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

While permission to reprint this publication is not necessary, it should be cited as:

Tipton, E., & Olsen, R. B. (2022). Enhancing the Generalizability of Impact Studies in Education. (NCEE 2022-003). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee.

This report is available on the Institute of Education Sciences website at http://ies.ed.gov/ncee.

IES Institute of Education Sciences

Enhancing the Generalizability of Impact Studies in Education

Elizabeth Tipton, Northwestern University
Robert B. Olsen, Independent Consultant

NCEE 2022-003
U.S. DEPARTMENT OF EDUCATION

Contents

Enhancing the Generalizability of Impact Studies in Education

This guide will help researchers design and implement impact studies in education so that the findings are more generalizable to the study's target population. Guidance is provided on key steps that researchers can take, including defining the target population, selecting a sample of schools–and replacement schools, when needed–managing school recruitment, assessing, and adjusting for differences between the sample and target population, and reporting information on the generalizability of the study findings.

Rigorous impact evaluations in education are designed to produce evidence on the causal effects of educational interventions. These studies are motivated by the need to inform decisions by educators and education policymakers. To ensure that impact evaluations in education produce internally valid estimates of the impact of the intervention, evaluators typically turn to experimental designs–Randomized Controlled Trials (RCTs)–or quasi-experimental designs. These evaluation designs help to ensure that impact findings reflect the impact of the intervention–and cannot be attributed to other factors.

However, each rigorous impact evaluation is conducted in a specific sample, and the generalizability of the study findings to the populations of interest to policymakers is typically uncertain. While some steps for improving the generalizability of impact evaluations have been identified (e.g., Tipton & Olsen, 2018), researchers would benefit from more comprehensive and detailed guidance on how to conduct impact evaluations to produce evidence that is more generalizable.

Why can't we just assume the impacts estimated in rigorous impact evaluations apply beyond the samples in which they were conducted? Because impacts may and often do vary. For some interventions, the impacts vary across students and schools–sometimes substantially (Weiss et al., 2017). Furthermore, numerous impact studies in education have found differences in impacts between different subgroups of students, and schools vary in their student composition. Therefore, when we find that an intervention works in particular schools that participated in a study, we cannot assume that it will work in other schools that were not part of the study. In fact, there is growing evidence that those assumptions are tenuous at best (for example, Bell et al., 2016; Orr et al., 2019).

Unfortunately, standard evaluation practice does not go very far in addressing this problem. Evaluators rarely define the *primary* target population of students and schools about which they are aiming to learn, however narrow or broad that population may be. Even fewer select sites or analyze the data with the explicit goal of estimating impacts for that target population. More typically, researchers select a sample based primarily on pragmatic considerations. Impact evaluations conducted under Institute of Education Sciences (IES) grants often recruit schools based on proximity to and relationships with researchers (Tipton et al., 2021). Impact evaluations conducted under IES contracts usually aim for some geographic dispersion but obtaining a representative sample from a target population is rarely a priority, and the resulting samples rarely resemble the broader population (Stuart et al., 2017). Although these studies sometimes compare the sample to the population of interest on basic characteristics, they rarely attempt to make statistical corrections for any differences.

Standard evaluation practice is affected by the fact that schools are not required to participate in rigorous impact studies, so obtaining a representative sample is challenging. In most impact studies, schools can opt out, and those that opt out may differ in a variety of ways from those that agree to participate. Furthermore, strategies used to obtain representative samples for surveys, like random sampling, may appear to be infeasible.

However, conducting impact studies with unrepresentative samples of schools and students raises concerns about the generalizability of the findings. These concerns may arise if certain types of students or schools are overrepresented in the sample, or if implementation fidelity is higher (or lower) in the sample than in the target population. From a statistical perspective, the resulting impact estimates may be biased in the sense that they systematically overstate or understate the true impacts in the target population. From a policy perspective, the impact estimates may systematically overstate or understate the potential benefits of adopting the intervention more broadly in the population of policy interest. In addition, the impact estimates may fail to capture the statistical uncertainty associated with generalizing beyond the sample to the population.

Guiding principles for generalization

To address these challenges, the IES Standards for Excellence in Education Research (SEER) provide guiding principles for education evaluations, including principles for facilitating the generalization of study findings. The SEER principles on generalization refer to the use of intentional sampling methods or other methods to “permit ready generalization of [study] findings to populations of interest” (see ies.ed.gov/seer/generalization.asp). Additionally, these SEER principles refer to statistical adjustments to support generalizing study findings to populations of interest. But these principles raise important questions for the practice of evaluation. Which sampling methods should evaluations use for generalizing to these populations? How can the generalizability of the study sample be assessed? And how should statistical adjustments be made in the analysis to support these generalizations? The purpose of this guide is to provide advice regarding the operationalization of these SEER principles.

Overview of the guide

This guide is designed to help evaluators enhance the generalizability of impact studies in education. It builds on literature on how to conduct impact studies for improved generalizability (Tipton & Olsen, 2018) by providing concrete recommendations aligned with SEER principles, along with a hypothetical example to illustrate how to implement these recommendations.

This guide is intended for producers of research–researchers who conduct impact evaluations of educational interventions and funders of those evaluations–but is motivated by the demand for rigorous evidence from consumers of research like local, state, and federal education policymakers.

This guide focuses on prospective, multi-site impact evaluations for which researchers need to select and recruit a certain number of sites. These studies face the challenge of recruiting an adequate sample of sites–and one similar enough to the study's primary target population. This guide offers recommendations on how to address this challenge. The guide also provides recommendations on how to assess and improve the generalizability of the study, given the sites that participate. These recommendations are applicable to both prospective impact evaluations that need to recruit sites *and* retrospective, quasi-experimental studies based on extant data. Figure 1 summarizes the recommendations presented in this guide for prospective, multi-site impact evaluations, and it lists these recommendations in the order in which prospective evaluations would naturally implement them–starting with defining the target population and developing a population frame. Recommendations about developing a sampling plan and recruiting the sample will not typically apply to retrospective impact studies–most begin with a sample from existing data–and they never apply to retrospective generalizations from completed studies. But most of the recommendations are applicable to any attempt to generalize from an impact evaluation.

Figure 1. Overview of the recommendations found in this guide

Recommendations	Steps
1. Define the target population	Identify potential moderators–characteristics that influence the impact of the intervention–and the primary target population of students and schools
2. Develop a population frame	Select and possibly combine data sources, explore the data, and refine the target population definition
3. Design a sampling plan	Determine the sample size, stratify the population, set recruitment targets, and design a plan for selecting schools within strata
4. Implement the sampling plan	Plan for recruitment, build and manage a team, screen out ineligible schools, and collect data on volunteers and decliners
5. Assess similarity	Compare the sample and population on observed potential moderators and explore threats from unobserved moderators
6. Adjust for differences	Reweight the sample, redefine the target population if necessary, and consider if generalizations are warranted
7. Report generalizability appropriately	Integrate generalizability into all facets of data collection and reporting

The recommendations in this guide apply to a wide range of impact studies in education. The guidance is applicable to three key types of studies: (1) randomized controlled trials that randomly assign schools, teachers, classrooms, or students; (2) regression discontinuity designs and quasi-experimental designs; and (3) evaluations of interventions targeted at schools, teachers, or students, regardless of whether the intervention is mandatory or voluntary. It also applies to studies that focus on a particular subgroup or subpopulation of students. The recommendations in this guide are applicable as long as the impacts of the intervention may vary across sites, so an impact study in an unrepresentative sample of sites would produce impact findings that do not generalize to the population.

For simplicity, this guide focuses on impact evaluations conducted in K–12 schools. The guidance provided here applies equally to impact studies conducted in a range of educational contexts, including early childhood centers and postsecondary institutions. But for ease of exposition, the narrative in this guide refers to the sites in which the evaluation is conducted as “schools.”

Throughout the guide, we illustrate the recommendations with a running example. This example is hypothetical but rooted in research on generalizability and our experience advising researchers on the topic. We have tried to balance providing an example specific enough to be clear and informative while broad enough to be useful to a wide range of impact evaluation contexts. We present the running example using *The Generalizer* (Tipton & Miller, 2021), a free web tool, to inform readers about simple software they can use to implement many of recommendations in this guide (see *A note about software*).

Finally, this guide focuses on one specific facet of external validity–generalizations of findings from a sample of schools to a target population of schools. This focus is driven by the challenges in obtaining samples of schools that represent the target population. The guide does not delve into challenges in obtaining representative sample of teachers and students within participating schools, either because the intervention is voluntary or the study is voluntary, and teachers and students can opt out of one or both. In addition, the guide doesn't cover other facets of external validity (see Shadish et al., 2002). These include methods for understanding and explaining variation in intervention impacts (for example, mediation analysis, moderation analysis); methods for predicting school and student specific treatment effects; methods for understanding the stability of effects across studies (for example, replication, meta-analysis); methods for making out-of-population predictions (transportability); and methods for addressing missing data that can influence the generalizability of the study's findings as well its internal validity.

A note about software:

Throughout this guide, we reference software tools that can be used to implement the methods provided in the guide. These include:

- *The Generalizer* (www.thegeneralizer.org; Tipton & Miller, 2021), a free web tool with two built-in datasets that can be used to (1) specify a target population, stratify the population, and develop a sampling plan, including lists of schools for recruitment, and (2) assess similarity between a sample of schools and a target population.
- Generalize (https://nustat.github.io/generalizeR; Ruel et al., 2022), a free R package that can be used with any population data and has the same capabilities as *The Generalizer* but can also estimate the average treatment effect using poststratification weights.

Elizabeth Tipton, an author of this guide, is also an author of both of these software tools. To our knowledge, no other tools with a focus on generalization in education research exist.

Recommendation 1. Select the Target Population for the Study

Goal: Define the target population of the study

Steps:

1. Identify the potential moderators that influence the impact of the intervention.
2. Identify the target population of students on which the study will focus.
3. Identify the target population of schools on which the study will focus.

Resources:

- Logic model for the intervention
- Prior evidence
- Developer of the intervention

What to report:

1. Potential moderators of the impact
2. The target population of students and schools on which the evaluation will focus

Clearly identifying the target population—that is, the population of primary interest in the study—is the single most important step you can take to enhance the generalizability of your impact study findings. Identifying the target population provides direction to the design, implementation, analysis, and reporting of study findings, as described in the remainder of this guide. For evaluations of educational interventions, the target population should usually be defined as the collection of students, teachers, and/or schools over which the study aims to estimate the average impact.

