The role of sampling and recruitment in improving the policy relevance of RCTs

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

Impact evaluations are hard, but great

Randomization is straightforward to do and easy to understand.

As a result, the intervention and comparison groups are equivalent on all *observed* and *unobserved* factors.

The results can be summarized succinctly in terms of the **average treatment effect.**

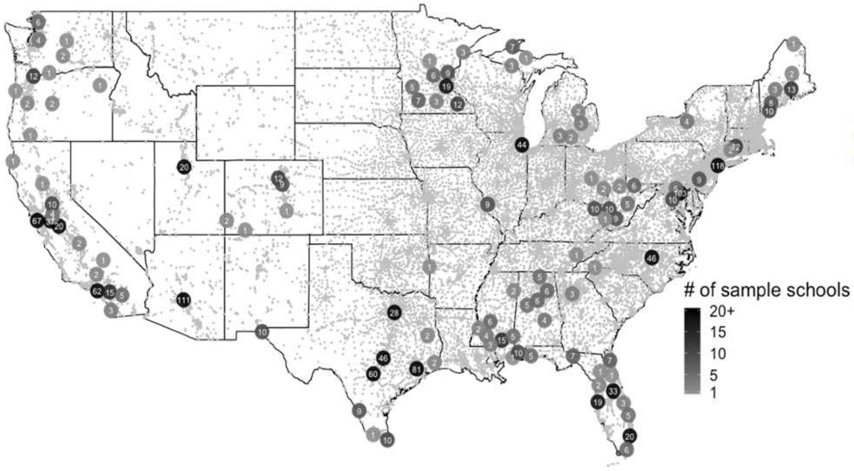
But not without problems

Impact evaluations are nearly always conducted on **samples of convenience**.

It is likely that the effects of an intervention **vary**.

This means that the intervention may be effective for some subgroups and *not* others.

e mean?



t **Δ** is simply the **weighted average of**

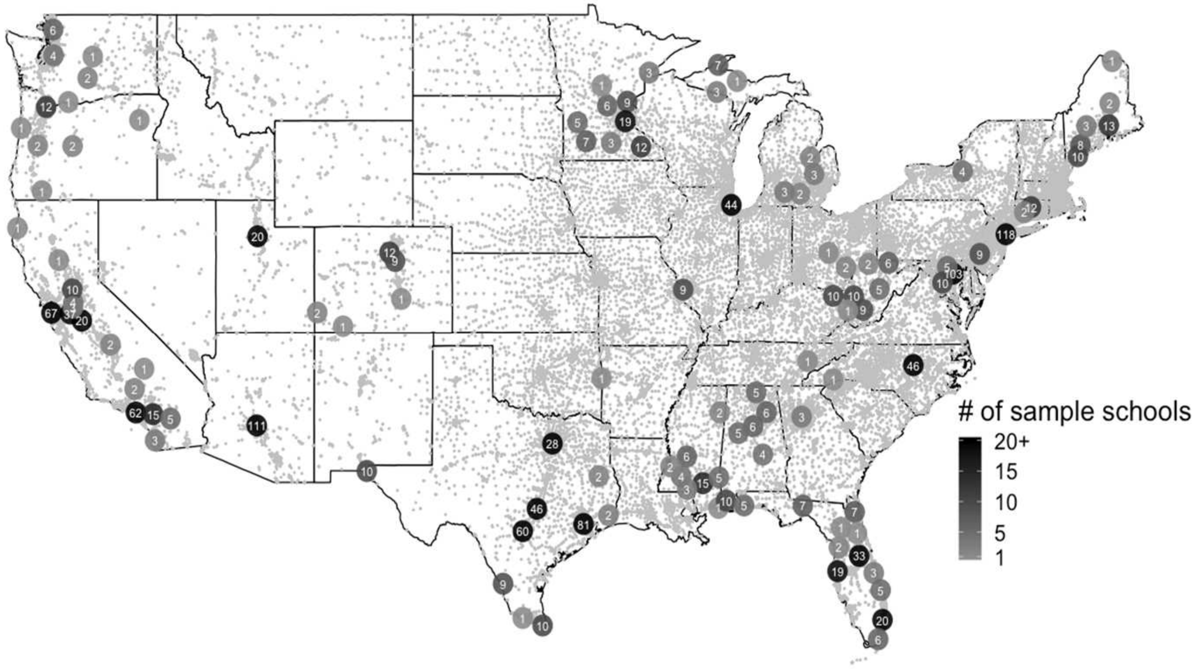
$$+ w_2\Delta_2 + \ldots + w_p\Delta_p$$

Unless the treatment effect is constant (or the sample is a representative), **Δ** will depend on the sample.

