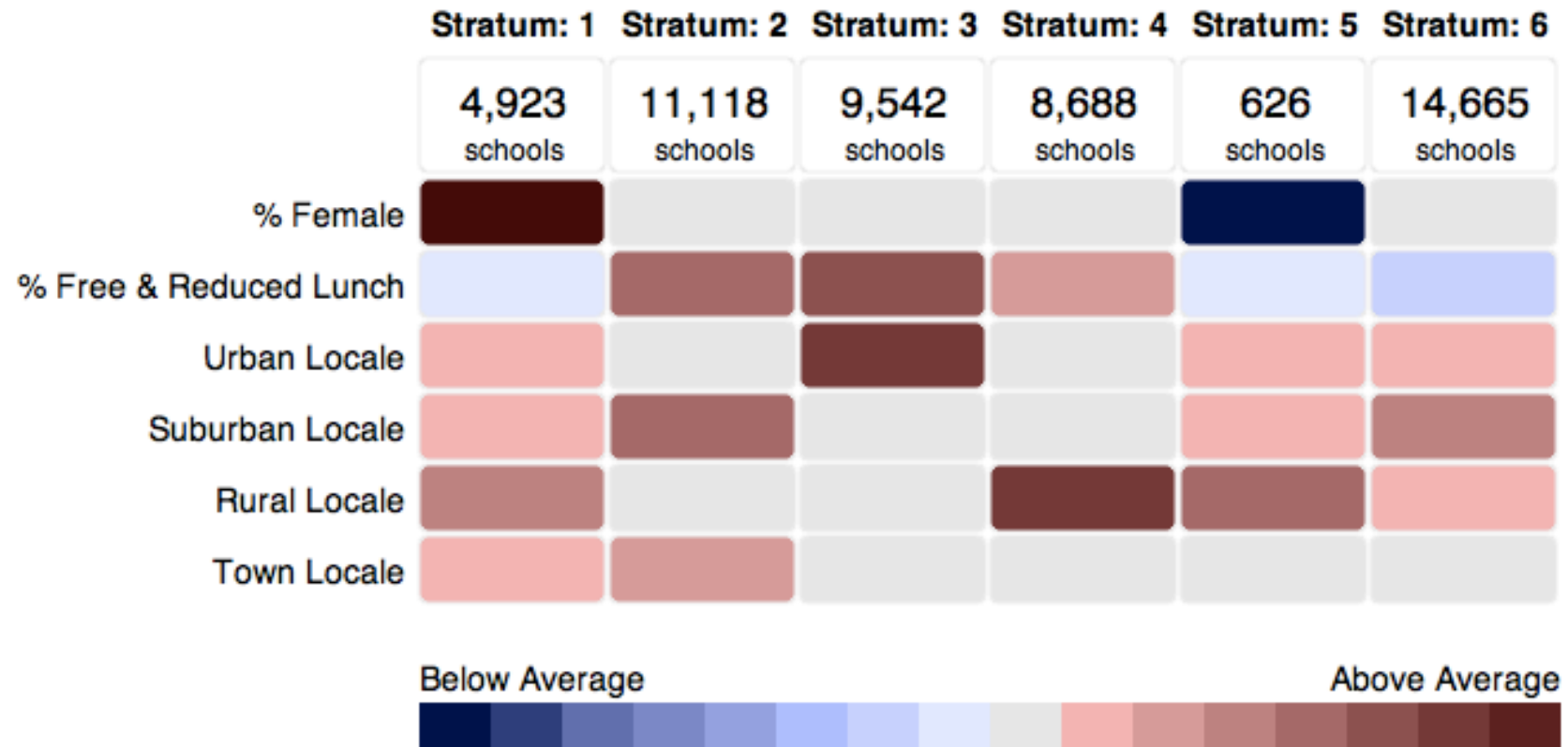
Cluster-analysis¹¹ is one approach:

- k -means with Gower's distance (when there are different covariate types) and standardized covariates.
- Results in strata of different sizes, some more homogenous than others.

In special cases, propensity scores can also be used¹².

11. Tipton (2014) 12. Tipton et al., 2014

Example Strata



Within-stratum recruitment

The total sample n can then be selected from these strata using:

- Proportional allocation
- Neyman allocation

Within each stratum, sites can be selected in a variety of ways:

- Ordered randomly;
- Ordered in terms of Euclidean distance from the stratum average vector;
- Selected based on convenience.

(Honestly, at this stage: Any way is better than current default.)