For studies that aim to estimate the impact on student outcomes, we recommend defining the target population in terms of students *and* schools (e.g., the students who receive the intervention and the schools in which the intervention is delivered). This will allow the study to estimate the impact of the intervention for the average student *or* the average school within the collection of students and schools that comprise the target population. More generally, defining the target population in terms of the schools included is also useful for any study that plans to recruit a sample of these schools that are willing to participate (e.g., willing to implement the intervention or to allow their teachers to participate).

Consider the factors that influence the impact of the intervention and use them to define the types of students and schools about which the study aims to learn. If necessary, narrow the target population to the types of students and schools you can realistically include in the study, while being careful to balance the trade-offs between feasibility and policy relevance. When defining the target population, obtain input from key members of the study team—and the study's funder and other key stakeholders, where appropriate–and communicate the target population definition to all members of the team. If the study funder is actively involved, make sure it agrees with the decision. Commit to your target population from the start and use it to inform all aspects of the study.

How to carry out the recommendation

1. Identify potential moderators that influence the impact of the intervention

The impact of the intervention will likely vary across populations of students and schools. The variables that influence the intervention's effectiveness are referred to as "moderator" variables, or sometimes as "treatment interactions" (see Weiss et al., 2014, for a general framework). Identifying potential moderators is a critical first step in defining the target population for your study and distinguishing it from other populations for which the intervention may be more effective or less effective.

The following factors may moderate the impact of the intervention that you plan to study:

- **Student characteristics.** What are the characteristics of the students who could participate in the intervention? Which of these characteristics are likely to be associated with the impact of this intervention?
- **Educational context.** What are the characteristics of schools that could implement this intervention? What are the characteristics of the districts and communities in which those schools are located? Which characteristics are likely to be associated with the impact of the intervention?
- **Counterfactual.** To which interventions or services–or "counterfactual"–will the tested intervention be compared? Many impact studies in education compare the intervention to a "business-as-usual" counterfactual that varies across schools in the study. What business-as-usual conditions might you expect to be associated with differential impacts?
- **Implementation.** How will the implementation of the intervention likely vary across schools in the study? Which dimensions of implementation will likely influence the impact of the intervention?

To understand the factors that moderate the impact of the intervention, you can:

- **Explore the intervention's logic model.** The logic model may identify or at least allow reasonable inferences about the types of students or schools most likely to benefit from the intervention.
- **Review prior evidence.** Evidence from prior studies on a version of the intervention may indicate the types of students or schools for which the intervention has been most effective or the contexts that have been previously studied. Importantly, keep in mind that while the statistical significance of a moderator certainly suggests it should be

included, the lack of statistical significance is not sufficient evidence to exclude a moderator because moderator analyses are very often under-powered.

- **Talk with the intervention developer and others knowledgeable about the intervention.** The intervention developer may have a strong sense, based on its experience developing and implementing the intervention, of the types of students and schools that benefit most from the intervention.

Finally, note that it is impossible to know *a priori* which variables moderate the impact of the intervention. In general, for this reason, it is wise to err on the side of including a potential moderator if you are uncertain.

2. Identify the target population of students on which the study will focus

In any impact evaluation, you must address the following question: for which students, or types of students, do we hope to learn about the impact of the intervention? It's not necessary to include all students that the intervention is designed to serve in the study's target population, but it helps to identify from the outset the population of students on which the study will focus. For example:

- An impact evaluation of Response to Intervention (RTI) may want to learn about the impacts of RTI services on K-5 students whose reading achievement is more than two grades below grade level.
- An impact evaluation of a virtual learning intervention may want to learn about its impacts on K-2 students with two particular disabilities.
- An impact evaluation of a federal mentoring program may want to learn about its impacts on high school students who participate in the program.

You should use the student characteristics identified earlier in Step 1—those likely to moderate the impact of this intervention—to define the target population of students. For example, if there are good reasons to expect the impact to depend on student's initial achievement, use prior achievement to define the target population of students on which the study will focus (e.g., students with below grade level mathematics achievement). There is no need to define the target population of students based on student characteristics that are not believed to moderate the impact of the intervention.

Finally, some studies might have multiple target populations of interest. For example, average treatment effect estimates may be desired for both an entire region and for states within the region. In this case, to apply the lessons from this guide, define the primary target population as the broadest of these—for example, the region. Tipton (2022) provides additional approaches in the case of multiple populations, but these are beyond the scope of this guide.

3. Identify the target population of schools on which the study will focus

To identify the types of schools on which to focus, first consider the schools of primary interest to the study's funder or primary stakeholder (see Stuart et al., 2017). For example, the U.S. Department of Education has sponsored evaluations to test educational interventions in a collection of schools that receive federal funding from a particular program. In these cases, the primary target population of interest to the funder–here, the federal government–may be all schools that receive or are eligible to receive funding from this program. Alternatively, school districts sometimes provide or obtain funding to evaluate an educational intervention in a collection of its schools. In these cases, the primary target population of interest to the funder–here, the school district–may be all district schools that could implement the intervention (e.g., all elementary schools in the district if the intervention focuses on elementary school grades).

Next, consider where students in this target population are enrolled. If your study is focused on a narrow target population (for example, students with a rare disability), students in the target population may be concentrated in a narrow set of schools. If so, you may want to restrict the population to include schools with a minimum number or share of target students.

Next, revisit the potential moderators that may vary across schools–such as those that capture the educational context–then consider the type of study you plan to conduct, and define the target population of schools in a way that would expand the evidence base for the intervention:

- **Initial efficacy trial.** Initial efficacy trials are the first rigorous impact study on the intervention. Because they typically aim to evaluate the intervention's effects under favorable circumstances, the target population should include schools where the values of the moderator variables identified in Step 1 above suggest that the impacts are likely to be large (for example, focusing on rural schools if urbanicity is a potential moderator and the impact is expected to be larger in rural areas than urban areas).
- **Pure replication studies.** Pure replication studies attempt to replicate prior evidence of impact in the same or a similar target population as previous studies. These studies should consider characteristics of prior study samples (for example, urban schools in the northeast) and define their target population to match.
- **Systematic replication studies.** Systematic replication studies, sometimes referred to as conceptual replications (Chhin et al., 2018; Coyne et al., 2016; Schmidt, 2009), vary one key dimension of the study, which may include the target population. If prior evidence strongly suggests that the intervention is effective for certain types of schools (for example, urban schools), a systematic replication study might decide to focus on other types of schools for which impacts may be different (for example, suburban or rural schools).

- **Effectiveness or scale-up studies.** These studies extend prior research to estimate the impact of the intervention when implemented on a broader scale and for a broad target population of schools (for example, all schools that could potentially implement the intervention).

Next, you should consider the types of schools that are feasible for you to recruit given the resources available for the study. For example, you may ideally wish to generalize to the target population of *all* schools in the United States, yet realistically, you may only be able to recruit in a single region or state. This means that you need to narrow your target population.

Finally, consider the tradeoffs between defining the target population broadly and defining it more narrowly. A broader population has advantages when:

- **The purpose of the study is to inform a single policy decision affecting a broad set of schools** (for example, an impact evaluation of a federally funded reading program that could be used to increase or decrease the program's funding). In these cases, define the target population broadly to match the population that would be potentially affected by the policy decision. This will maximize the policy relevance of the study findings.
- **The study could potentially inform a collection of state or local policy decisions** (for example, an impact evaluation of a curriculum that could by adopted by districts throughout the country). In these cases, defining the population broadly enough to include all schools potentially affected by state or local policy decisions that could be informed by the study findings would make the study relevant for a larger set of decisionmakers.

At the same time, there are good reason to define the population more narrowly for some studies. When an impact evaluation is being conducted for a single local policymaker (e.g., a single school district) to inform local policy decisions, defining the population narrowly to include local schools will provide the most relevant information for that policymaker. In some cases, however, even when the study could inform decisions by a wider range of policymakers (e.g., school districts throughout the country), you may choose to define the population more narrowly if your research team is only able to recruit schools from a narrow population. The decision of how narrowly or broadly to define the target population is important because–when combined with later recommendations in this guide–it directly affects how broadly or narrowly the findings from the study can be generalized.

Example

Intervention

A team of cognitive psychologists has worked for several years developing and refining a new reading curriculum for early elementary school students (K-2). The program is classroom based and includes teacher professional development, whole class lessons, small group instruction, and reading materials in which Black and Hispanic families are well represented in the content of the materials (for example, including reading passages focused on Black and Hispanic individuals). Better representation of Black and Hispanic families in reading materials is intended to improve student engagement and learning for Black and Hispanic students.

Potential moderators

Based on the literature on reading instruction, the research team expects the intervention effect to be associated with student background characteristics (for example, primary language, family socioeconomic status) and their representation in books and reading materials (for example, race, ethnicity, gender). Additionally, because prior studies focused on large urban schools, the team is not sure if the program effects might differ in other settings (for example, rural schools, smaller schools, smaller districts). Finally, the team expects that the size of the intervention effect will depend, too, on the 'business-as-usual' reading program they would have used otherwise. For this reason, the team has identified primary language, socioeconomic status, race, ethnicity, gender, urbanicity, school size, and business-as-usual reading program as potential moderators of the intervention.

Target population

Given the nature and goals of the curriculum, the evaluation will focus on a target population of public schools that meet all three of the inclusion criteria listed below:

- Serve students in grades K-2 since this is the appropriate grade level for the intervention.
- Serve a student population that is at least 50 percent non-White, since the intervention is focused on improving the outcomes of Black and Hispanic students and is expected to have a larger effect on achievement for these students than for White and Asian students.
- Located in the southeast, for logistical feasibility. In principle, the research team is interested in learning about the effects of the intervention nationwide. But in practice, the team will focus on schools in the Southeast because the team includes a principal investigator (PI) in Georgia and a co-PI in Florida, so recruiting in the Southeast will be more logistically feasible.

In the next section (Recommendation 2), this target population will be operationalized and enumerated using data.