Thus, in general the SATE and PATE differ: $\Delta_s \neq \Delta_p$

Samples differ from target populations

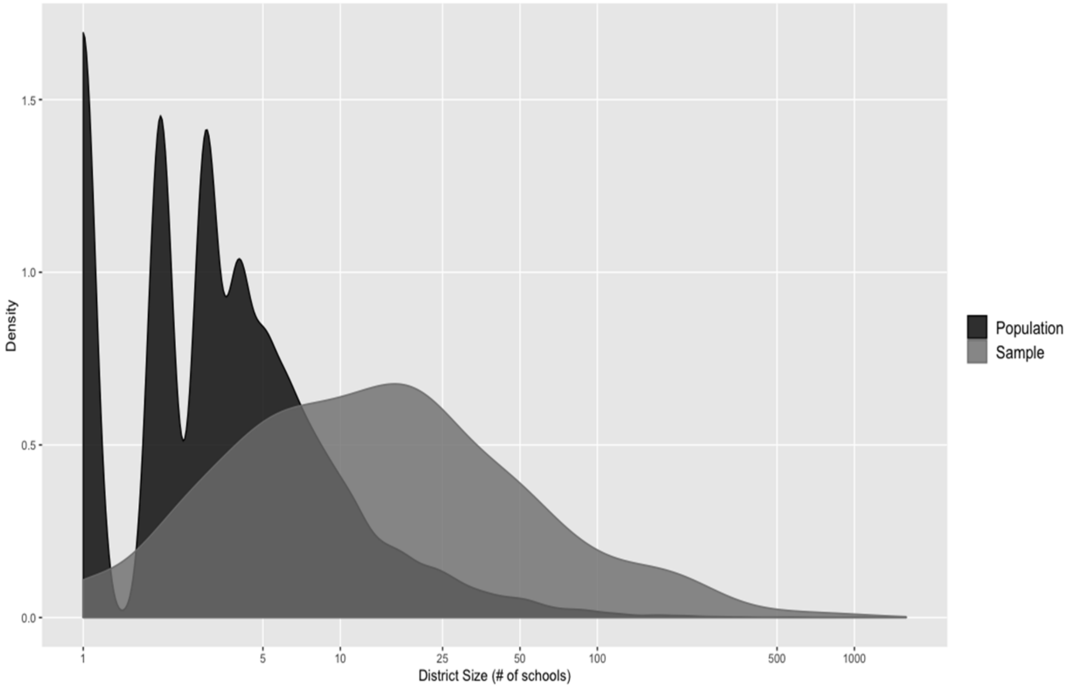
Locations of schools in 34 RCTs funded by IES between 2011-2015



** Tipton, Spybrook, Fitzgerald, Zhang, & Davidson (2020)*

We often include ‘easier’ to recruit sites

Size of school districts in 34 RCTs funded by IES between 2011-2015



Researchers prefer large school districts.

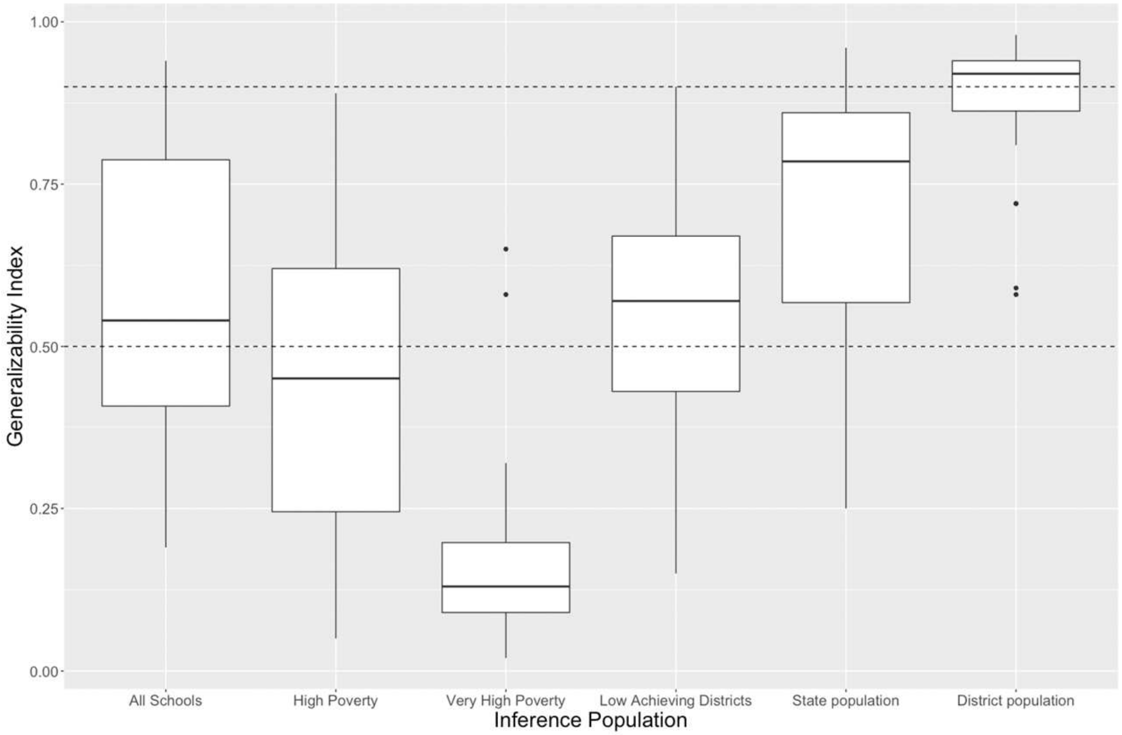
Large districts tend to bring with them more schools.