The Generalizer: A free webtool

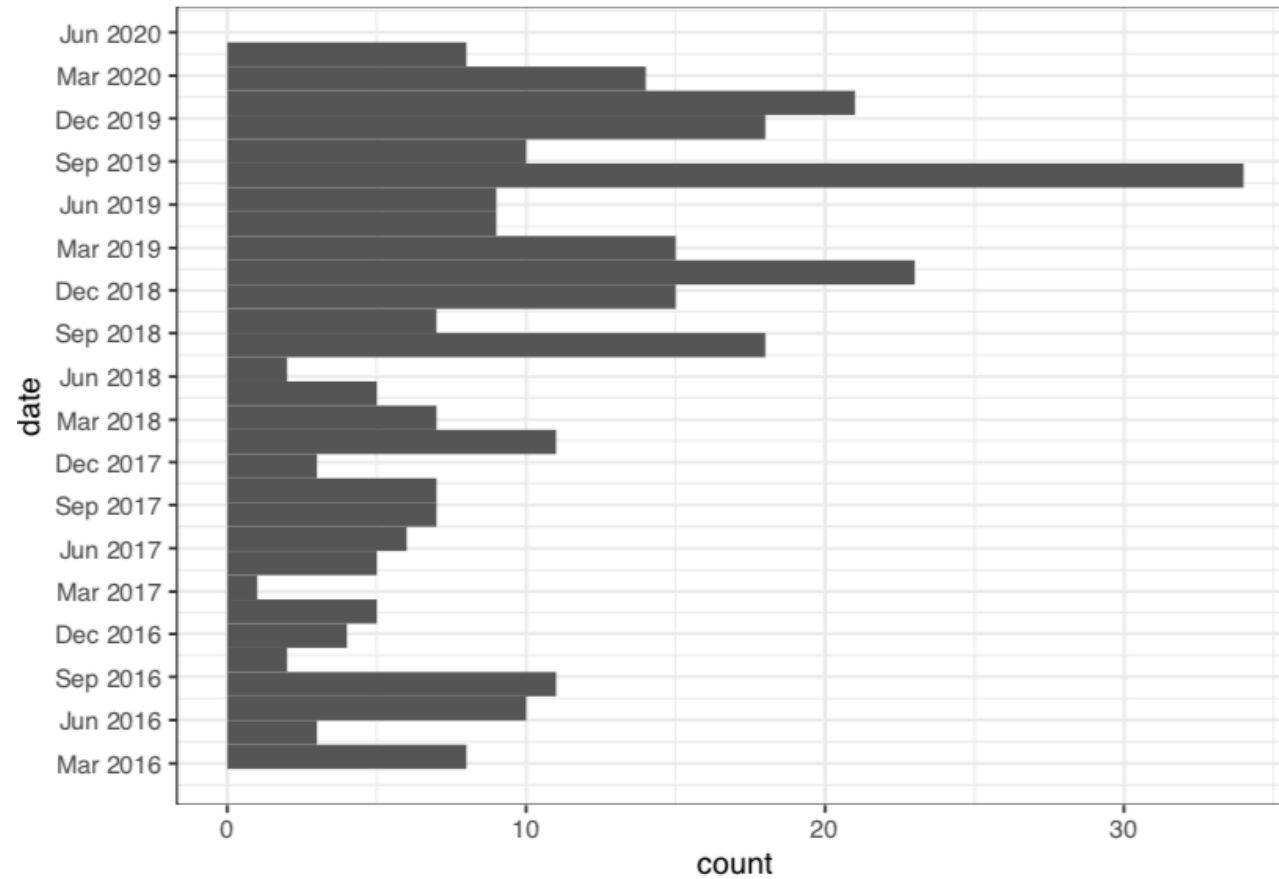
I realized no one would implement these methods if it wasn't easy.
What about a free, easy to use, webtool?

Design principles for the user experience

- Stand-alone (doesn't require users to read anything else)
- Data + Statistics + Teaching
- Make it simple (language, design, errors)
- Make it useful (reports, tables, figures, data)

www.thegeneralizer.org

Usage has increased over time



Examples of education studies using this approach

Year	Intervention	Population	Selection Process
2011	Open Court Reading, Everyday Math	Schools like those using the programs in the US	Purposive
2015	National Study of Learning Mindsets	9 th graders in public HS in the US	Probability
2015	Khan academy in CC	Community colleges in CA	Purposive
2015	ASSISTments	Public middle schools in Maine	Purposive
2015	Reasoning Math	Public middle schools in WV	Purposive
2015	PACT	Public 6 th grade classrooms teaching US history in US	Purposive
2017	ASSISTments	Public middle schools in WV	Purposive
2019	Early Math	Head Start and Public Pre-K in US	Purposive + Probability

What's next?

What about testing moderators?

All methods for generalizing assume that we have a handle on **why treatment effects vary across sites**.

But we don't.

This is because sample sizes in RCTs are based upon the goal of testing hypotheses about the ATE, not about moderators.

Moderator tests are an afterthought and are greatly under-powered.

How are site-level moderators tested?

Assume that we have $i = 1, \dots, n_j$ students in $j = 1, \dots, J$ sites:

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_2 X_j + \gamma_3 T_j * X_j + r_j + e_{ij}$$

Assume also that:

$$e_{ij} \sim N(0, \sigma^2) \text{ and } r_j \sim N(0, \tau_{|X,T}^2).$$

Our goal is to estimate the **Standardized Effect Size Difference (SESD)**:

$$\delta_{ds} = \gamma_3 / \sqrt{(\sigma^2 + \tau^2)},$$

Standardized how?

The SESD is standardized in relation to SD in Y as:

“The effect of a 1-unit change in X on the SD of Y.”

But the scale of X matters:

- If X is dichotomous with probability Q in one group, then

$$V(X) = Q(1 - Q) \leq \frac{1}{4}$$

- If X is continuous, V(X) can be large or small.

The solution I propose is that we should standardize X in relation to the **population standard deviation**.

Minimum detectable SESD

Dong, Spybrook, and Kelcey (2018) provides us with:

$$MDES D(|\delta_{ds}|) = M_v \sqrt{\frac{(1 - R^2_{|X})\rho n + (1 - \rho)}{P(1 - P)S_x^2 J n}}$$

Where:

$M_v = t_{\alpha/2} + t_{1-\beta}$ for two-tailed tests

$\nu = J - 4$

S_x^2 is the variation in X in the sample

J = number of sites

n = average number of units in each site

ρ = ICC

$R^2_{|X}$ = prop. of $V(Y)$ explained by X

P = prop. in treatment

Role of X in the sample

We can rewrite this as:

$$MDES D(|\delta_{ds}|) = \frac{MDES D_p(|\delta_{ds}|)}{r_x}$$

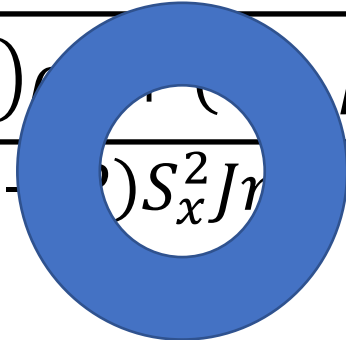
Where:

- $MDES D_p(|\delta_{ds}|)$ is the MDES D when the *sample is selected randomly* from the population.
- $r_x = S_x / \sigma_x$ is the sample SD of X scaled by the population SD.