Recommendation 2. Develop a Population Frame

Goal: To develop a population frame with data on schools that could potentially be included in the study

Steps:

1. Select a primary data source for the population frame
2. Add data from other sources, if needed
3. Explore the population frame data and refine the definition of the target population, if needed
4. To the extent possible, restrict the data to schools that are within the target population

Resources:

- *The Generalizer*, a free web tool (www.thegeneralizer.org; Tipton & Miller, 2021), includes the Common Core of Data (CCD) and the Integrated Postsecondary Data System (IPEDS)
- generalize, an R package (https://nustat.github.io/generalizeR; Ruel et al., 2022), also includes CCD and IPEDS data

What to report:

1. Inclusion/exclusion criteria for the study
2. The total number of schools included in the population frame
3. Summary statistics on potential moderators in the population frame

To facilitate the selection and recruitment of schools into the sample, assemble a dataset–called a population frame–that enumerates and describes the schools in the target population. In this frame, each row should include a school in the target population, variables that capture potential moderators, and contact information. We recommend constructing the population frame at the school level, even when the target population is defined at the teacher level or the student level, and even when the study aims to include samples of teachers and/or students. In any case, a list of eligible schools is a useful tool in supporting the study recruitment effort because school permission is typically required before asking teachers and students from a school to participate in the study. Assembling a population frame may require merging data from different sources. To the extent possible, exclude schools that fall outside of the target population. But when this is not possible due to data limitations, ineligible schools can be excluded from the study later during the school recruitment phase.

How to carry out the recommendation

1. Select a primary data source for the population frame

Select the most recent data source that includes all of the schools in the target population and as many of the potential moderators identified previously as possible. This data source could come from the federal government or from states, as described below and summarized in Table 1:

- **National data.** The federal government collects annual data on many educational institutions, including public and private schools as well as pre-K programs and postsecondary institutions. These data typically include enrollment; demographics (for example, racial/ethnic, gender); location (for example, urban, rural); and funding (for example, per pupil revenue). These data are useful both for studies focused on national target populations and, when selecting a subset of the data, for studies focused on state or local target populations.
- **State data.** To date, the federal government has funded the development of state K-12 longitudinal data systems in 49 states. These data systems often include information on teachers and students, including on student outcomes (for example, aggregate state test scores, accountability metrics).

Table 1. Common data sources available for developing K-12 population frames

Data source	Scope	Level	Information
Common Core of Data (CCD)	National	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
ED*Facts*	National	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)	National	School district	Indicators of social, economic, and housing conditions for school-age children and their parents
Private School Universe Survey (PSUS)	National	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)	National	School district, school	Measures of academic achievement and achievement gaps for public schools
State longitudinal data systems	State	Student	Longitudinally linked student achievement measures from state accountability tests and other information

For impact evaluations conducted in K-12 public schools, the logical starting place to build a population frame is the Common Core of Data. The Common Core of Data (CCD) is a census of all public schools and districts nationwide, and it includes information on the number of students enrolled—by grade level and in different categories (for example, by race and ethnicity)—as well as the number of teachers, the type of school, the location of the school, and other variables. The CCD's school and district files can be easily merged to construct a population frame of schools that also includes district characteristics. The Generalizer (Tipton & Miller, 2021) can be used to access CCD data and follow many of the remaining recommendations made in this guide.

For impact evaluations conducted outside of K-12 schools, the logical starting place will vary. For example, for impact evaluations conducted in

- K-12 private schools, start with data from the Private School Universe Survey
- Postsecondary institutions, start with data from the Integrated Postsecondary Data System (IPEDS), also currently available in *The Generalizer*
- Head Start centers, start with data from the Early Childhood Learning & Knowledge Center at the U.S. Department of Health and Human Service's Administration for Children and Families

For some types of institutions, such as state pre-K programs, there may be no national list, and assembling a complete list may be infeasible or cost prohibitive. For studies conducted in these types of institutions, we recommend starting with the American Community Survey (ACS) to identify a population frame of counties in which all of the institutions in the target population are located. Then select a representative sample of counties (see Recommendation 3) and assemble or collect data needed to identify the institutions that are part of the target population—and if possible, data on potential impact moderators—within the selected sample of counties. These data will comprise the population frame for the evaluation.

2. *Add data from other sources, if needed*

The primary data source selected for the population frame may lack important information, including variables needed to screen out ineligible schools—schools that are not part of the target population—and variables on potential impact moderators. Building the final population frame for the study therefore often involves merging together data from various sources.

Given the data limitations of the CCD, researchers may want to add the following information to the population frame:

- Community characteristics (for example, poverty, housing) from the ACS
- Program usage from program developers

- State policies from the Education Commission of the States webpage (www.ecs.org) as well as state Department of Education websites
- District policies "scraped" from district websites (Wright et al., 2021)
- Identification of schools that receive funding from some program, particularly for studies motivated by the interests of the program funder

If the available data sources do not capture important moderators or eligibility factors that define the target population, consider conducting an initial survey of potentially eligible schools to obtain this information. An initial survey may be feasible if this set of schools is sufficiently small, the survey is inexpensive to administer, or the budget for the study is sufficiently large.

3. To the extent possible, restrict the data to schools that are within the target population

In many studies, variables in the population frame will allow your team to distinguish schools that are part of the target population from schools that are not. For example, if you define the target population to include urban schools in three Northeastern states, the data from the CCD can be used to restrict the population frame to those schools.

However, in other studies, some of the factors that define the target population may not be available in the population frame data. For example, the target population may be restricted to schools that are currently not offering teachers professional development (PD) that is similar to that provided by the intervention, but data on the PD offered by schools in the population frame may not be available. In these cases, accept that the population frame will include some schools that fall outside the target population. Later, we provide guidance on how to screen out these schools during the recruitment process (see Recommendation 4), assess the generalizability of the sample relative to the true target population, excluding ineligible schools (see Recommendation 5), and adjust for differences between the sample and this target population (see Recommendation 6).

4. Explore the population frame data and refine the definition of the target population, if needed

Once the population frame has been assembled, explore the data to examine the characteristics of the population frame through summary statistics (e.g., means, standard deviations) and visual displays (e.g., histograms). This analysis may reveal issues with data quality and completeness that need attention.

Exploring the population frame may reveal information about the target population that leads you to revisit and refine the definition of the target population. For example, it may reveal segments of the population in which the intervention would be difficult to implement, such as rural schools that may lack the resources necessary to support implementation. Alternatively, it may reveal a population that is either more similar to the target populations from previous studies—or more different from those populations—than you intended, given the contribution that the study aims to make to the evidence base. In these cases, consider refining the target population based what you learned from exploring the population frame data. In practice, the process of exploring the data in the population frame and refining the target population is more iterative than linear. For example, if rural schools are removed from the target population, exclude them from the population frame, produce and explore new summary statistics and visualizations, and consider whether additional refinements are needed to arrive at a final target population definition and frame.

Example

To operationalize the target population criteria specified earlier, the team used *The Generalizer* web tool (Tipton & Miller, 2021), which includes both 2018-19 CCD data and some data from the 2018 ACS. In *The Generalizer*, the population frame was constructed to include Title I "regular" public schools serving K–2 students in the Southeast (Alabama, Florida, Georgia, Kentucky, Maryland, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia). *The Generalizer* automatically excludes schools with fewer than 20 students or 2 teachers. Charter schools and schools with more than 50 percent of White students were excluded. The final population frame included 4,149 schools.

Using *The Generalizer*, the team operationalized the potential moderators by selecting various aggregate measures of student demographic characteristics and other school and district characteristics. The team selected four demographic measures aggregated to the school level (percentage of students eligible for free or reduced-price lunch and percentages of female, Black, Hispanic, and Native American students). Note that eligibility for free or reduced-price lunch is included as an indicator of socioeconomic status. In addition, the team selected two demographic measures aggregated to the district level because school-level measures were unavailable from the CCD (percentage of English language learners and percentage of students speaking a language other than English at home). Finally, the team included contextual variables in the population frame (urbanicity [rural, urban, town, suburban]; number of students per school, and number of schools per district).The population frame did not include any information on the reading programs currently used by schools. Since this is unavailable, the team will collect information on this during recruitment. Characteristics of the population frame are in Table 2.

Table 2. Characteristics of the population frame (from *The Generalizer*)

Potential moderator	Data source	Mean (standard deviation)
Students eligible for free or reduced-price lunch (%)	CCD (school)	78.2 (23.9)
Female students (%)	CCD (school)	48.4 (2.0)
Black students (%)	CCD (school)	50.2 (20.2)
Hispanic students (%)	CCD (school)	24.8 (16.4)
Native American students (%)	CCD (school)	0.6 (3.2)
English language learners (%)	CCD (district)	8.3 (4.1)
Home language other than English (%)	ACS (district)	17.7 (11.5)
Urban school (%)	CCD (school)	38.1 (37.5)
Suburban school (%)	CCD (school)	35.6 (35.5)
Town school (%)	CCD (school)	10.0 (22.7)
Rural school (%)	CCD (school)	16.3 (28.0)
Number of students per school	CCD (school)	562.1 (163.4)
Number of schools per district	CCD (school)	120.4 (90.4)

The goal will be to recruit a sample of schools into the study that is, on average, similar to the population frame on the same characteristics found in Table 2. In this defined population frame–Title I public schools serving primarily Black and Hispanic students in grades K-2 in the Southeast–the average school serves predominately low-socioeconomic status (SES) students (78.2 percent). The average school in this population frame is found in a large school district (including about 120.4 schools). In these schools, on average, about half of the students are Black (50.2 percent) and quarter are Hispanic (24.8 percent). These schools vary in their location, with the majority in urban (38.1 percent) and suburban (35.6 percent) locales. Just over a quarter of the schools are in towns or rural areas (10.0 percent and 16.3 percent, respectively). Finally, in the districts including these schools, about 8.3 percent of students are English language learners (ELLs) and 17.7 percent of families speak a language other than English at home.

Recommendation 3. Design a Sampling Plan

Goal: To develop a plan for obtaining a representative sample from the target population

Steps:

1. Determine the number of schools needed for the study.
2. Stratify the schools in the population frame based on potential moderators.
3. Set recruitment targets for each stratum.
4. Design a plan for selecting schools within strata.