They are more often urban.

They have very different resources and students.

Studies may not generalize well

Comparisons of study samples to 6 populations in each of the 34 RCTs funded by IES between 2011-2015



1. Studies did not report clearly what their target populations were.

2. Most studies did **not** represent policy populations well.

3. They did best at representing the districts they were in.

Solutions to this problem

1. **Define** the target population and recruit to represent it.
2. **Assess** the similarity between the sample and target population and **adjust** for any differences in the impact estimate.
3. **Determine** if there is a sub-population that the sample better represents than the whole.
4. **Report**, report, report. You as evaluator know more about the target population and recruitment than anyone else – help others understand where results likely apply and where they do not.

Design a better study

Design-based approach

Studies should plan for the types of generalizations they would like to be able to make. This includes defining:

- 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.

Define the target population

The target population needs to be:

- Enumerated (a list of all sites)
- Described using clear inclusion / exclusion criteria.

Importantly, population definitions can be **broad or narrow**.

Example: *This study seeks to determine the average treatment effect of [intervention] in the population of 80,726 'regular' public elementary schools in the US in 2017-18. The population includes schools serving students in K-6; alternative schools and federal schools were excluded.*

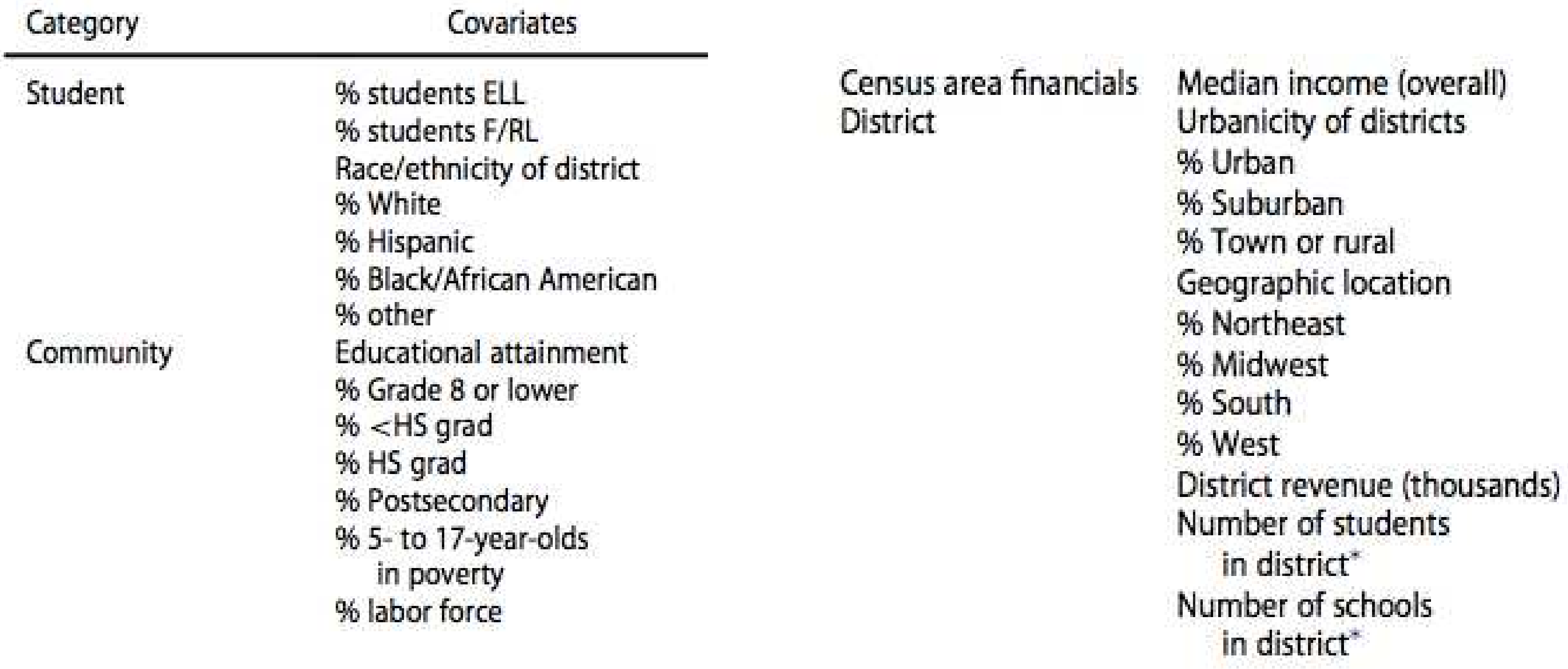
Identify potential moderators

If treatment impacts vary, which variables might moderate the effect? We need the sample and target populations to have similar distributions of these moderators.

It's impossible to know these in advance. But we can develop a sense of the **potential moderators:**

- Requires data on these for all sites in the population
- Demographics, variables related to implementation, variables related to outcomes

Example potential moderators



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	Median income (overall)
District	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*

From: Tipton et al, (2016)

Divide the population into strata

Stratification is an easy tool for decreasing variance and increasing similarity.