Thus it is clear that the MDES D is **as effected by the increase in the sample size as by the standard deviation of X.**

Minimum detectable SESD

Dong, Spybrook, and Kelcey (2018) provides us with:

$$MDES D(|\delta_{ds}|) = M_v \sqrt{\frac{(1 - R^2_{|X|}) (1 - \rho)}{P(1 - \rho) S_x^2 J n}}$$


Where:

$M_v = t_{\alpha/2} + t_{1-\beta}$ for two-tailed tests

S_x^2 is the variation in X in the sample

n = average number of units in each site

$R^2_{|X|}$ = prop. of $V(Y)$ explained by X

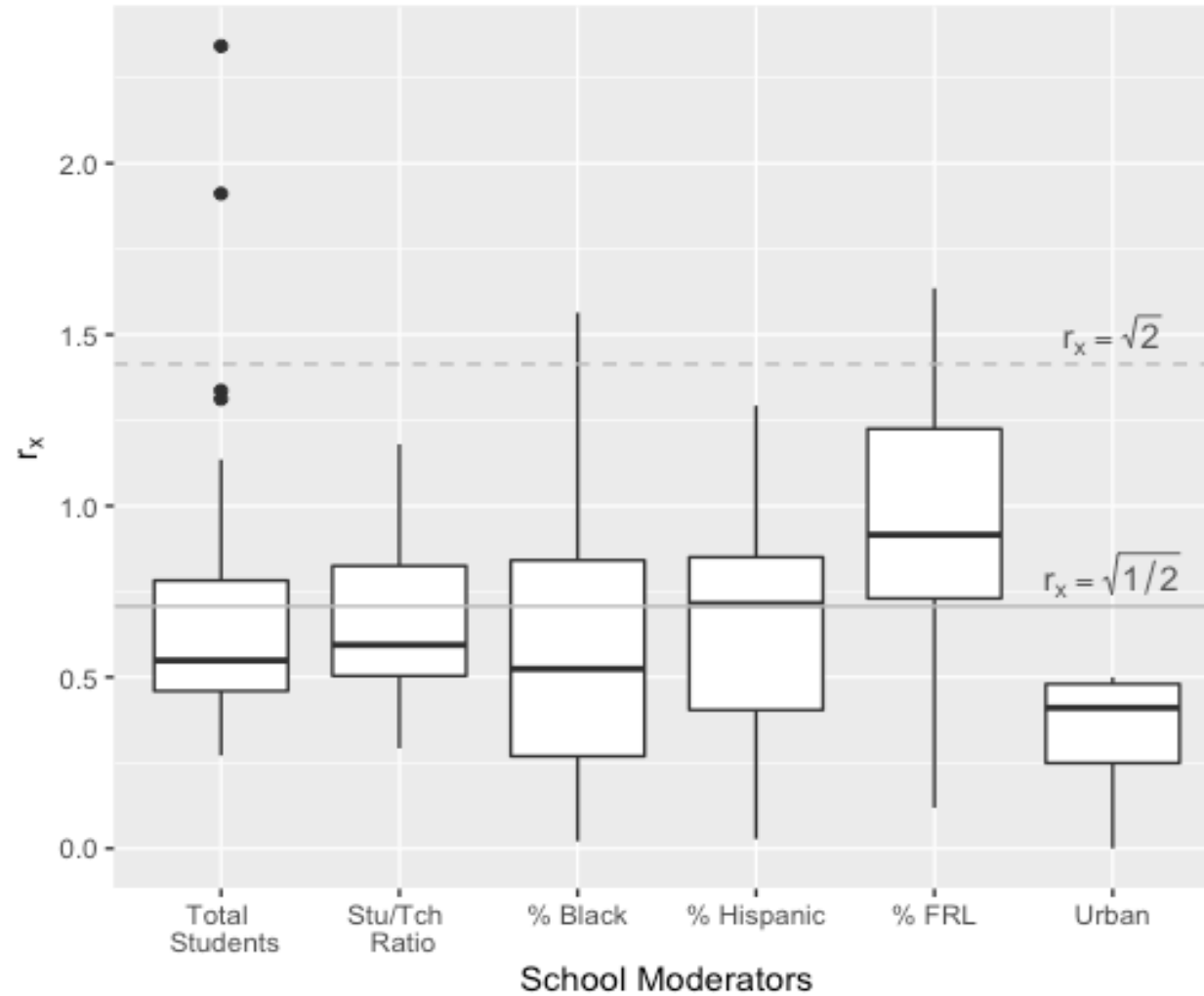
$\nu = J - 4$

J = number of sites

ρ = ICC

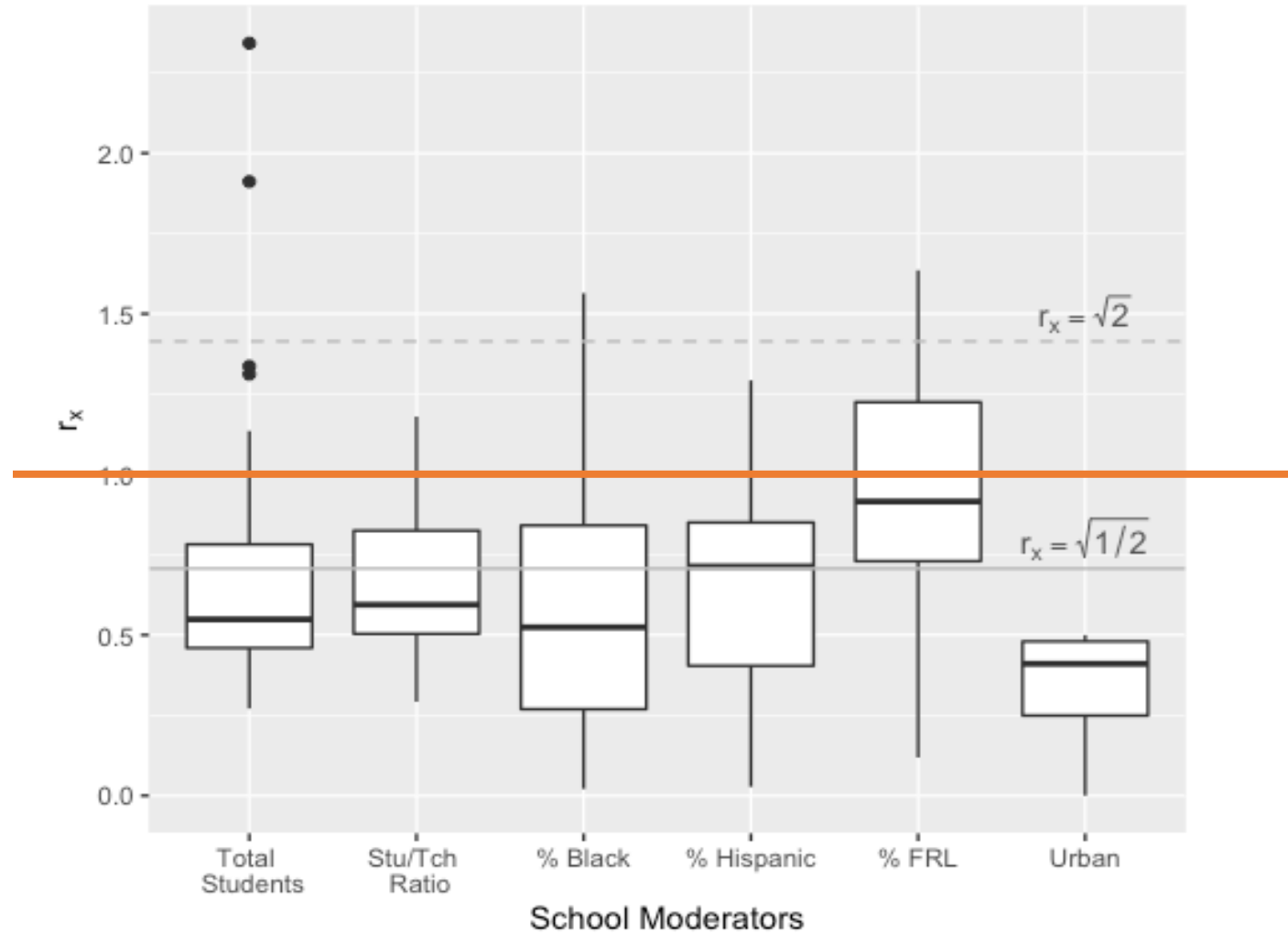
P = prop. in treatment

Example: 19 RCTs in education



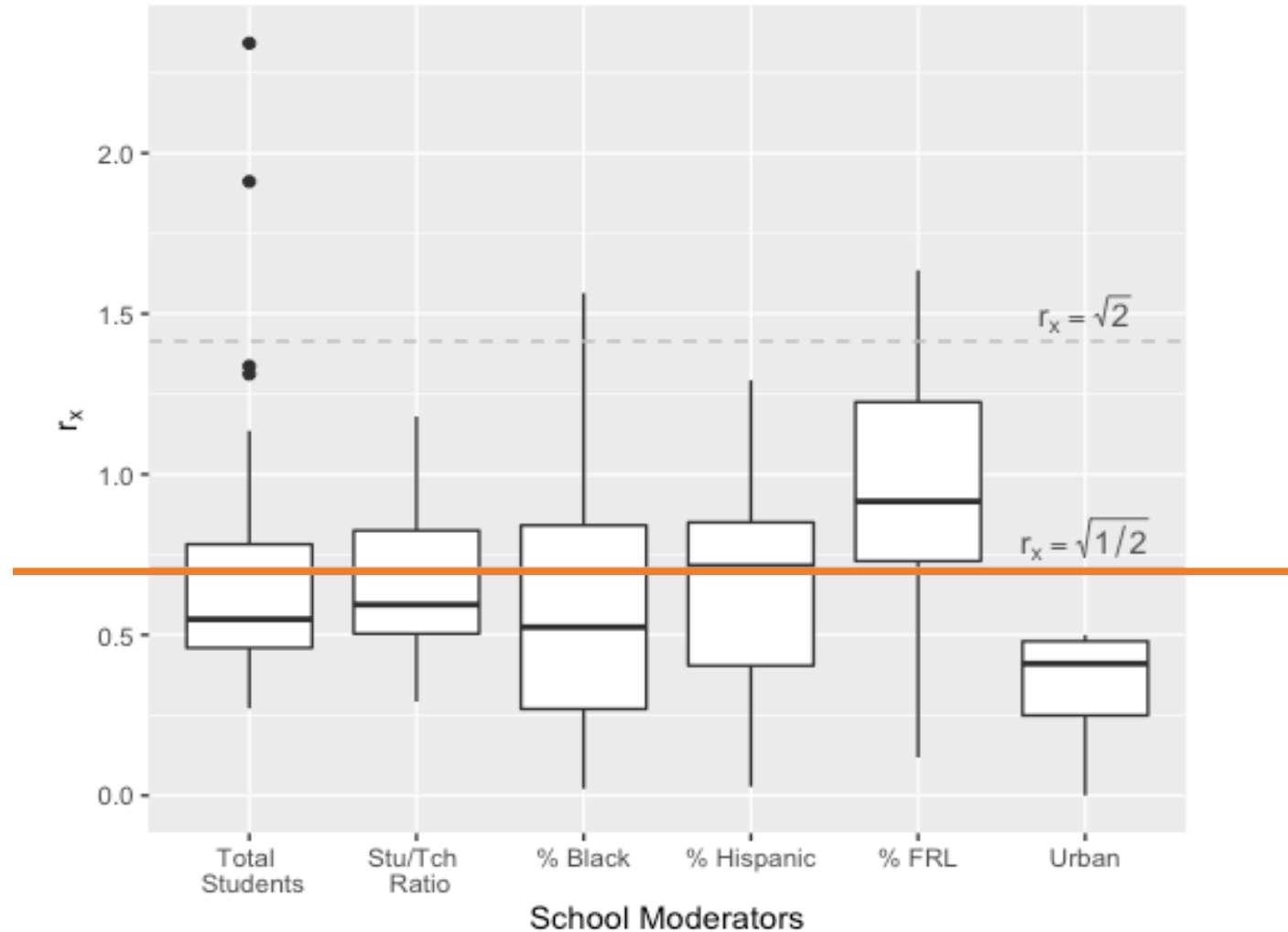
Example: 19 RCTs in education

**Most studies
have small
variation in X**



Example: 19 RCTs in education

Many studies
have *very*
small
variation in X



What can we do to improve power?