Resources:

- *The Generalizer*, a free web tool (www.thegeneralizer.org; Tipton & Miller, 2021)
- generalize, an R package (https://nustat.github.io/generalizeR; Ruel et al., 2022)
- Statistical power analysis (e.g., PowerUp!; Dong & Maynard, 2013 and Optimal Design; Spybrook et al., 2011)

What to report:

1. The variables used to construct the strata and the method used for creating the strata (such as k-means clustering)
2. The number of strata defined, the proportion of schools from the population frame in each stratum, and the number of schools to be recruited from each stratum
3. The strategy for sampling schools within strata

To produce findings that generalize to the target population, design a sampling plan for selecting a sample of schools that is as representative of the target population as possible. To that end, use data on potential moderators to stratify the schools in the population frame and set recruitment targets for each stratum. Then choose a sampling approach for selecting a representative sample from each stratum. Developing and implementing a sampling plan designed to produce a representative sample of the population frame will increase the likelihood that the final sample is similar to the target population and that the study findings apply to that population.

The recommendations described below focus on selecting a sample of schools. But larger impact studies often require multiple school *districts*—or would benefit from including multiple districts to broaden the generalizability of the study findings. For these studies, use the steps described below with minor modifications to select a sample of districts from which to select and recruit a representative sample of schools (see the *Advanced techniques* at the end of this section for more details). Finally, although this guide does not directly address the sampling of classrooms, teachers, or students within schools, the same principles and approaches could easily be extended to these situations.

How to carry out the recommendation

1. Determine the number of schools needed for the study

Before selecting schools, you must determine the target number of schools and students to include in the sample. These targets should be set to ensure that the study has adequate statistical power to detect the impact of the intervention. The target sample size can be calculated using standard formulae (for example, Schochet, 2008; Hedges & Rhoads, 2009) and downloadable software (e.g., PowerUp!, Dong & Maynard, 2013; Optimal Design, Spybrook et al., 2011). These calculations require assumptions, ideally based on prior evidence, for key parameters (such as the intraclass correlation), and these should be aligned with the target population to the extent possible (for example, Hedges & Hedberg, 2007, 2013).

2. Stratify the schools in the population frame based on potential moderators

Stratification is a tool that can help in recruiting a sample that is representative of the population frame in terms of potential moderators. In stratification, the population frame is divided into strata—also called 'blocks' or 'bins'—and part of the sample is recruited from each. More specifically, stratification entails setting fixed sample size targets for each stratum (for example, 6 schools from Stratum A and 14 schools from Stratum B). It is commonly used in survey research to increase precision; it is also used in evaluation research to ensure geographic and other forms of diversity in the sample. Overall, stratification improves the similarity between the sample and target population—thus reducing bias and increasing precision.

Categorical variables lend themselves easily to stratification. For example, the population of schools nationwide could be divided into six strata by all combinations of school level—elementary, middle, and high—and locale—urban versus other. Stratification using continuous variables is more complicated because there are many possible ways of setting thresholds that divide continuous variables into discrete strata. One approach would be to first create categorical versions of each continuous variable—e.g., 'high SES' versus 'low SES' schools and large schools versus small schools—and then to cross them with each other to create strata—e.g., small and high SES schools, small and low SES schools, large and high SES schools, and large and low SES schools. Using this approach, however, results in a large number of strata; when there are p potential moderators, this would mean 2^p strata, which is infeasible if that is more strata than the number of schools needed for the study. Dimension-reduction methods such as cluster analysis, latent variable analysis, and factor analysis can be used to construct a smaller number of strata and estimate the proportion of the variance of potential moderators that is explained by those strata. These approaches take advantage of the fact that many potential moderators are correlated with one another. *The Generalizer* will allow you to use k-means clustering (see Tipton, 2014b) to divide the population frame and describe the characteristics of each cluster.

Whatever method you use to stratify the population, you must choose the number of strata into which to divide the population. Additional strata provide more protection against obtaining a sample that differs from the population frame on potential moderators because they increase the proportion of the variation in the moderators explained by the strata. However, the inclusion of additional strata can also increase the resources required for recruitment, since additional strata impose additional constraints on the recruitment process. Therefore, you face a trade-off. Additional strata are only advisable if the additional variability explained is large relative to the additional constraint. For this reason, it is helpful to consider a few options (e.g., 4, 5, or 6 strata) and the proportion of variation between strata (versus within strata).[1] In the ideal, *all* of the variation would be between strata, indicating that schools in the same stratum have exactly the same values of the potential moderators, but this is never possible in real data.

There has been no research to date on the optimal number of strata for selecting representative samples of sites for impact evaluations. However, related literature that focuses on a different problem—the selection of comparison units for quasi-experimental designs—has established rules of thumb for the number of strata (for example, Cochran, 1968; Rosenbaum & Rubin, 1984). Therefore, for now, we generally suggest following those rules of thumb and defining four to six strata for selecting and recruiting schools.

Finally, once the strata have been defined, look at summary statistics or figures that help you understand how schools in the strata differ from each other (for example, see Figure 1). This information may help school recruiters in a particular stratum better understand the types of schools they need to contact; it is also useful for describing the population frame in study reports and other publications.

3. *Set recruitment targets for each stratum*

Decide how many schools should be recruited from each stratum. In most cases, set recruitment targets proportional to the population within each stratum.[2] Doing so will reduce the likelihood that the statistical (for example, weighting) adjustments described in Recommendation 5 will be needed to make the sample comparable to the population frame. For example, suppose that the study's power analysis suggests that your study should include

[1] There are several available statistical metrics that can be used as well in making this decision (for a review, see Tibshirani et al., 2001).

[2] At a minimum, however, the evaluation should include enough schools from each stratum to estimate a treatment impact. In a cluster randomized trial that randomizes schools within stratum, this would mean a minimum of two schools per stratum.

50 schools, and suppose that 20 percent of schools in the population frame are in Stratum 1. The study team should then aim to recruit 20 percent of the sample—or 10 schools—from Stratum 1.

However, in other cases, it may make sense to oversample schools from certain strata. Impact evaluations that will focus on certain subgroups of schools within the target population (for example, especially low-performing schools) may need to oversample schools of that type to reach an adequate sample size. In addition, impact evaluations that need a substantial number of eligible students in each participating school—either for statistical power or to deliver the intervention (for example, in a classroom setting)—may want to oversample larger schools. However, it is important to note the following tradeoff: when schools are oversampled from selected strata, you need to use weights to estimate the average treatment effect, and all else equal, those weights reduce the statistical power of the analysis. To achieve the same precision as a proportional sample, you would need to recruit a larger overall sample.

4. Design a plan for selecting schools within strata

Within strata, choose a method for selecting schools that is designed to produce a representative sample. You have two options:

1. **Probability sampling**. Probability sampling involves setting the probability of selection into the sample for each school in the population frame and then selecting schools with those probabilities. This probability may be equal for all strata or set higher for strata from which the researchers plan to oversample (see the previous discussion under Step 3).
2. **Balanced sampling.** Balanced sampling involves selecting schools to maximize the degree of similarity between the sample and the population frame on observed moderators.

In either case, you can rank order and recruit schools within strata in order until sample size targets overall and within strata are reached. For the most common form of probability sampling—stratified random sampling (see Olsen & Orr, 2018)—schools can be ordered randomly within strata. For balanced sampling, schools can be ordered within strata based on their distance to the stratum mean, from most similar to the stratum mean to least similar to the stratum mean (for example, using Euclidean distance on the set of moderators; see Tipton, 2014b). *The Generalizer* rank orders schools using this balanced sampling approach.

Recruiting from ordered lists for each stratum facilitates the process of selecting replacement schools when initially selected schools decline to participate. For example, building on an earlier example, suppose that the study aims to recruit 10 schools from Stratum 1. In this case, recruiters should attempt to persuade the first 10 schools on the ordered list for Stratum 1 to participate in the study. However, if four of those schools opt out of the study, recruiters can turn to the schools ranked 11-14 and recruit them to participate. Replacement schools may differ from the schools they replace in unobserved ways, yielding a sample that differs from the population frame on some variables that moderate the impact of the intervention. But selecting replacements within the same strata as the original schools provides some protection against the bias from schools self-selecting into or out of the impact evaluation.

In some cases, research funders require prior evidence of support from districts and schools before sponsoring the study. Thus, you may have to commit to including certain schools before the study is funded and the recruitment process formally begins. In these cases, simply identify the strata to which these schools belong and count these schools toward meeting the targets for those strata. For example, if the study aims to recruit 10 schools from Stratum 1, but 2 of the schools to which the study has pre-committed come from Stratum 1, include those 2 schools in the sample and recruit an additional 8 schools from Stratum 1 after funding has been received and the recruitment process begins.

Example

The research team conducted a statistical power analysis to determine that they needed 40 schools to detect an average impact of 0.15 standard deviations, the Minimum Detectable Effect Size set for the study. They used *The Generalizer* to develop options for dividing the schools in the population frame into strata; here they focused on creating between four and six strata. After comparing the proportions of variance explained by each of the stratification options and discussing resources and planning, the team decided to select the option with five strata. *The Generalizer* then used cluster analysis to divide schools in the population frame into five strata based on the moderators included in the frame. Stratum membership explained 43.8 percent of the variance in these moderators across schools. Importantly, this means that the remaining 56.2 percent of the variation in these potential moderators is *within* strata, indicating that the strata are not homogenous. The result is that there is still significant variation between schools within strata, which means that within-strata recruitment needs to proceed carefully to achieve a sample representative of the stratum-specific population frame.