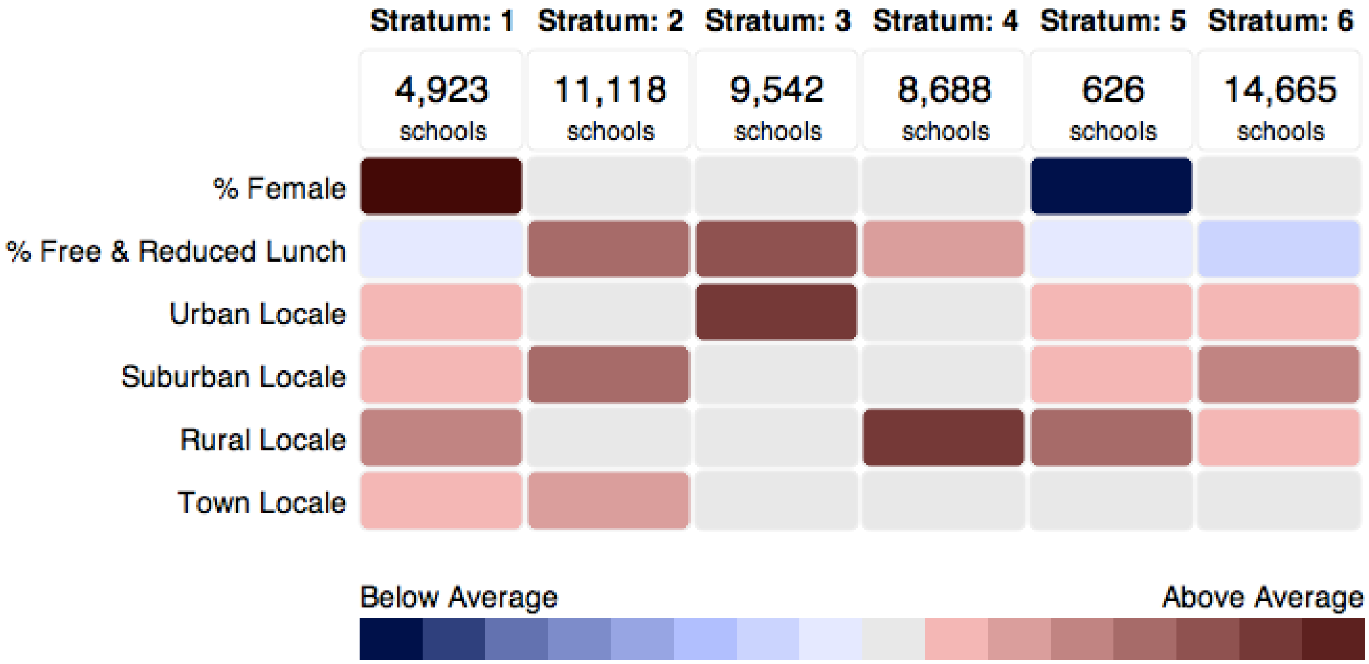
Cluster-analysis[1] is one approach:

- 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 used[2].

1. Tipton (2014); 2. Tipton et al., 2014

Example strata



From: The Generalizer.

Recruit from the strata

The total sample *n* can then be recruited from these strata using proportional allocation. *That is, if Stratum 1 contains 20% of the target population, then 20% of the sample should be recruited from the stratum.*

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

- Randomly
- Similarity to the the stratum average
- Based purely on convenience

Collect information on recruitment

This process gives recruiters goals and requires the development of **strategies**.

Incentives, resources, and goals may differ by stratum.

Information can be collected during recruitment, providing information on **refusals**, which can later be analyzed.

For example, in 2 scale-up studies, we[1] found that schools in medium sized school districts and those serving predominately low-income students were more likely to agree to be in the studies.

[1] Tipton, Fellers, Vaden-Kiernan, Borman, Caverly, & de Castilla (2016)

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	9th 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 6th 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

Assess and Adjust

Assess similarity

Regardless of your recruitment approach, ask:

> **Is the sample like the target population in terms of potential moderators?**

How can you do this?