The simple answer: Maximize $r_x = \frac{S_x}{\sigma_x}$:

- Recruit sites that are extremely different on the covariate under study.
- Ideally, divide your sample evenly across these extremes.

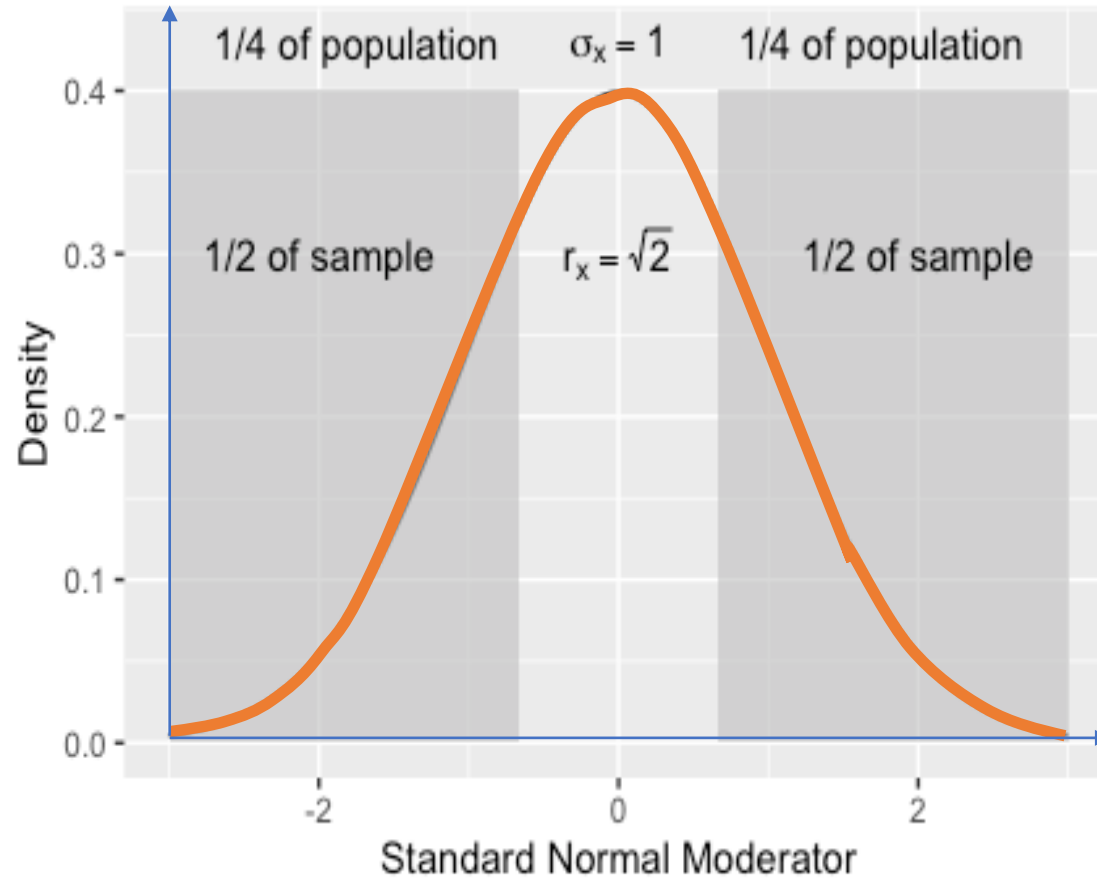
How can we do this?

- Use population data to identify these extremes.
- Use response surface modeling and optimal design methods when there are multiple variables (e.g., D-optimality criteria).

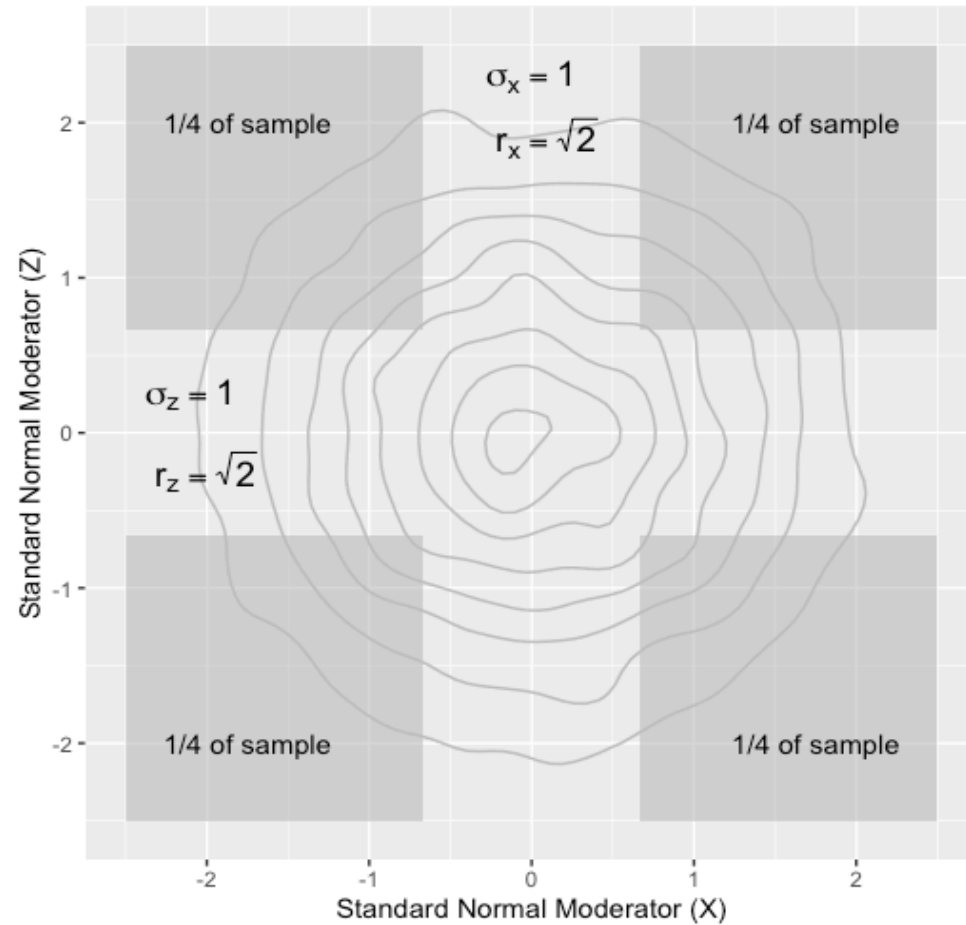
Simplest case: Single continuous moderator



