Designing & Planning for Generalizable Impact Studies

- Elizabeth Tipton
- Northwestern University

- *Presented at IES Cluster Randomized Trial Workshop*
- *Northwestern University*

Generalizability

Efficacy studies *used to* focus entirely on:

- Internal validity
- Statistical conclusion validity

But results of these studies are intended to be useful *in practice, in schools*.

If treatment effects vary, it matters a lot who is in such a study.

The crux of the problem

The *average* treatment effect Δ is simply the **weighted average of subgroup** averages Δ_j.

$$\Delta = w_1\Delta_1 + w_2\Delta_2 + \cdots + w_k\Delta_k$$

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:

$$\boldsymbol{\Delta_s \neq \Delta_p}$$

IES *Required* (for Initial Efficacy)

IES now recognizes this and includes requirements for generalizability in their RFPs.

You must describe the:

- **Characteristics of your sample**
- Research design and methods
- Power analysis
- Data analysis plan

IES *Recommended* (for Significance)

Describe the **population intended to benefit** from this intervention and how your sample does or does not represent this larger population, including

- The **learners** who should benefit, either directly or indirectly, from this intervention
- The **education personnel** who will implement the intervention and how they will implement it
- The **heterogeneity of the sample** you propose compared to the target population.

IES *Recommended* (for Research Plan)

Related to variation:

Identify factors that might lead to the effect of the intervention **varying** across the learners and settings in your target population and the variables available to measure these factors.

Related to target population:

Define and enumerate **who would benefit** from the intervention. IES does not expect individual projects to be generalizable to the U.S. population as a whole; instead, your target population may represent a narrow segment of the larger U.S. population.

Identify the inclusion/exclusion criteria you will use during sample recruitment. Discuss how these may narrow the target population studied and influence the generalizability of the results to the target population.

IES *Recommended* (for Research Plan)

Related to your sample:

Describe the **setting** in which the study will take place, including the size and characteristics of the setting and/or the surrounding community, and how this will help better identify the learners or settings for which the intervention is most likely to work.

Explain how your work **with this sample** will contribute to a larger body of knowledge on promising interventions and whether your work will contribute to expanding what is known about learners from diverse backgrounds and experiences, including learners from underrepresented communities or populations.

Describe **the setting(s)** in which the research will take place (provide letters of agreement in Appendix E) and discuss how they will allow you to draw conclusions about the education settings your research is intended to inform. Describe strategies to reduce attrition, if applicable.

IES *Recommended* (for Research Plan)

Related to recruitment:

Detail the **procedure that will be used to recruit** a specific sample that represents the target population in need of the proposed intervention.

Describe the **sample recruitment procedure** that will be used to **ensure similarity** between the sample and target population.

Describe **strategies to increase the likelihood that participants** (for example, schools, educators, and/or learners) will join the study and remain in the study over the course of the evaluation.

IES *Recommended* (for Research Plan)

Related to analysis:

Describe how you will **measure the generalizability of your findings** by contrasting your sample's characteristics with the characteristics of the target population.

Describe your **plans for adjusting** for any mismatch between your sample and the population.

Whoa!

This seems like a lot of requirements.

It is.

The good news is that there is a guide that can help.

I'll give an overview today.

Overview of the guide

IES Institute of Education Sciences

Toolkit

Enhancing the Generalizability of Impact Studies in Education

NCEE 2022-003
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation and Regional Assistance

U.S. Department of Education
Miguel Cardona
Secretary

Institute of Education Sciences
Mark Schneider
Director

National Center for Education Evaluation and Regional Assistance
Matthew Soldner
Commissioner

Thomas Wei
Amy Johnson
Project Officers

The Institute of Education Sciences (IES) is the independent, non-partisan statistics, research, and evaluation arm of the U.S. Department of Education. The IES mission is to provide scientific evidence on which to ground education practice and policy and to share this information in formats that are useful and accessible to educators, parents, policymakers, researchers, and the public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other IES product or report, we would like to hear from you. Please direct your comments to ncee.feedback@ed.gov.