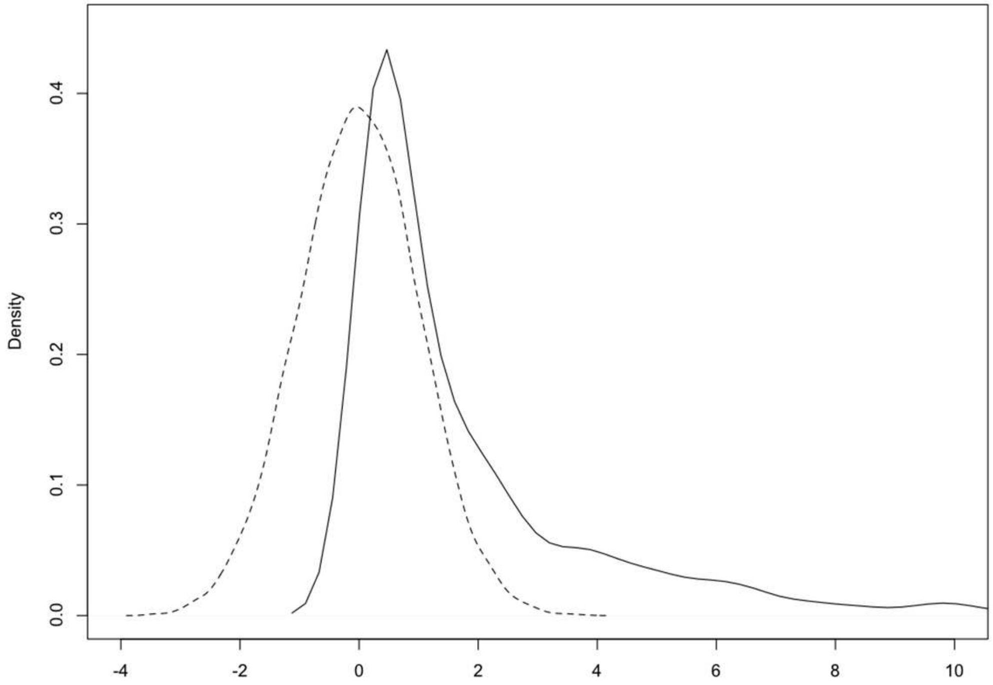
- Calculate SMDs between the sample and population on potential moderators
- Calculate a global measure of similarity

Globally compare the sample and population

For each site, estimate $s_i(x)$ the probability that the site is in the evaluation given their potential moderators,

$$\log\left(\frac{s_i(x)}{1 - s_i(x)}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

Compare the distributions of these probabilities in the sample and population.



Summarize the similarity

The **generalizability index** is a global measure of similarity. Formally it is defined as,

$$\beta = \int \sqrt{f_s(s) f_p(s)} ds$$

This is the *geometric distance* between the distributions of potential moderators.

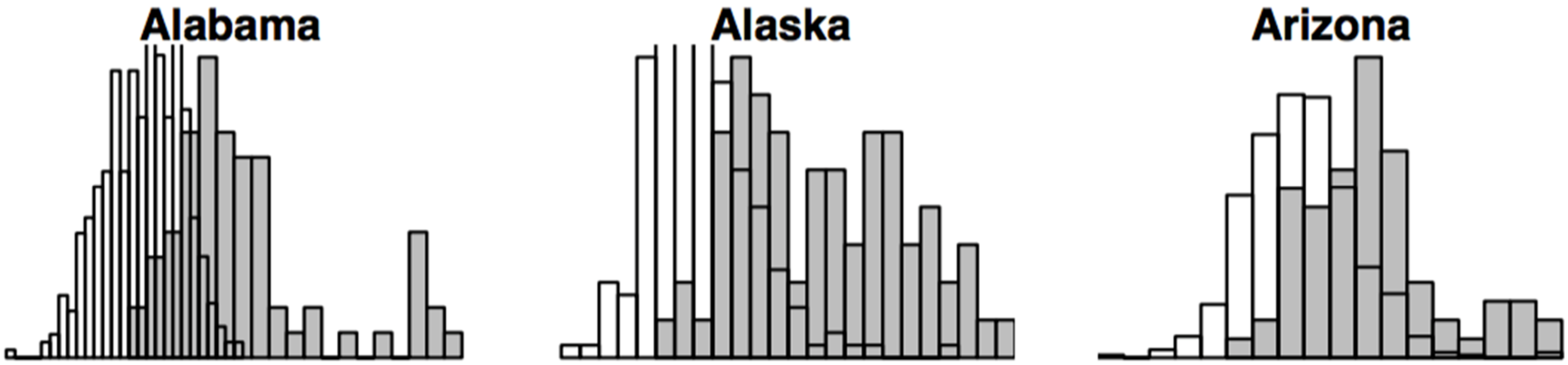
The index takes values between 0 and 1:

- 1 indicates the sample is an exact miniature of the population,
- 0 indicates the sample and population share no common features.

We can interpret it by multiplying by 100:

"the sample of schools in our study is 90% similar to the population of schools in the US."

Generalizability index examples



Not very similar

Quite similar

0.31

0.52

0.79

Begin to think about adjustment

Conclusion	Index value	Decision
The sample is as similar to the target population (on the potential moderators) as a random sample of the same size.	> 0.90 (ish)	The usual population ATE estimator is fine.
The sample differs from the target population and these differences should be adjusted for.	0.50 < index < 0.90	A population ATE estimate will require adjusting for differences.
The sample is very different from the target population.	< 0.50	A population ATE estimate is not possible without extrapolations.