Simplest case: Single continuous moderator

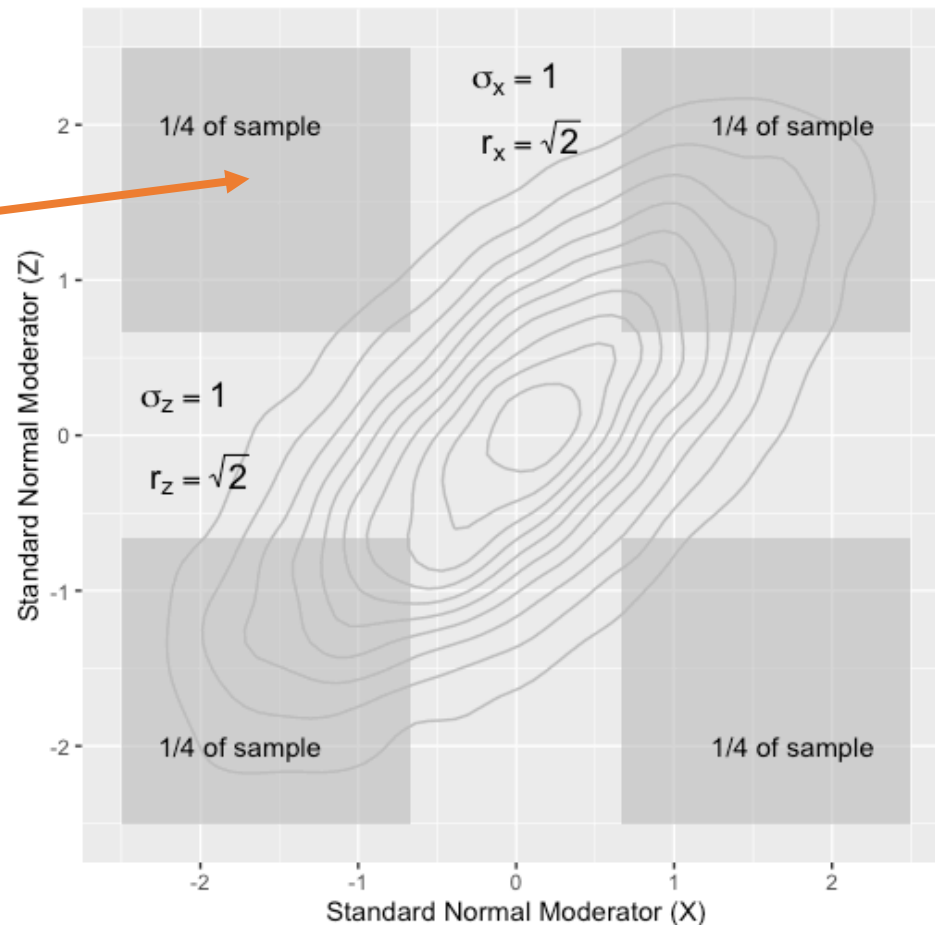


More complex: 2 independent moderators



More complex: 2 correlated moderators

Note: It's really hard to get these (and unlikely to get them without trying)



But what about in real studies?

In industrial experiments, typically the levels of the covariates **X can be chosen by the experimenter.**

But when we're trying to understand treatment effect moderators, **we have to work with the moderator levels observed in the population:**

- The full range of a covariate may not be observed
- Moderators may be highly correlated, making it hard to de-alias
- We have small samples (n) yet we may have several moderators of interest (p), leaving very few replicates in each stratum