This report was prepared for the Institute of Education Sciences (IES) under Contract 91990020F0052 by Mathematica. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

February 2022

The guide

The guide was:

- authored by Elizabeth Tipton (Northwestern University) and Rob Olsen (George Washington University)
- supported by the Institute of Education Sciences

It provides:

- 7 recommendations on how to enhance the generalizability of impact studies in education, including randomized controlled trials (RCTs) and quasi-experimental designs (QEDs)
- Examples to illustrate these recommendations

3 Recommendations (planning)

1. Select the target population	*Identify the collection of students and schools about which the impact study aims to learn*
2. Develop a population frame	*Assemble a dataset that includes a list of the schools from the target population*
3. Design a sampling plan	*Develop a plan for selecting a representative sample of the schools from the target population*

+ 4 more (Implementation)

4. Implement the sampling plan	*Select schools according to the sampling plan and recruit them to participate*
5. Assess the similarity between the sample and the target population	*Use the population frame data to compare schools in the sample to schools in the target population*
6. Adjust for differences between the sample and the target population	*Use weighting or regression methods to adjust for differences in observed moderators*
7. Report generalizability appropriately	*Report sufficient information to help readers understand the study's generalizability*

1. Select the target population

What is a “target population”?

Impact evaluations usually aim to estimate (among other things) the average impact of an intervention

- But average over whom, or what?

A well-defined objective for an impact study is to estimate the intervention’s average impact **for a target population**

- The study’s target population *defines the collection of students and/or schools over which the study aims to estimate an average impact*

Use moderators to identify populations

The average impact of the intervention will vary across populations defined by the values of "impact moderators"—factors that influence the direction and magnitude of the intervention's impact

- If urbanicity and the racial composition of enrolled students influence the intervention's impact in schools, then the average impact will be different in a population of urban, majority black schools than in a population of suburban, majority white schools

So, identifying potential impact moderators is a key first step in defining the target population for your study

How to identify impact moderators?

To identify potential impact moderators

- Review the intervention's logic model
- Consider empirical evidence from the literature
- Consult experts like the intervention developer

Note that the evidence available to select moderators may be limited (but don't let that stop you from making your best guesses)

Use substantive considerations

Identify the target population for which evidence on the intervention's impact would be most useful:

- *To inform a specific policy decision by a particular funder*, identify the full set of students and schools that would be directly impacted by the decision (e.g., all schools nationwide that received funding from the federal program being evaluated)
- *To fill a gap in the evidence base*, identify the populations of students or schools for which prior evidence is lacking (e.g., rural schools if all prior studies have focused on urban schools)

Narrow the population, if necessary

Restrict the target population based on **pragmatic** considerations

For example, exclude schools from districts or states that:

- Are located far from the study team, making their inclusion in the study too costly
- Cannot provide the necessary data

Tradeoffs to consider

Defining the target population broadly can elevate a study's aspirations or applicability to a wide range of decision-makers

But if the study can't reasonably recruit a sample that represents that broad population, those aspirations may be unachievable

Researchers face a tradeoff between a broader and potentially more important population that they can't fully represent and a narrower population that they can, but that is of less policy interest

2. Develop a population frame

What is a population frame?

A **population frame** is a dataset that includes a list of units from the target population

- In studies that plan to recruit schools to participate, the population frame should include a list of the schools for recruitment
- Ideally, this list will include schools that are part of the target population ("eligible schools") and exclude schools that are not ("ineligible schools")

What is included?

Population frame data should include the following information about each school in the population:

- Identifying information (e.g., name, location, NCES id)
- Contact information (e.g., address, phone number)
- Potential impact moderators—those that can be assembled at reasonable cost
- Other variables needed to distinguish eligible schools from ineligible schools

What schools are in the frame?

A population frame should include:

- Include schools that are in the population (as much as possible)
- Exclude schools that aren't in the population (as much as possible)

But a population frame doesn't have to be perfect to be useful

- Studies need some way of identifying eligible schools
- Ineligible sites can be screened out during the recruitment process

Some population frames for K-12

Data source	Level	Information
Common Core of Data (CCD)	School district, school	Numbers and types of districts and schools, student enrollment, federal program participation (e.g., Free and Reduced Lunch), teacher counts, district expenditures, and other information
EDFacts	State, school district, school	General information and state-reported performance data for federal education programs (e.g., Title I, IDEA)
American Community Survey—School District Demographic System (ACS-SDDS)	School district	Indicators of social, economic, and housing conditions for school-age children and their parents
Private School Universe Survey (PSUS)	School	Religious orientation, level, total enrollment, length of school year and school day, single-sex or coeducational, program emphasis, and other information
Stanford Education Data Archive (SEDA)	School district, school	Measures of academic achievement and achievement gaps for public schools

3. Design a sampling plan

An overview

Obtaining a perfectly representative sample *is infeasible* for impact studies where participation is voluntary

To obtain a sample that is as representative *as possible*:

- Divide schools into **strata** based on potential impact moderators
- Set **recruitment targets** within each stratum to ensure proportional representation of schools across strata
- **Select** schools—and replacement schools—for recruitment to obtain a sample that is as representative of the target population as possible

How should the sample be stratified?