The simplest reweighting

	Stratum 1	Stratum 2	Stratum 3	Stratum 4
Population	20%	30%	40%	10%
Sample	10%	50%	20%	20%
Estimated Treatment impact	**0.29**	**-0.13**	**0.49**	**0.08**
Standard error	***.18***	***.04***	***.08***	***.10***

$$\hat{\tau}_p = \sum_{i=1}^{k} w_{pi} \hat{\tau}_i = .20(.29) + .30(-.13) + .40(.49) + .10(.08)$$

$$SE\left(\hat{\tau}_p\right) = \sqrt{\sum_{i=1}^{k} w_{pi}^2 SE\left(\hat{\tau}_i\right)^2} = \sqrt{.20^2\left(.18^2\right) + .30^2\left(.04^2\right) + .40^2\left(.08^2\right) + .10^2\left(.08^2\right)}$$

More complex adjustments

If you didn’t plan for generalization from the outset, there may *not* be strata to use for adjustment.

One approach is to instead **post-stratify** based on $s_i(x)$, the *probabilities that sites were in the sample (*Tipton, 2013).

Another approach is to use **inverse-probability** of selection weights, i.e., $w_i = 1/s_i(x)$ (Stuart et al, 2011).

A few caveats

1. Adjustments affect both the population ATE estimate and its precision.
2. Adjusted estimates tend to be less precise (larger standard errors).
3. The more similar the sample is to the population, the smaller the effect of the adjustments.
4. You can only adjust for *observed potential moderators:*
   - There may be other, unobserved moderators
   - Some of the potential moderators may not be actual moderators

Determine the population actually studied

Adjustment is not always possible

The effectiveness of adjustment methods is limited in practice because of **under-coverage,** which occurs when some population sites have 0 probability of being in the sample.

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.*

This is ***the*** generalization problem in the real world**.**

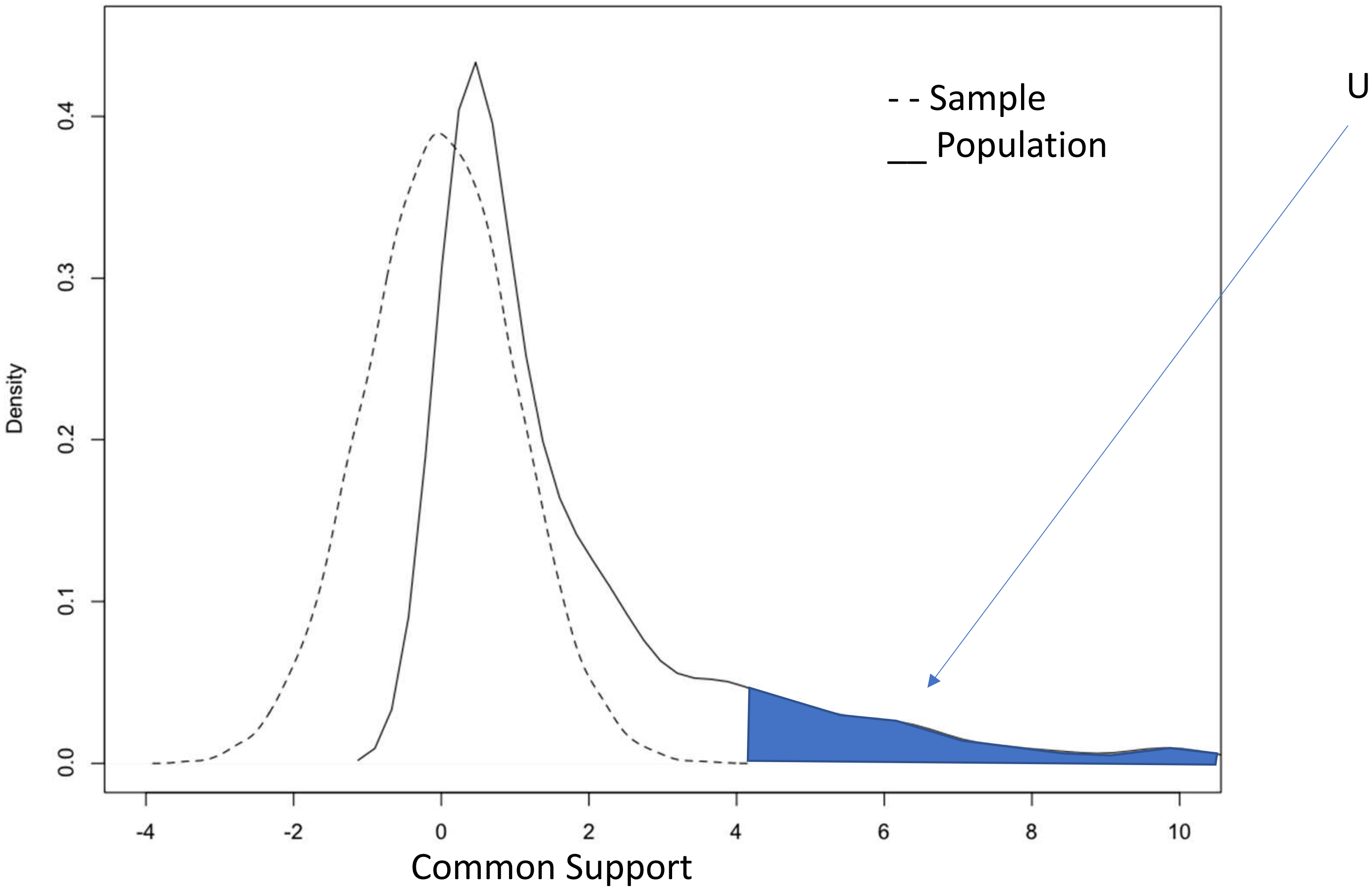
Example of the problem

	Stratum 1	Stratum 2	Stratum 3	Stratum 4
Population proportion	20%	30%	40%	10%
Schools to recruit (n=40)	40x20% = 8	12	16	4
Actual Sample #1	7	14	13	6
Actual Sample #2	10	0	20	10
Actual Sample #3	20	2	8	10