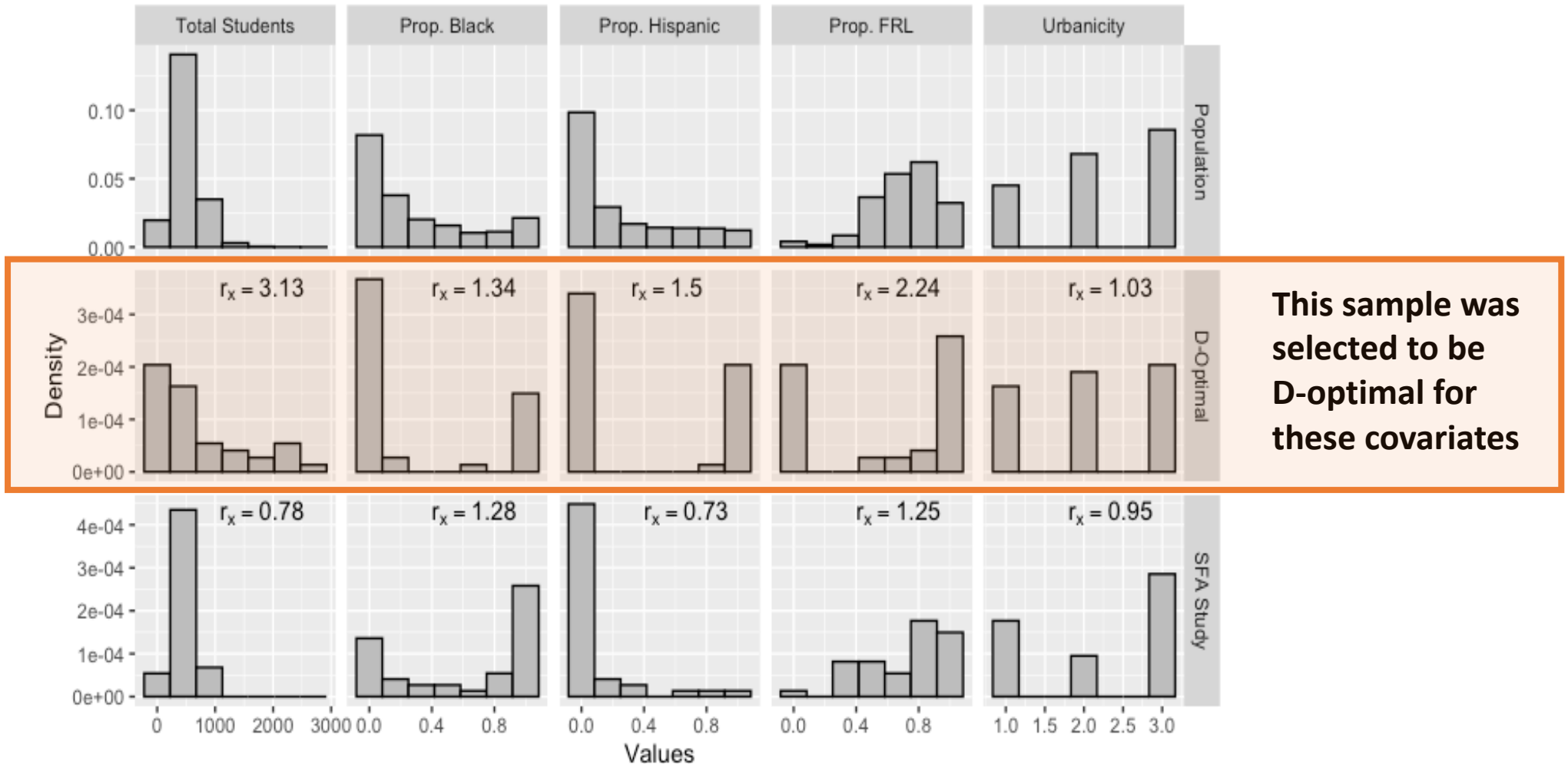
Optimal design principles

In experimental design, there are optimality criteria we can use for these situations:

- **A-optimality**: Minimize the average variance of the estimates of the regression coefficients
- **D-optimality**: Minimize $|(\mathbf{X}'\mathbf{X})^{-1}|$ (which takes into account covariance)

These designs can be implemented in R using the **AlgDesign** package.

Properties of an optimal sample of $n = 40$



Recall that
$$r_x = S_x / \sigma_x$$

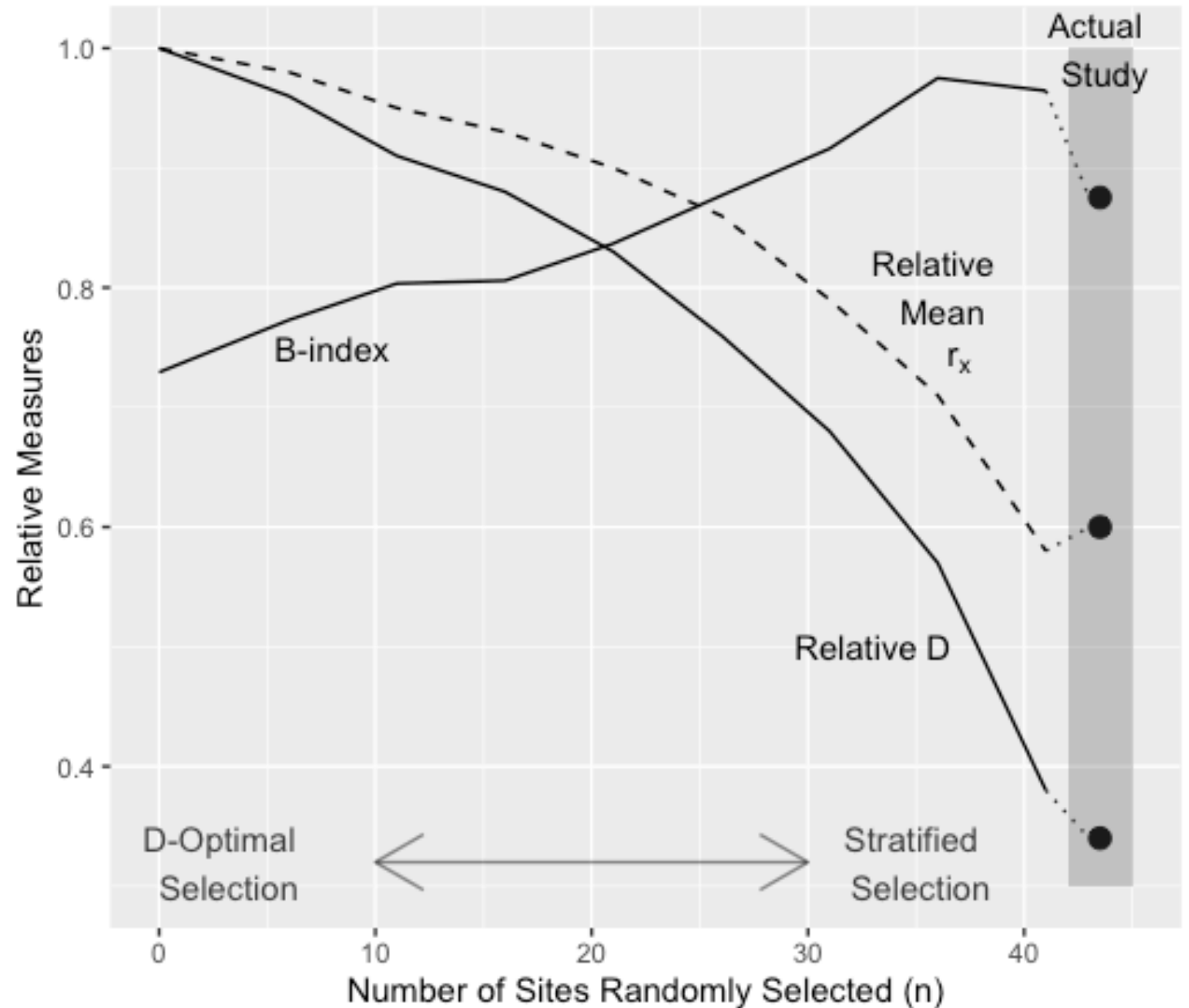
Augmentation strategy

For the PATE, the ideal is to recruit proportional to the population (e.g., strata).