Categorical variables lend themselves to stratification:

	Elementary School	Middle School	High School
Urban	ES & urban	MS & urban	HS & urban
Other	ES & other	MS & other	HS & other

Continuous variables can be divided into categories based on some threshold (e.g., at the median to ensure adequate numbers of schools in each stratum)

What if there are a lot of moderators?

Stratification is challenging when the number of moderators is large but the number of sites to be selected is small

- For example, all possible combinations of 5 binary variables would create 32 strata, and many studies do not even include 32 schools

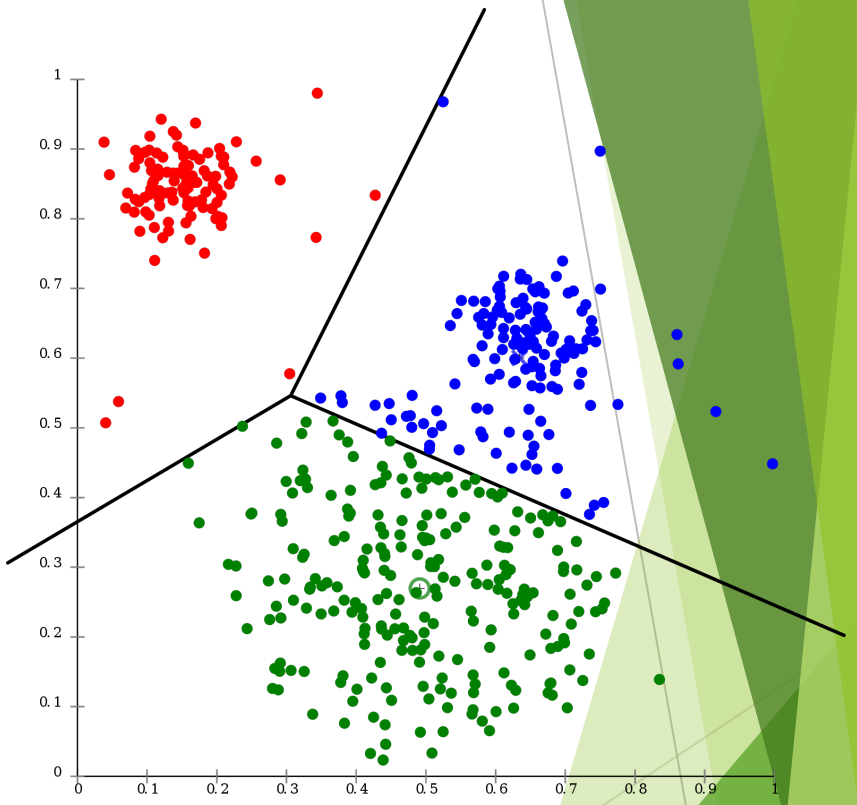
To address this challenge, impact studies can use **cluster analysis**

Cluster analysis to create strata

Cluster analysis is a statistical technique for grouping similar observations.

One approach is to use "**k-means clustering**" to group similar schools based on observed moderators in the population frame.

Researchers can use cluster analysis to create strata for sampling.



How many strata?

How many clusters / strata should we define?

- More strata are better for producing generalizable samples
- But too many strata are unwieldy to implement given recruitment

The guide recommends testing 4, 5, and 6 clusters, and choosing based on the percent of variance explained across all moderators

- For example, choose 5 clusters if the percent explained is much higher than for 4 clusters and only slightly lower than for 6 clusters

How should schools be selected?

The use of strata alone can help yield samples that are as representative as possible, accounting for school participation decisions

But how the sample is selected **within** the strata also matters. Two options covered in the guide are:

- **Probability** sampling—that is, setting the probability of selection for each school and selecting schools with those probabilities
- **Balanced** sampling—that is, selecting a sample of schools nonrandomly to match the population on observed moderators

Probability sampling?

Probability or random sampling is used often in surveys and occasionally in impact studies (e.g., the Head Start Impact Study)

The guide covers one approach to random sampling. Within strata:

- Schools are **sorted by a random number**
- Schools are selected in sort order for recruitment—both initially and as replacements when initially selected schools decline to participate

Balanced sampling?