Table 3. Characteristics of population frame divided into five strata

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
Total number of schools	285	1,307	947	203	1,407
Students eligible for free or reduced-price lunch (%)	4.8 (13.3)	79.9 (17.2)	75.9 (16.4)	78.9 (14.2)	93.1 (10.1)
Female students (%)	49.2 (3.6)	48.1 (2.4)	48.2 (2.4)	47.9 (2.8)	48.7 (2.9)
Black students (%)	53.7 (29.1)	37.3 (15.7)	28.2 (21.7)	32.2 (34.9)	78.7 (16.2)
Hispanic students (%)	19.0 (17.6)	23.6 (14.1)	48.2 (20.0)	62.4 (33.7)	6.1 (6.0)
Native American students (%)	0.2 (0.2)	1.4 (7.5)	0.3 (0.6)	0.0 (0.1)	0.3 (0.7)
English language learners (%)	8.0 (4.7)	6.4 (3.4)	14.7 (4.3)	19.1 (0.0)	4.3 (3.6)
Home language other than English (%)	11.1 (5.1)	11.5 (5.0)	30.6 (8.6)	73.8 (0.0)	8.1 (6.5)
Urban school (%)	76.8 (42.9)	33.1 (47.1)	22.8 (42.0)	17.7 (38.3)	48.3 (50.0)
Suburban school (%)	8.8 (28.3)	35.8 (48.0)	67.6 (46.8)	81.3 (39.1)	13.0 (33.6)
Town school (%)	6.7 (25.0)	10.5 (30.6)	4.0 (19.6)	0.0 (0.0)	15.5 (36.2)
Rural school (%)	8.8 (28.3)	20.7 (40.5)	5.6 (23.0)	1.0 (9.9)	23.2 (42.3)
Number of students per school	553.4 (218.3)	596.2 (189.5)	689.2 (229.7)	673.9 (344.2)	431.8 (152.8)
Number of schools per district	120.8 (76.7)	66.5 (57.9)	191.8 (102.4)	529.0 (0.0)	64.1 (69.9)

Table 3 provides characteristics of these strata and can be used to describe these different subsets of the population frame. At the top, the total number of schools in the population frame in each stratum is listed. Distinctive features of schools in each strata are listed below:

- **Stratum 1:** Urban and about 50 percent Black
- **Stratum 2:** In smaller school districts and about 60 percent Black or Hispanic
- **Stratum 3:** In larger suburban districts and about 75 percent Black or Hispanic
- **Stratum 4:** In larger suburban districts, almost entirely Black or Hispanic, and in communities where English is not the home language
- **Stratum 5:** Smaller schools, about 75 percent Black, and almost entirely eligible for free or reduced-price lunch

The research team also used *The Generalizer* to divide the total sample size target of 40 schools into the stratum recruitment targets shown in Table 4. Based on proportional allocation, these range from 2 schools (Stratum 4) to 14 schools (Stratum 5).

Finally, when this plan was developed, the team had already recruited 10 schools to be part of the study; these schools were included as letters of support in the grant proposal. The team located these 10 schools in the data and found that they fell in Strata 2 and 3, with 5 schools in each. Therefore, the research team needed to recruit 8 additional schools from Stratum 2 and 4 additional schools from Stratum 3 to reach the sample size targets of 13 and 9 for Strata 2 and 3, respectively.

Table 4. Stratum allocations in the population frame and sample

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
Number of schools in the population frame	285	1,307	947	203	1,407
Percentage of schools in the population frame	7	32	23	5	34
Number of schools in the study sample (Proportional allocation, *n* = 40)	3	13	9	2	13

Moving forward, the plan was to use balanced sampling within strata to identify and recruit selected schools. To do so, using the same potential moderators as identified in Table 2, the research team first calculated averages for each variable within each stratum. Within each stratum, for each school, they then calculated the standardized Euclidean distance between the school and the average school in the stratum. Based on these distances, the team ranked the schools within each stratum from first (most similar to the stratum average) to last (least similar). These rankings were provided by *The Generalizer*. As indicated in the next recommendation, recruitment within each stratum then proceeded from the top of this list until the full sample was successfully recruited.

Advanced techniques

Until now, we have focused on the simplest scenarios in which there is a single target population, a sample can be selected from a population frame that includes sites in the target population, and the primary or only goal of the study is to estimate the average treatment effect in the same target population. However, additional methods may be worth considering in special cases:

1. **The study aims to generalize to a target population over a broad geographic area.** Studies that aim to learn about the average impact in a broad population, like the entire country, likely require samples spread throughout the country to justify the underlying assumptions needed for generalizations. To obtain a geographically diverse sample of schools at reasonable cost, research teams can select two-stage samples by selecting a sample of school districts and then a sample of schools within those districts. Although two-stage sampling is often used in large-scale randomized trials conducted for the U.S. Department of Education (for example, Gleason et al., 2019; Heppen et al., 2020; Herrmann et al., 2019), the generalizability of these studies would likely be improved by

using probability or balanced sampling at both stages to select representative samples of districts and schools. Although rare in impact studies in education, two-stage probability sampling of this general type was used in the Head Start Impact Study to select a representative sample of program grantees and then a representative sample of Head Start centers within grantees (Puma et al., 2010).

2. **The study design necessitates selecting schools from outside of the target population.** Generalizations to target populations to which the sample does not belong are referred to as *synthetic generalizations* and require stronger assumptions. For example, Tipton et al. (2016) considered two impact evaluations that each defined the target population to include schools that were already implementing the intervention, but the study design required a sample of schools that were *not* already implementing it. Specifically, each study was a randomized controlled trial designed to compare schools assigned to the intervention to other schools assigned to a “business-as-usual” control group. This design requires a sample of schools that were not already implementing the intervention—that is, for which “business-as-usual” does not involve the intervention—to permit a comparison between the intervention and business-as usual. In these instances, begin by creating a population frame for the target population (only including all schools already implementing the program). Then create another population frame that includes schools that are *not* in the target population (all schools *not* currently implementing the program). For example, Tipton and colleagues partnered with the intervention developers to identify schools already using the program (based on sales data), then removed these schools from CCD data to identify schools that were not already using the program. Finally, use propensity score methods to reweight the second population frame to be more similar to the first. Alternately, use these methods to stratify the schools in the second population frame to support sampling and recruitment—in particular, (1) define strata in the first population frame; (2) set sample size targets for each stratum based on the share of schools from the first frame that fall into each stratum; (3) sort schools from the *second* population frame into those strata; and (4) select schools from the second population frame to satisfy the sample size targets set in step (2).
3. **There are multiple target populations**. Given the expense of impact evaluations, it may be more cost effective to design a single study to estimate the average treatment effect for multiple target populations. One example is to estimate separate average treatment effects for each of the 50 states. When there are multiple target populations, the optimal sampling method for one population may be at odds with what is best for one or more of the other populations. Tipton (2022) provides a discussion of this problem and compromise approaches to selecting the sample.
4. **There are multiple goals for the study.** This guide focuses on sampling methods for estimation of the population average treatment effect. Other goals might include estimating subgroup treatment effects or testing hypotheses regarding moderators of treatment effects. The optimal sampling method for estimating the average treatment effect, however, is not always optimal for estimating these other parameters. Tipton (2021)

provides a compromise approach that can increase statistical power for testing moderators. Furthermore, Tipton et al. (2019) provide an overview of a sampling approach when the goal is to estimate both the average effect and subgroup effects, which requires oversampling small subgroups.

If you are interested in these advanced methods, we advise you to partner with a sampling statistician.

Recommendation 4. Implement the Sampling Plan

Goal: To recruit a sample of schools that represents the target population

Steps:

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.’

Resources:

- Population frame data with strata indicators and contact information for each school
- Spreadsheet for tracking recruitment from *The Generalizer* (www.thegeneralizer.org; Tipton & Miller, 2021)

What to Report:

1. The total number of schools contacted and the total number that agreed to participate
2. The most common reasons reported by schools for not participating
3. Of those contacted, a comparison of differences between those schools that agreed to be in the study and those that did not

While designing a sampling plan has its challenges, implementing the plan–successfully recruiting schools into the study–is more challenging but critical to the success of impact evaluations in education. Districts may refuse to allow researchers access to their schools, or schools may decide that they are not interested in participating. Participating in impact evaluations typically requires time from busy staff. It may also require testing educational interventions that schools are not interested in implementing. Since participation in most studies is voluntary, districts and schools must be persuaded to participate. How to obtain voluntary study cooperation goes beyond the scope of this guide. However, to obtain high cooperation rates, study teams need to consider ways to make study participation worthwhile for participating districts and schools.

To successfully recruit schools, start by developing a recruitment plan that includes adequate staff, training, and resources. The recruitment plan should also address how to handle schools that are found to be ineligible for the study during the recruitment process (e.g., if a school falls outside of the target population).

How to carry out the recommendation

1. Budget time and resources for recruitment as early as possible

To prepare for recruiting the sample, develop a plan and a budget to support the effort. The plan may need to cover the following:

- **A planning year.** The first year of a study might focus on preparing materials and then recruiting schools. This planning year allows time for the development of relationships with new schools that might otherwise not be included. It also allows time for submitting research applications that districts may require, obtaining approval from the responsible Institutional Review Board(s) and, for studies conducted for federal agencies, obtaining approval from the Office of Management and Budget under the Paperwork Reduction Act.
- **Multiple cohorts of recruitment.** Because it can be difficult to recruit a large number of schools in one year, consider planning to spread the recruitment process over multiple years. This results in a multi-cohort design, which is not uncommon in impact evaluations.
- **School- and teacher-specific incentives.** Typically, studies include incentives for participation in a study. These often include financial incentives–for the school, the teacher, or students–as well as access to programs (for example, free curriculum); desirable professional development; and additional classroom support. To encourage participation in randomized trials, consider offering the program to the control group for free after outcome data are collected for the study. Also consider giving larger incentives to schools from strata in which recruitment is more difficult.
- **Relationships with partner organizations.** For example, these might be school districts as well as organizations focused on teacher professional development or networks. By partnering with these organizations, you may find they help to recruit schools that would not otherwise agree to participate.
- **Support from the evaluation's funder.** The funder may be able to provide a letter of support encouraging participation by districts and schools. In some cases, the funder may even be willing to contact school districts to encourage their assistance and cooperation.
- **Additional recruiters.** Given how difficult recruitment can be, multiple recruiters will likely be necessary. In the next section, we provide more details on putting together and managing a team.

2. Build and manage a recruitment team

Once a study is funded and designed, the task of recruitment begins. Successful recruitment involves a team with the following:

- **Recruiters who visit schools.** Hire or select recruiters with the experience and skills to negotiate effectively with school decision-makers. These recruiters might include former teachers, former school leaders, or those with survey experience. Recruiters need to be able to think on their feet, develop relationships, and advocate for the study.

- **A recruitment lead who oversees recruiters and timelines.** The team lead oversees the implementation of the sampling plan (including stratum goals) by recruiters in the field. Ideally, this lead has a deep understanding of the intervention under study and the target population, as well as prior experience recruiting schools.