For moderators, the ideal is to recruit the extremes.

These are at odds.

An compromise is to recruit for the PATE for *most* of the sample, but augment this with some extremes.



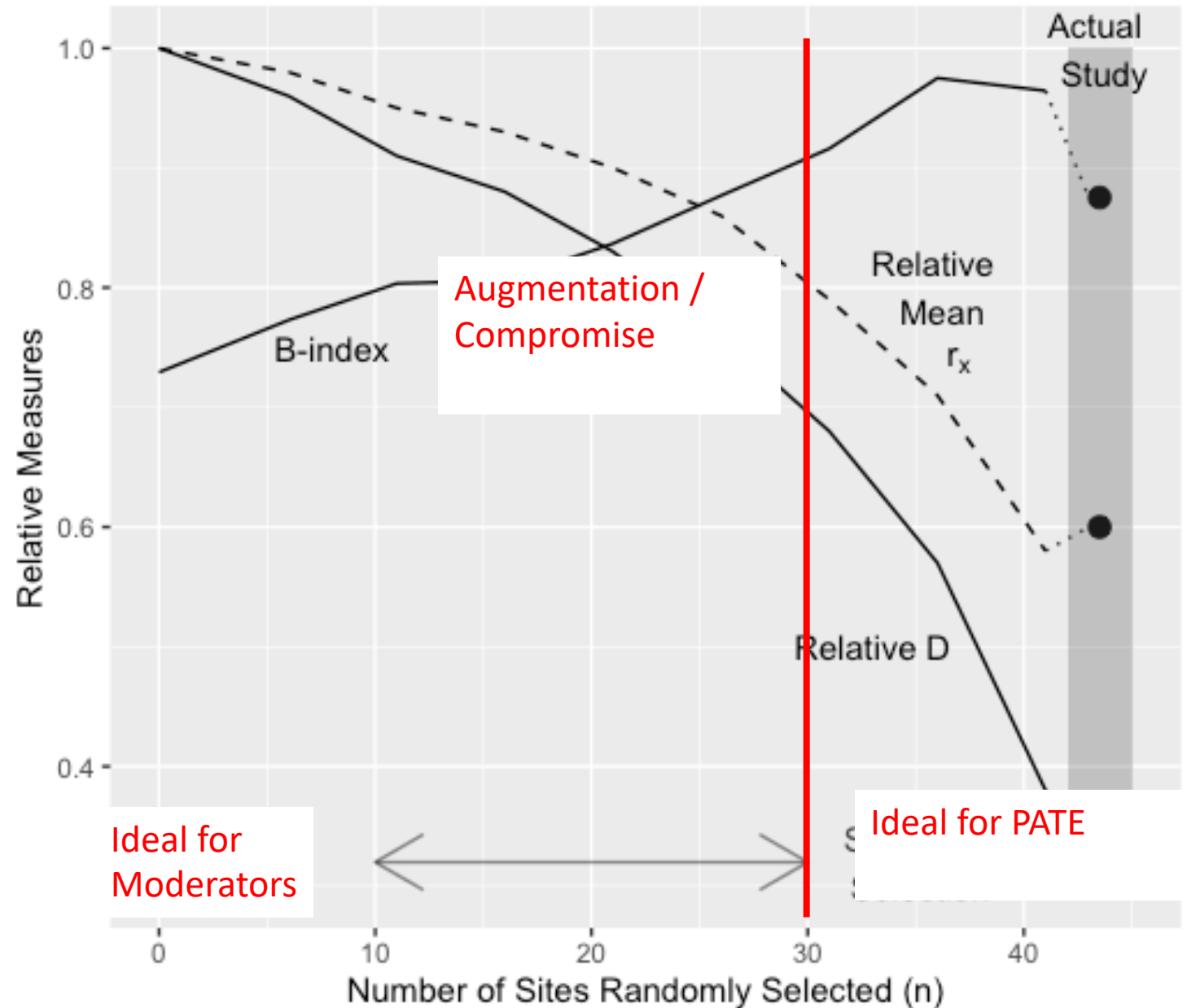
Augmentation strategy

For the PATE, the ideal is to recruit proportional to the population (e.g., strata).

For moderators, the ideal is to recruit the extremes.

These are at odds.

An compromise is to recruit for the PATE for *most* of the sample, but augment this with some extremes.



Conclusion

Take home points

We know as statisticians that:

- The sample matters and that the SATE and PATE can differ.
- Research design matters.

How do we change practice?

- Understand users and their constraints
- Develop simple tools
- Think the “front end” as much as the “back end” in methods development

Thank you!

Elizabeth Tipton

tipton@northwestern.edu

www.bethtipton.com

<https://stepp.center>

@stats_tipton

Abstract

Field experiments are now common in education, development, and the social sciences. Generalizing from the results of a field experiment to a policy relevant population, however, is difficult when the effect of the intervention varies across people and institutions. As a result, statisticians are increasingly interested in the development of methods for generalizing treatment effects, as well as testing moderators of treatment impacts. Yet much of this methodological development has focused on analytic approaches, neglecting the role that the sample itself plays in these analyses. This is particularly important given that nearly all field experiments are conducted in samples of convenience. In this talk, a practical approach to recruitment and sample selection is introduced that is population focused and easily implementable when population data is available. This sample selection approach is extended to include optimal designs for estimation of treatment effect moderators. Throughout the talk, examples from education and psychology experiments will be included.