"Balanced sampling" is a general term to describe strategies for selecting samples that are like the target population on observed characteristics (e.g., potential moderators)

The guide covers one such strategy. Within strata:

- Schools are **sorted by the multivariate distance between each site and the stratum mean, from smallest to largest**
- Schools are selected in sort order for recruitment—both initially and as replacements when initially selected schools decline to participate

4. Implement the sampling plan

Threats to generalizability during recruitment

School recruitment will not always yield a sample that is as representative of the target population as we would like:

- **Refusal rates** may be higher in some strata than others
- Within strata, refusals may be **related** to both observed **moderators** and unobserved moderators at the school or district levels
- If so, the sample may **not represent** the target population well

The recruitment effort should be designed to anticipate and mitigate these challenges

Steps in recruiting schools

1. **Budget** time and resources for recruitment as early as possible
2. Build and manage a recruitment **team**
3. Screen out schools that are **ineligible** for the study
4. **Collect** and report data on ‘volunteers’ and ‘decliners’

What strategies can mitigate threats?

Mitigation strategies include:

- **Training recruiters** to intensively recruit schools higher on the list (e.g., closer to the stratum means) before recruiting schools lower on the list
- **Financial incentives** and reallocation of recruitment effort toward schools in "hard-to-recruit" strata
- **Tracking** of school participation decisions and reasons for declining
- **Collection and reporting of information** on recruited schools that agreed to participate and recruited schools that declined

5. Assess the similarity between the sample and the target population

Generalizability bias: What is it?

Generalizability bias results when the sample average treatment (SATE) differs from the population average treatment effect (PATE), where:

- SATE = the *true* ATE in the sample
- PATE = the *true* ATE in the population

It arises when both (a) treatment effects vary in relation to moderators and (b) these moderators are not similar in distribution in the sample and population.

Generalizability bias may arise from differences in observed moderators (e.g., those in the population frame data) and differences in unobserved moderators

How to assess generalizability

Compare recruited schools to the target population:

- Calculate the **standardized mean differences** in potential moderators between the sample and the population using data from the population frame
- Calculate a global measure of representativeness – the ***generalizability index*** – across all the moderators

Example (SMDs)

Table 7. Comparison of study sample and population frame on potential moderators

Potential moderator	Population frame mean	Sample mean	Population frame standard deviation	Standardized mean difference
Students eligible for free or reduced-price lunch (%)	78.2	86.3	23.9	0.34
Female students (%)	48.4	48.5	2.0	0.05
Black students (%)	50.2	56.0	20.2	0.29
Hispanic students (%)	24.8	21.9	16.4	-0.18
Native American students (%)	0.6	0.2	3.2	-0.13
English language learners (%)	8.3	7.0	4.1	-0.32
Home language other than English (%)	17.7	12.0	11.5	-0.50
Urban school (%)	38.1	42.2	37.5	0.11
Suburban school (%)	35.6	41.0	35.5	0.15
Town school (%)	10.0	8.6	22.7	-0.06
Rural school (%)	16.3	10.4	28.0	-0.21
Number of students per school	562.1	633.7	163.4	0.44
Number of schools per district	120.4	130.6	90.4	0.11

Note: The standardized mean difference is the difference in means between the population and the sample divided by the population standard deviation.

Calculating the index

- Let $Z = 1$ if a school is in the sample (and $Z = 0$ if it is not)
- Let $X_1, X_2, \ldots, X_p$ be moderators of the treatment effect
- Estimate

$$l = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_p X_p$$

- Examine the empirical densities of these two distributions:

  $f_p(l)$: Distribution of logits for all population schools

  $f_s(l)$: Distribution of logits for all sample schools

Calculate the index

- Calculate the index (see Tipton, 2014):

$$B = \int_{-\infty}^{\infty} \sqrt{f_p(l) f_s(l)},$$

- This looks complex, but you don't need to do this manually! It is built into software ☺

Interpreting the index

B index	Generalizability	Sample …
[0.9, 1]	Very high	No adjustment is required: Sample is as similar to the population on these moderators as a random sample
[0.8, 0.9]	High	Can be reweighted (with little variance penalty) to achieve a generalizable estimate of the PATE
[0.5, 0.8]	Medium	Can be reweighted (with a large variance penalty) to achieve a generalizable estimate of the PATE
[0, 0.5]	Low	Is too different from the population to produce a generalizable estimate of the PATE

What about unobserved moderators?