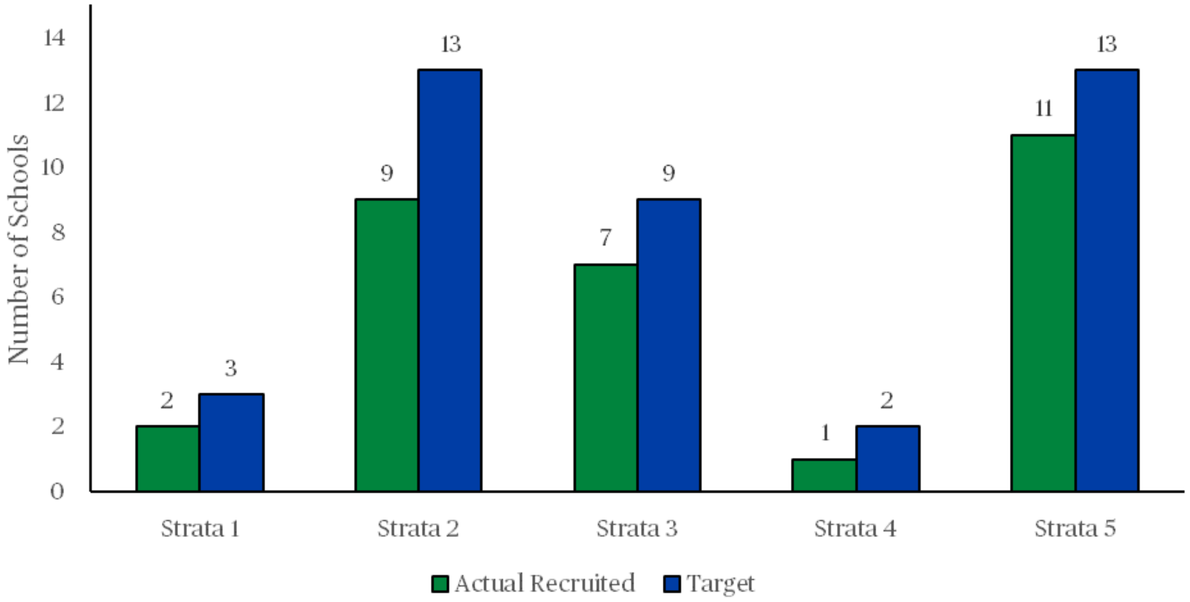
Once hired, teams have five important activities:

1. **Develop recruitment materials.** These include materials for training recruiters, recruitment scripts and protocols for in-person visits, brochures, and materials to provide to contacted schools, and materials for parents and consenting students.
2. **Train and practice before going into the field**. The lead should develop a training process that provides recruiters with a deep understanding of the intervention and study procedures as well as of the need for sample representativeness. Furthermore, the training should provide opportunities for recruiters to practice thinking on their feet when they encounter different hurdles to participation. Ideally, this training is supported by well-developed training materials, protocols for recruitment, and study materials to be provided to schools.
3. **Manage recruitment lists.** In general, recruiters should try hard to recruit the top ranked schools within a stratum before proceeding to new schools. Although a stratum might include more than 1,000 possible schools, it is not ideal to provide recruiters with this entire list. Instead, the recruitment lead should initially provide recruiters with only the top-ranked schools in each stratum to focus their attention on those schools and increase the chances of successfully recruiting a set of schools that is very similar to the target population. For example, the lead should consider providing a recruiter for only the first five ranked schools for the stratum. If the recruiter does not meet the recruitment goal for the stratum, then the lead provides another five schools. This process encourages recruiters to focus on gaining cooperation from the schools that the research team would most like to include in the evaluation and not just on obtaining cooperation from an adequate number of schools.
   - **Continuously monitor recruitment**. As schools agree to be in the study, recruiters should maintain information in real time to assess progress toward recruitment targets (per stratum and overall). In particular, this monitoring might include recruitment targets for each stratum as well as the total sample that has agreed to be in the study within each stratum. See Figure 2 for an example. By periodically examining this information the team lead can determine how well the recruitment efforts are going as well as address any problems that arise. For example, monitoring may reveal that recruitment is moving very slowly in one stratum. In response, the team lead could reallocate other recruiters to this stratum, including the most seasoned recruiters; offer additional financial incentives for participation; and consider other ways of

making participation attractive to schools in this stratum (e.g., offering to give briefings to district staff).

4. **Encourage continuous and open communication**. Recruitment is difficult, and the team will often need to strategize together. Once recruitment begins, it may become clear, for example, that recruitment is more difficult in some strata than others, and you may need to update strategies, resources, and incentives based upon this information.

Figure 2. Actual recruited to date versus target number of schools



3. *Screen out schools that are ineligible for the study*

In many evaluations, the population frame may not include all of the variables needed to determine whether schools are eligible for the study, or these variables may include missing values or errors for some schools in the population frame. In these cases, you may find that some of the schools recruited to participate are ineligible for the study–that is, they fall outside of the study's target population. For example, perhaps the contacted school is already implementing the intervention under study. Or, perhaps, the intervention requires school resources that are not available, such as a library, a school nurse, a guidance counselor. In these instances:

- **Exclude the school from the study.** Schools that are not part of the target population should not be part of the study sample.

- **Document that the school is ineligible (and why).** This step is important later when assessing the similarity between the sample and target population (see Recommendation 5)–*excluding ineligible schools*–and making statistical corrections for differences between the sample and the target population (see Recommendation 6).

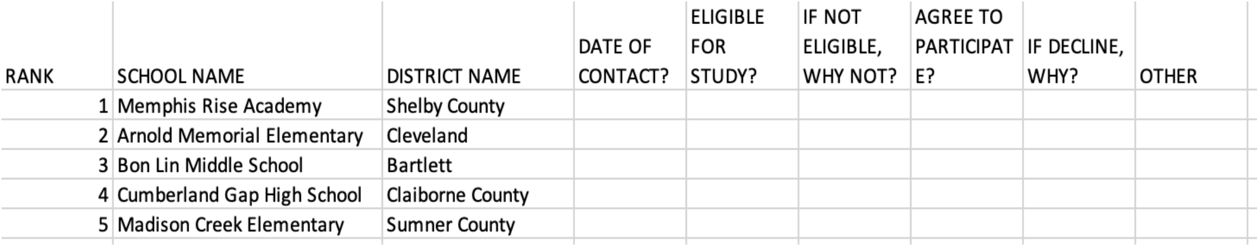
4. *Collect and report data on 'volunteers' and 'decliners'*

When recruited to participate in an evaluation, some schools will accept and volunteer to participate, but other schools will decline and opt out of the study. Declining to participate is akin to 'non-response' in survey research, which raises concerns that volunteer schools may differ in important ways–observed or unobserved–from schools that decline to participate. In these instances:

- **Document when schools decline to participate in the study.** This record will help recruiters focus on schools that have not yet decided whether to participate; it also facilitates later reporting of the "take up" rate–the percentage of recruited schools that agreed to participate (much like the response rate on a survey).
- **Compare volunteers to decliners using data on all potential moderators from the population frame.** Substantial differences would suggest self-selection or nonparticipation bias that should be addressed at a later stage (see Recommendation 6).

To document the recruitment effort, recruiters should maintain a database regarding the schools that were contacted, if they were eligible for the study (including why not), if they agreed to be in the study, and if not, information given regarding why. *The Generalizer* provides automatically a .csv file that can be used to track these reasons for each school in the population frame; an example of such is provided in Figure 3. We recommend identifying a small set of likely reasons for nonparticipation and asking nonparticipating schools if those reasons apply. This will allow you to identify and report schools' most common reasons for declining to participate. At the end of the study, these data can be analyzed to assess the differences between eligible and ineligible schools, and differences between schools that agree to participate and those that decline, and to explore the implications of these differences for generalization (Recommendation 5).

Figure 3. Recruitment tracking spreadsheet (exported from *The Generalizer*)



RANK	SCHOOL NAME	DISTRICT NAME	DATE OF CONTACT?	ELIGIBLE FOR STUDY?	IF NOT ELIGIBLE, WHY NOT?	AGREE TO PARTICIPAT E?	IF DECLINE, WHY?	OTHER
1	Memphis Rise Academy	Shelby County						
2	Arnold Memorial Elementary	Cleveland						
3	Bon Lin Middle School	Bartlett						
4	Cumberland Gap High School	Claiborne County						
5	Madison Creek Elementary	Sumner County						

Example

The team

While the sampling plan was being developed, the PI began hiring and training a team of recruiters. This team consisted of three recruiters; one recruiter was in Georgia, near the PI; another in Tennessee; and the third in Florida. In addition, one of the PI's graduate students, who had previously been an elementary school teacher, took on the recruitment team lead position.

Recruitment

A total of 120 schools were recruited to take part in the study. Ultimately, 40 of these agreed to participate in the study, including the 10 schools that committed to participate before the study began. Eighty of the 120 schools did not ultimately take part in the study:

- 10 schools responded but were found to be ineligible for the study.
- 10 schools did not respond to repeated attempts to contact them and thus implicitly declined to participate, but were assumed to be eligible for lack of evidence to the contrary.
- 60 responded and were found to be eligible for the study, but refused to participate and provided a reason for declining.

The most common reason given for not participating was recent turnover in leadership (31 percent) or that the school was already invested in its current reading program (29 percent). The remaining reasons varied, including concerns with staffing and the recent implementation of other new programs.

Recruitment was more difficult in Strata 2 and 5—which included more rural schools—than in the other strata; in Stratum 2, only 6 schools joined the study (versus the 8 required), and in Stratum 5, only 9 joined as well (versus the 13 required). To compensate for these differences, because recruitment was a bit easier in Strata 3 and 4, researchers recruited slightly larger samples there (7 versus 4 and 5 versus 2, respectively). In Stratum 1, researchers recruited the required three schools.

Comparison of volunteers versus decliners

Among the 110 eligible schools, the 40 schools that agreed to participate in the study were compared to the 70 schools that declined by calculating standardized mean differences (SMDs) between the population frame and the sample for each potential moderator in the population frame. The SMDs ranged from -0.72 for the percentage of students with a home language other than English to 0.64 for the number of students enrolled per school (Table 5).

In addition, recruiters asked schools for information on their current reading program. In some cases, the reading program was a single, comprehensive curriculum, whereas in others it included a variety of components from different sources. This information was collected for eligible schools that volunteered to be in the study and, to the extent possible, for those that declined. Results indicate that schools declining to be in the study were more commonly using one of two well-established reading programs than those that volunteered (Table 6). This suggests that differences between the business-as-usual program should be adjusted for in the analyses (see Recommendation 6, Advanced Methods, Section 1 for more information).