Compared to an unadjusted estimator (that is biased!):

- Sample #1: the reweighted estimator has a standard error about **2%** larger.
- Sample #2: this is an example with a coverage error. **We cannot reweight.**
- Sample #3: the reweighted estimator has a standard error about **64%** larger!

Selection probability version



Under-coverage

A strategy to use

It’s *not* helpful to just restrict generalizations to an overlap region:

- “You can generalize to 70% of the population in the common support region” is not helpful.

Better: use inclusion criteria

- Try different simple exclusion criteria
- For each criterion, check how similar the sample is to the (new) population
- Repeat

Open Court Reading Experiment

In 2011-12, a scale-up evaluation of OCR began recruitment.

The goal was to generalize to the population of schools *that purchase the OCR curriculum* (Tipton et al, 2014).

But recruitment was very difficult and the sample selected clearly differed from this population (Tipton et al, 2017).

Question: Was there some sub-population that the study could generalize well to?

Possible inclusion/exclusion rules

Sub-population restricted to include only units with covariate values LESS than the MAX observed in the Experiment for:

1. % FRL
2. % Hispanic
3. % Black
4. % White
5. % Other
6. % ELL
7. Total district revenue
8. % less than 9th grade education
9. % less than HS education
10. % less than HS graduate
11. % post-secondary education
12. % poverty
13. % employed in labor market
14. District median income
15. Total students in district
16. Total schools in district
17. % Urban
18. % Rural or town
19. % Suburban
20. Northeast
21. Midwest
22. South
23. West