- To explore threats from unobserved moderators:
  - Calculate and assess the **opt out rate** among schools selected and recruited to participate
  - Summarize and report on moderators that are observed for the sample but not observed for the population
  - If possible, test for the presence of unobserved moderators by estimating the unexplained variation in impacts across sites
  - Treat the generalizability index for observed moderators as an **upper bound** on the similarity between the sample and population frame on unobserved moderators

6. Adjust for differences between the sample and the target population

What should you do if index < .9?

B index	Generalizability	Sample …
[0.9, 1]	Very high	No adjustment is required.
[0.8, 0.9]	High	Apply post-stratification weights
[0.5, 0.8]	Medium	Apply post-stratification weight (or use propensity score weights)
[0, 0.5]	Low	Redefine the target population using propensity score weights, then reweight

Post-stratification

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Total/Sum
N_j	200	400	300	100	1000
n_j	30	30	20	20	100
N_j/n_j	6.67	13.33	15	5	40
$\boldsymbol{W_{ij} = \frac{N_j/n_j}{N/n}}$	0.67	1.33	1.5	0.5	NA
$\boldsymbol{W_j = n_j W_{ij}}$	20	40	30	10	100

Propensity score approaches

- Begin with the distributions of logits
- Divide into 5 subclasses / strata
  - Based on quintiles of the $f_p(l)$ distribution

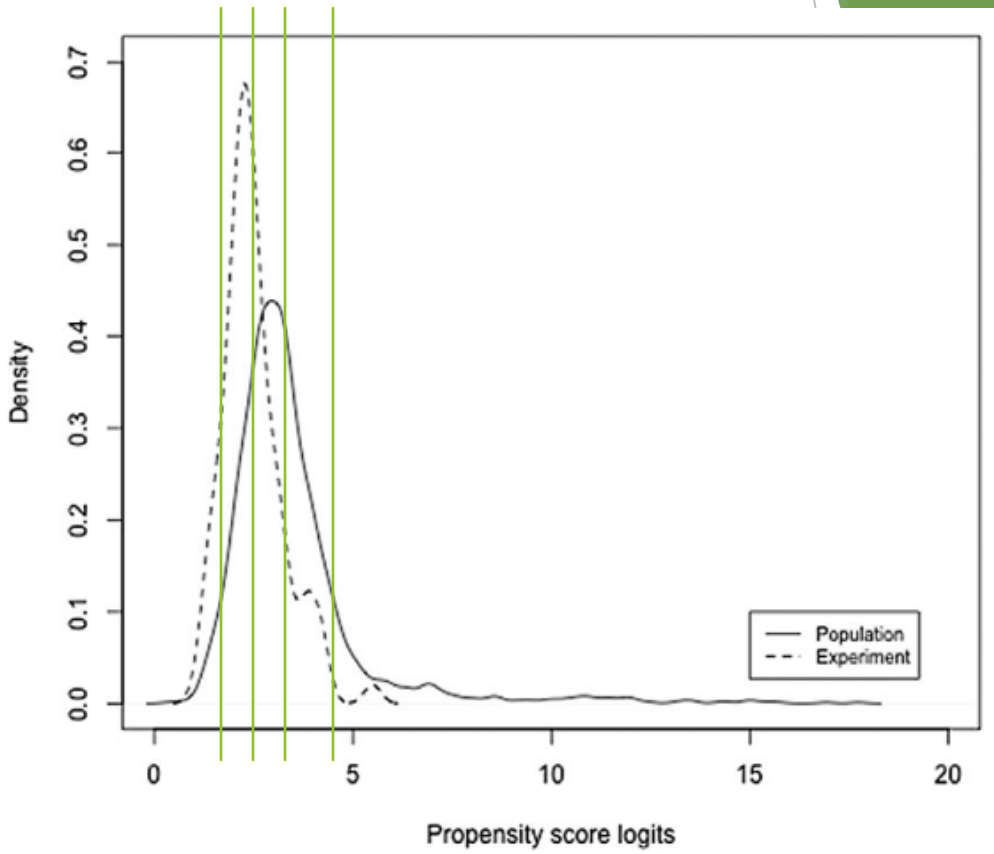


FIGURE 1. *Distribution of propensity score logits for population and experimental schools, when the predicated category is nonmembership in the experiment.*