Table 5. Comparison between eligible schools joining and declining participation in the study

Potential moderator	Volunteers mean (*n* = 40)	Decliners mean (*n* = 70)	Pooled standard deviation	Standardized mean difference
Students eligible for free or reduced-price lunch (%)	86.3	98.0	23.9	0.49
Female students (%)	48.5	48.6	2.0	0.05
Black students (%)	56.0	64.5	20.2	0.42
Hispanic students (%)	21.9	17.6	16.4	-0.26
Native American students (%)	0.2	0.3	3.2	-0.03
English language learners (%)	7.0	5.2	4.1	-0.44
Home language other than English (%)	12.0	3.7	11.5	-0.72
Urban school (%)	42.2	47.8	37.5	0.15
Suburban school (%)	41.0	48.5	35.5	0.21
Town school (%)	8.6	6.7	22.7	-0.08
Rural school (%)	10.4	2.0	28.0	-0.30
Number of students per school	797.1	738.3	163.4	0.64
Number of schools per district	221.0	145.1	90.4	0.16

Note: The table excludes 10 schools that were sampled but found to be ineligible during the recruitment process. The standardized mean difference is the difference in means between volunteers and decliners divided by the pooled standard deviation.

Table 6. Current reading program (business-as-usual) at eligible schools

Potential moderator	Percentage of volunteers (*n* = 40)	Percentage of decliners (*n* = 70)
Reading program 1	10.0	7.1
Reading program 2	12.5	17.1
No specific program	60.0	42.9
Combination of programs	17.5	5.7
Reading program being evaluated	0	8.6
Information not available	0	18.6

Note: The table excludes the 10 schools that were sampled but found to be ineligible during the recruitment process.

Recommendation 5. Assess the Similarity Between the Sample and the Target Population

Goal: To determine the degree of similarity—and identify possible differences—between the sample and target population

Steps:

1. Consider limitations in the population frame before conducting the analysis.
2. Compare the sample and target population using data from the population frame.
3. Explore the potential threats to generalizability from unobserved moderators.

Resources:

- Data from population frame on potential moderators
- *The Generalizer*, a free web tool (www.thegeneralizer.org; Tipton & Miller, 2021)
- generalize, an R package (https://nustat.github.io/generalizeR; Ruel et al., 2022)

What to report:

1. Comparisons between the sample and target population (as operationalized in the population frame) in terms of potential moderators (for example, standardized mean differences)
2. A measure of overall similarity (such as the generalizability index)

Once the sample has been identified, compare schools in the sample to schools the target population using data on potential moderators from the population frame. The goal is to identify large differences between the potential moderators in the sample and target population that suggest that the average treatment effect estimated in the sample may not generalize the average treatment effect in the target population. For this reason, report any differences in observed variables that may moderate the intervention's impact (see Recommendation 7) and consider and acknowledge potential differences in unmeasured moderators. Identifying differences in measured moderators can prompt your study team to make statistical adjustments that may improve the generalizability to the target population (see Recommendation 6).

How to carry out the recommendation

1. *Consider limitations in the population frame before conducting the analysis*

During recruitment, you may have found and excluded schools that are not eligible for the study (e.g., because the data necessary to screen them out from the start were unavailable). If so, the population frame may include other ineligible schools that cannot be identified. If ineligible schools have different characteristics than eligible schools, the population frame data may not accurately represent the characteristics of the target population, and the comparisons proposed in this section between the sample and population frame may either overstate or understate the true differences between the sample and target population.

To assess this potential problem, compare the characteristics of eligible schools to the characteristics of ineligible schools using data from the population frame on schools recruited to participate. In particular, produce a table like the one used to compare schools that agreed to participate to those that did not (see Table 5 in the previous section). If eligible and ineligible schools are similar to each other in the potential moderators, or the share of recruited schools found to be ineligible is very small, it would suggest that the inclusion of ineligible schools in the population frame is inconsequential for describing the characteristics of the target population, and you can proceed to compare the sample to the population frame without any statistical adjustments. However, if the two groups differ substantially on any potential moderators, and the share of recruited schools found to be ineligible is nontrivial, see Recommendation 6, *Advanced techniques,* for guidance.

2. *Compare the sample and target population using data from the population frame*

Recall that Recommendation 1 was to identify the target population—that is, the types of students about which the study aims to learn and the types of schools to which the study hopes to generalize—and Recommendation 2 was to develop a population frame that contains the target population. For students and schools, as well as the school districts in which they are located, compare the sample of schools that agreed to participate in the study to the target population using data on potential moderators included in the population frame (see Recommendation 2, Step 2). In particular:

- **Calculate and report standardized mean differences (SMDs) for each potential moderator.** Standardize each moderator included in the population frame using the population frame standard deviation across strata. Then calculate the SMD by taking the difference in means of the standardized variable between the sample and the population frame. The SMD estimates the degree of similarity between the sample and target population. Calculating SMDs will allow you to identify the potential moderators for which the differences are the largest and to identify the subset of variables for which statistical adjustments (Rubin, 1990) may be required (see Recommendation 6).

- **Calculate and report the generalizability index.** Calculate the generalizability index using data on each moderator from the population frame. This index was proposed by Tipton (2014a) as a global measure of similarity between the sample and target population. The index ranges from 0 to 1, with higher values indicating greater similarity in the distributions of potential moderators between the sample and population frame. Values can be interpreted by multiplying by 100; for example, if the index equals 0.80, the schools in this study are 80 percent similar to the population frame on the potential moderators. For studies with about 40 schools, index values greater than 0.90 roughly correspond to the same degree of similarity that would be expected in a random sample of the same size, indicating that no statistical adjustments are required.[3] Index values below this indicate that some statistical adjustments will be required in order to estimate the population average treatment effect well, and index values below 0.50 indicate samples that are so different that inferences to the target population are not warranted (see Recommendation 6).

To implement the steps above, you must be able to identify the schools that belong to the target population in the population frame. However, in many cases, the population frame may include some schools that fall outside the target population for reasons discussed in earlier sections (e.g., when data for one or more eligibility criteria are not available). *Advanced techniques* at the end of this section explains how to compare the sample to the target population under these scenarios.

3. *Explore the potential threats to generalizability from unobserved moderators*

The methods described in the previous section apply when moderators are observed for both the sample and the target population. But some moderators may not be observed for either, and others may be observed only for the sample because they were collected as part of the study.

Importantly, the degree of similarity between the sample and population frame on observed variables may overstate the similarity overall, including both observed and unobserved variables. If the sample differs from the population frame on observed variables, it likely also differs on a host of unobserved variables. Furthermore, we should expect larger differences

[3] Tipton et al. (2017) provide an approximate sampling distribution for the generalizability index. This indicates that when there are $p = 20$ covariates, for $n = (40,60,80,100,120)$ this upper threshold should be (.90,.94,.95,.96,.97). If $p = 10$, this threshold should be slightly higher, (.94,.96,.97,.98,.98). The theory is based on tests regarding how likely a large difference is in a random sample and makes assumptions regarding the covariate distributions. For this reason, it is more reasonable to treat this threshold as a ‘rule of thumb’ than as a hypothesis test.

on unobserved variables than observed variables if observed variables were used to stratify the sample (see Recommendation 3). For this reason, after examining observed variables, you should consider unobserved variables that may threaten the generalizability of the study findings. While the role of unobserved variables cannot be assessed directly, the following steps may be useful:

- **Calculate and assess the opt out rate among schools selected and recruited to participate.** The "opt out" rate means the proportion of eligible schools contacted to be in the study that declined to participate. In addition to this rate, provide evidence collected from recruiters regarding the reasons that schools declined participation (Recommendation 4) since they may influence the generalizability of the study findings. For example, if many schools declined to participate because they reported insufficient resources to implement the intervention well, it may suggest that schools in the study have greater resources–and could expect larger impacts–than the average school from the population frame.
- **Summarize and report on moderators that are observed for the sample but not observed for the population.** Impact studies often collect data on the students, teachers, and schools that participate in the study, including characteristics that may moderate the impact of the intervention. Reporting information on these potential moderators may support informed conjectures about possible differences between the sample and the population; they can definitely support generalizations to other populations for which those characteristics are observed.
- **If possible, test for the presence of unobserved moderators.** When assignment to the intervention is within schools, you may directly estimate the variation in treatment impacts across schools–for example, using models with fixed, school-specific intercepts and random treatment effects that follow a normal distribution (see Bloom et al., 2017). These models can be used to calculate the proportion of the cross-school impact variation that are explained by observed moderators. If observed moderators explain only a small share of this variation, then unobserved moderators are likely responsible for much of the variation in impacts across schools.
- **Treat the generalizability index for observed moderators as an upper bound on the similarity between the sample and population frame on unobserved moderators.** The balanced sampling methods described in the previous section, which make use of observed moderators in the population frame, are likely to yield a sample that is more similar to the population for observed moderators than for unobserved moderators. Therefore, the generalizability index constructed using observed moderators will likely overstate the similarity between the sample and population frame on unobserved moderators.

Example

Of the 120 schools recruited for the study, the 10 schools found to be ineligible were found to be similar to the 110 eligible schools on observed moderators: the standardized mean difference for each moderator was less than 0.25 (not shown here).[4] Combined with the small share of schools found to be ineligible, this evidence suggest that the population frame is adequate for describing the target population and assessing the generalizability of the sample.

The 40 schools that agreed to participate in the study were compared to the population frame by (1) calculating SMDs for each potential moderator in the population frame and (2) calculating the generalizability index. The standardized mean differences between the population frame and the sample ranged from -0.50 for the percentage of students with a home language other than English to 0.34 for the percentage of students eligible for free or reduced-price lunch (Table 7). The generalizability index was calculated to be 0.82, suggesting that some statistical adjustment is needed to support generalization (see Recommendation 6).

Additional information on schools that declined participation suggests that the business-as-usual reading program used in the comparison condition varied from what was common in the sample schools. These comparisons (found in Table 6) indicate that declining schools were more likely to be using one of two well-established programs. This difference between participating and declining schools suggests that there may be unobserved differences between the sample and the population frame. Best practice would be to adjust for this moderator using additional weights (see Recommendation 6, *Advanced techniques, Adjust for volunteer bias when additional moderators are available in the selected sample only*). The research team should be clear in these limitations when they report findings.

[4] When standardized mean differences are smaller than 0.25, regression can be used to adjust for any residual differences without threat of extrapolation (Rubin, 2001).