Not important for bias reduction, but perhaps useful for inclusion criteria

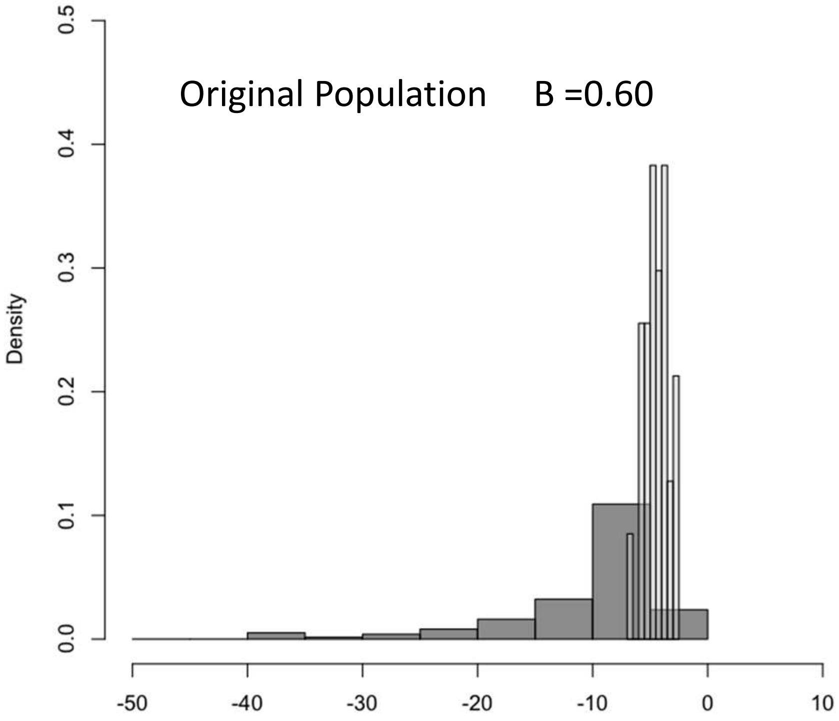
Table of comparisons

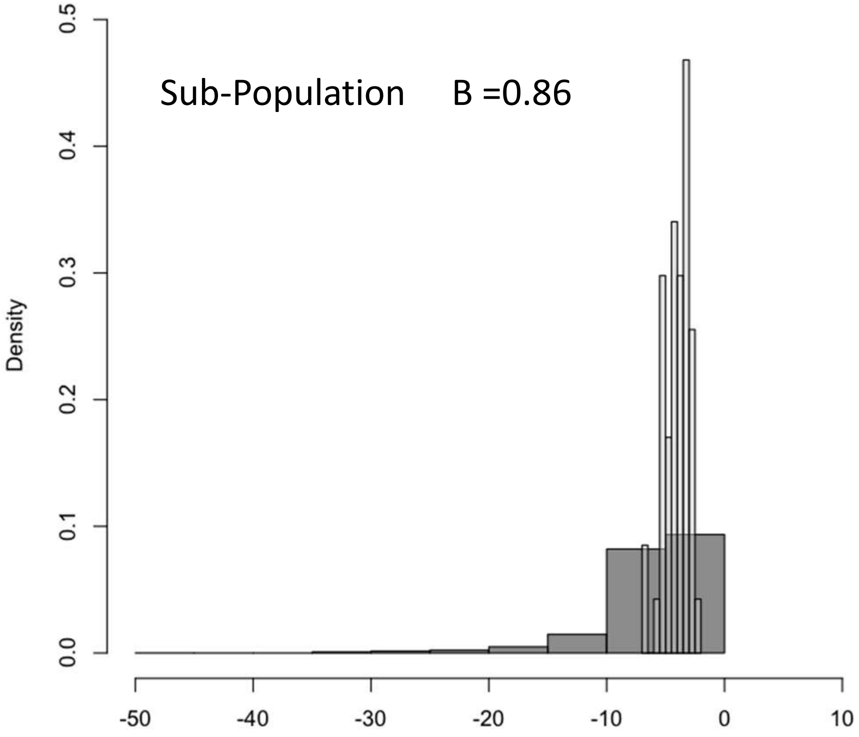
Criterion	B-index	% of Original Population
15	0.77	48%
7	0.76	45%
16	0.75	56%
11	0.69	66%
13	0.65	83%
14	0.64	73%
2	0.63	82%
8	0.61	86%
6	0.61	77%
…	…	…
10	0.55	96%

Criterion	B-index	% of Original Population
14	0.82	38%
11	0.81	41%
13	0.77	47%
2	0.77	46%
15	0.77	48%
7	0.76	45%
8	0.75	49%
20	0.75	48%
16	0.75	55%
…	…	…
10	0.71	52%

Criterion	B-index	% of Original Population
2	0.87	29%
15	0.84	33%
8	0.84	32%
7	0.83	31%
12	0.83	35%
11	0.83	35%
13	0.82	36%
1	0.82	38%
14	0.82	38%
…	…	…
6	0.77	30%

Comparison (before vs after)





Example final results

Covariate	|SMD| P vs S	|SMD| P_0 vs S
Minimum	0.02	0.01
Average	0.33	0.16
Maximum	0.80	0.48
> 0.28	9	2
Generalizability index	0.61	0.87
Population %	100%	29%

Final inference population:

“Of the schools that typically purchase OCR, the results of this study generalize to those in smaller districts serving communities with lower incomes, but not serving large Hispanic student populations.”

What to do and report

Be clear about generalization

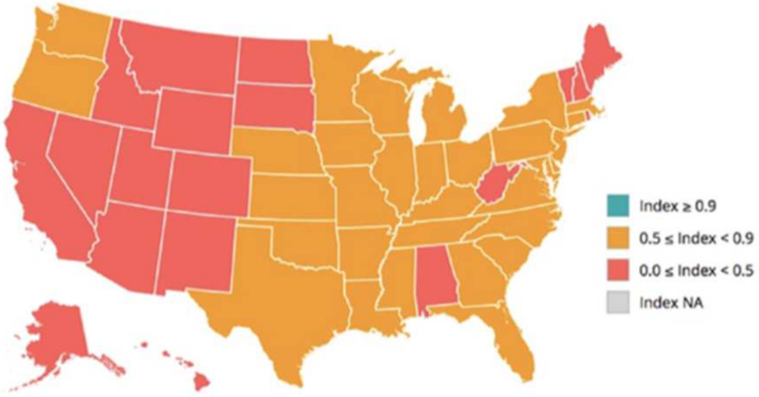
1. When possible, have hard conversations about generalization at the beginning of the study.
   - What is the goal of the study? How will the results be used? Who needs these results? Who makes decisions about implementation of the intervention?
   - Look for population data. Be creative. Combine sources.
   - Stratify. Try to recruit a representative sample.
   - *Include this in your study reports.*

2. Collect data on recruitment.
   - Who did you contact? Who agreed? Why didn’t sites want to be in the study? Did those that agreed differ from those that didn’t?
   - What incentives did you try? Did they work?
   - *Include this in your study reports.*

Be clear about generalization

3. Compare your sample to relevant target population(s) and adjust for differences.
   - Summarize similarity globally and on potential moderators.
   - When necessary, adjust for differences.
   - *Include this in your study reports.*

4. Help readers understand where result apply and where they don't.
   - Provide clear inclusion/exclusion criteria.
   - If you can't generalize to the ideal target population, figure out where you can generalize.
   - *Include this in your study reports.*



Resources to help

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516-524.

Coburn, K., Ackerman, B., Tipton, E., & Chao, B. **generalize**: An R package for planning, assessing, and estimating population treatment effects.

Tipton, E. & Miller, K. The Generalizer: A webtool for planning and assessing generalization. https://www.thegeneralizer.org

Thank you!

Elizabeth Tipton
tipton@northwestern.edu
www.bethtipton.com
https://stepp.center
@stats_tipton