Propensity score subclassification

	Subclass 1	Subclass 2	Subclass 3	Subclass 4	Subclass 5	Total/ Sum
p	(0, 0.01]	(0.01, 0.08)	(0.08, 0.12]	(0.12, 0.19]	(0.19, 0.20)	NA
N_j	200	200	200	200	200	1000
n_j	3	8	20	30	39	100
N_j/n_j	66.67	25	10	6.67	5.13	40
$\boldsymbol{W_{ij} = \frac{N_j/n_j}{N/n}}$	**6.67**	**2.5**	**1**	**0.67**	**0.51**	**NA**
$\boldsymbol{W_j = n_j W_{ij}}$	**20**	**20**	**20**	**20**	**20**	**100**

Reweighting is great, right?

- Keep in mind that reweighing:
  - improves similarity / reduces bias from
  - observed moderators.
- It comes at a cost:
  - Reweighting tends to increase standard errors
  - The lower the generalizability index, the larger the increase in standard errors
- It can't fix everything:
  - What about unobserved moderators?

Other approaches

- Inverse probability of selection weighting:
  - Weight each sample unit $1/p_i$ (where p_i estimated from logit model)
  - This is like subclassification with a lot of strata
- Regression approaches
  - Could treat this as a prediction problem
  - Beneficial for additional covariates that you don't have full knowledge of in the population
  - Useful in larger samples
  - More in the guide

How do you know this works?

- The theory suggests this will increase similarity. But it doesn't always do so perfectly. You need to assess this.
- To do so, compare unweighted and weighted SMDs for the moderators. You'd like the weighted ones to be < .25 SDs
- You might try a few different methods to see which one improves the similarity:
  - E.g., post-stratification (using your strata)
  - E.g., propensity score subclassification (with different #s of strata)
  - E.g., inverse probability weighting

Example

Table 9. Comparison of sample to target population after reweighting adjustment

Potential moderator	Target population mean	Reweighted sample mean	Target population standard deviation	Adjusted standardized mean difference	Standardized mean difference
Students eligible for free or reduced-price lunch (%)	78.2	80.6	23.9	0.10	0.34
Female students (%)	48.4	48.5	2.0	0.05	0.05
Black students (%)	50.2	52.0	20.2	0.09	0.29
Hispanic students (%)	24.8	25.2	16.4	0.02	-0.18
Native American students (%)	0.6	0.5	3.2	-0.03	-0.13
English language learners (%)	8.3	8.0	4.1	-0.07	-0.32
Home language other than English (%)	17.7	17.4	11.5	-0.03	-0.50
Urban school (%)	38.1	39.9	37.5	0.05	0.11
Suburban school (%)	35.6	35.7	35.5	0.00	0.15
Town school (%)	10.0	9.2	22.7	-0.04	-0.06
Rural school (%)	16.3	15.5	28	-0.03	-0.21
Number of students per school	562.1	579.8	163.4	0.11	0.44
Number of schools per district	120.4	127.9	90.4	0.08	0.11

Note: The adjusted standardized mean difference is the difference in means between the population and the reweighted sample divided by the population standard deviation.

Undercoverage

- When is there trouble?
  - If the generalizability index < .5
  - If you were unable to recruit any schools in one of your strata
- What to do?
  - You'll need to explore your data and see if there is a way to redefine your target population so that all schools have a >0 probability of being in the study
  - Keep in mind here that you need this target population to be interpretable to non-experts!

7. Tools to help

www.thegeneralizer.org

- Free webtool
- Helps with:
  - Defining a target population
  - Building a population frame
  - Developing a sampling plan
  - Assessing generalizability
- Data available for:
  - K-12 (CCD)
  - Higher-Ed (IPEDS)

GeneralizeR package

- Free R package
  - https://nustat.github.io/generalizeR/
- Works with any population data (that you have)
- Helps with:
  - All that The Generalizer does
  - + More assessment tools
  - + Adjustments (reweighting approaches)

Conclusions

Take home points

1. **Generalizations will be made.**
   - The question is not “do we want to generalize?” but instead “do we want to **lead** the generalizations?”
2. **Plan for generalization.**
   - Every study has an inference population: some are broad, while others are narrow.
   - Even when your best efforts fail, you will be in a better situation for post-hoc statistical adjustments.
3. **Help others understand where results generalize.**
   - Be sure to report your **assumptions**: Which covariates did you compare on. Why these?