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.

Advanced techniques

In the best case, you can assess the generalizability of the sample to the target population using the data assembled in developing the population frame. However, if some of the recruited schools are found to be ineligible during the recruitment process, as described earlier (see Recommendation 4), presumably other schools in the population frame that were *not* selected for recruitment would have also been found to be ineligible. In this scenario, it will not be possible to restrict the population frame to just those schools that are in the target population, so it will not be possible to assess the generalizability of the sample to the full target population as described earlier in this section.

However, in some cases, it may be possible to assess the generalizability of the sample to a representative subset of the target population—those that were selected for recruitment. Suppose that you selected a representative sample of schools from the population frame, following Recommendation 3. Further, suppose you were able to contact each of these schools during the recruitment process—and, importantly, were able to assess their eligibility during recruitment. Under these circumstances, calculate SMDs and the generalizability index of the schools participating in the study *relative to schools selected for the sample and found to be eligible*. These estimates will approximate the SMDs and generalizability index relative to all schools in the true target population as long as sampled schools were selected to be representative of the schools in the population frame.

However, in many studies, the analysis described above is complicated by the fact that some recruited schools do not respond at all to the study’s recruiters: their eligibility for the study is unknown. If the share of recruited schools with unknown eligibility is relatively modest, you may want to consider simply imputing their eligibility by applying standard imputation techniques to the data assembled for the population frame. But if this share is large, it may be difficult to produce convincing evidence of the sample’s generalizability.

Recommendation 6. Adjust for Differences Between the Sample and the Target Population

Goal: Estimate the average treatment effect for the target population, adjusting for any differences between the sample and population on potential moderators

Steps:

1. Use post-stratification when the sample is fairly but not perfectly similar to the target population (e.g., 0.50 < generalizability index < 0.90).
2. Redefine the target population when the sample is not similar to the population frame (generalizability index < 0.50). Consider restricting inferences to the sample only.

Resources:

- Data from population frame on potential moderators
- generalize, an R package (https://nustat.github.io/generalizeR; Ruel et al., 2022)

What to report:

1. The final estimation method implemented (for example, post-stratification) and the extent to which this method increases similarity between the sample and target population
2. The final target population for which generalization is possible
3. The estimate of the average treatment effect for the target population as well as its standard error
4. Assumptions required for the estimation method to result in an unbiased estimate

The results of the analyses conducted in Recommendation 5 might indicate that the study sample differs from the population frame in terms of measured and unmeasured moderators. When this occurs, the impact in the sample will differ from the impact in the population frame. When these differences are with respect to observed moderators, you should make and report statistical adjustments to address those differences. These adjustments will typically reduce the bias when generalizing from the sample to the population frame. They will also typically reduce the precision of the estimates (that is, increase the standard errors), as mentioned earlier.

Note that although many of the methods covered in this section should be accessible to a wide range of education researchers, implementing the methods presented in the *Advanced Methods* section will benefit from advanced statistical expertise.

How to carry out the recommendation

Recommendation 5 described how to calculate a generalizability index that summarizes the degree of similarity between the sample and population frame. This index value offers a guide regarding if adjustments are required and the types of adjustments that are useful. Here, our recommendations for statistical adjustments depend on the value of this index value (see Tipton, 2014a; Tipton et al., 2017 for the development of these). Recall that when the index is greater than about 0.90, no adjustment to improve generalizability is recommended because the sample and population frame are already similar to one another on the potential moderators.[5]

1. *Use post-stratification when the sample is fairly but not perfectly similar to the population frame (0.50 < index < 0.90)*

The most common situation is one in which there is a mismatch between the stratum allocations in the sample of schools that agreed to participate in the study and population frame. Post-stratification can be used to adjust the sample for these differences: schools in underrepresented strata are up-weighted and schools in overrepresented strata are down-weighted. In particular, for each stratum, schools are assigned a weight equal to the number of schools in the population frame from the stratum divided by the number of schools in the sample from that stratum. These weights can be used in estimating the average treatment effect—for example, with the 'generalize' (Ruel et al., 2022) package in R or with survey analysis procedures (for example, SVY in Stata, PROC SURVEYREG in SAS).

The benefit of post-stratification is that it typically reduces bias in the impact estimate. Before conducting this analysis, recalculate SMDs and the generalizability index for the weighted sample. Ideally, the degree of similarity between the weighted sample and population frame should be much improved with the weights. If the weighting adjustment does not result in adequate similarity (e.g., index above 0.90) between the reweighted sample and population frame, you may use more advanced methods (see the *Advanced techniques* at the end of this section).

It is important to realize that post-stratification often reduces precision. This reduction in precision is proportional to the generalizability index (Tipton, 2014a). We recommend using post-stratification, despite the reduction in precision, when the value of the index is between 0.50 and 0.90, to produce impact estimates that better generalize to the population frame.

[5] Although no adjustment is necessary for purposes of generalizability, adjustments for differences between the treatment and control groups may be necessary or may be useful for improving the precision of the impact estimate.

2. *Redefine the target population when the sample is not similar to the population frame (index < 0.50)*

If school recruitment is particularly difficult for some types of schools, the generalizability index may fall below 0.50 and the SMDs may be very large for one or more potential moderators. Large differences between the sample and population frame typically indicate that there is undercoverage–a portion of the population (perhaps an entire stratum) that is simply not represented at all in the study. Undercoverage can result when the study team was unable to recruit any schools from one of the strata. Although it may seem that you could solve this problem by collapsing empty strata into other strata, doing so would rely on the untestable assumption that the impact is the same in the empty stratum as in the stratum with which it was combined. When the generalizability index is below 0.50, generalizing from the sample to the population frame is not possible without extrapolation.

In this scenario, we recommend proceeding using the following steps:

1. **Acknowledge these differences in any reports or publications.** In particular, report the generalizability index, identify the individual moderators for which the sample is very different from the population frame (SMDs greater than 0.25; see Rubin, 2001), and indicate that generalizing to the intended population frame may not be feasible.
2. **Consider narrowing the target population.** For example, if most of the large suburban schools in the population were in Stratum 5, and no schools from this stratum agreed to participate in the study, consider using propensity score methods to exclude schools from the population that fall outside of the "region of common support"–the values of potential moderators covered by schools in the sample. For more details on the use of propensity score methods for this purpose, see *Advanced techniques* at the end of this section. Although narrowing the target population narrows the inferences that can be made from the study findings, broader inferences would require out-of-sample extrapolations that are hard to justify.
3. **Assess the similarity between the sample and the narrower target population.** First exclude schools that do not belong to the narrower target population from the population frame. Then recalculate the generalizability index. If the index is greater than 0.50, the treatment effect for the narrower population can be estimated well with poststratification adjustments. However, if this is not the case, you may use more advanced methods (found later in this section).

3. *Consider restricting inferences to the sample only*

This guide focuses on methods for designing, conducting, and estimating impact estimates for a target population. These methods require adequate data and assumptions regarding the potential moderators. In some cases, however, you may feel these assumptions are untenable; for example, when recruitment is very difficult, you may suspect that there are large unobserved differences between schools that volunteered to be in the study and those that declined. In these cases, you should consider whether it is reasonable to estimate an impact for the target population or whether, instead, the study should only focus on the impact estimate in the sample. If the latter, your study reports should be very clear that the results only apply to the sample included in the study and that generalizations beyond this sample are not supported statistically.

Example

Because the generalizability index (0.82) is in the medium range–greater than 0.50 but less than 0.90– the team began by checking to see if a post-stratification adjustment would be effective at reducing these differences. This involved the following steps.

1. Develop post-stratification weights

The first step was to calculate post-stratification weights that up-weight schools in strata that are underrepresented in the sample and down-weight schools in strata that are overrepresented in the sample. The team constructed raw and normalized weights (Table 8). The raw weights sum to the total number of schools in the population frame, and the normalized weights sum to the number of strata (five), so the simple average of these weights across the five strata is 1.

Table 8. Stratum weights for analysis

	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
Population	285	**1,307**	947	203	1,407
Sample (*n* = 40 schools)	3	11	12	5	9
Raw weight	285/3 = 95.00	1,307/11 = 118.82	947/12 = 78.92	203/5 = 40.60	14,07/9=156.33
Sum of raw weights	489.67	489.67	489.67	489.67	489.67
Normalized weight[a]	5×95.00/489.67 = 0.97	5×118.82/489.67 = 1.21	5×78.92/489.67 = 0.81	5×40.60/489.67 = 0.41	5×156.33/489.67 = 1.60

[a] The normalized weight equals the stratum weight divided by the simple average of the five stratum weights.

2. Evaluate if post-stratification weighting increases similarity

Next, the team applied these weights to each of the potential moderators. To do so, letting X_{ijk} be the *k*th moderator value for school *i* in stratum *j*, they calculated the reweighted mean $\bar{X}_{wk}$ using

$$\bar{X}_{wk} = \frac{1}{40} \sum_j \sum_i w_{ij} X_{ijk}$$

Where W_{ij} are the normalized weights from Table 8. Next, the team calculated the adjusted SMD for each moderator by dividing this reweighted value by the population standard deviation for the same moderator (Table 9).

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
Students eligible for free or reduced-price lunch (%)	78.2	80.6	23.9	0.10
Female students (%)	48.4	48.5	2.0	0.05
Black students (%)	50.2	52.0	20.2	0.09
Hispanic students (%)	24.8	25.2	16.4	0.02
Native American students (%)	0.6	0.5	3.2	-0.03
English language learners (%)	8.3	8.0	4.1	-0.07
Home language other than English (%)	17.7	17.4	11.5	-0.03
Urban school (%)	38.1	39.9	37.5	0.05
Suburban school (%)	35.6	35.7	35.5	0.00
Town school (%)	10.0	9.2	22.7	-0.04
Rural school (%)	16.3	15.5	28	-0.03
Number of students per school	562.1	579.8	163.4	0.11
Number of schools per district	120.4	127.9	90.4	0.08

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

Overall, this simple post-stratification adjustment appeared to be effective in increasing similarity between the sample and target population. After adjustment, the largest absolute standardized mean difference is 0.11 (compared to 0.50 prior to adjustment), and only two of the potential moderators have values above 0.10 (compared to 11 prior to adjustment). This improvement can also be found in the reweighted generalizability index, which is now 0.93.

3. Use weights to estimate the population average treatment effect and standard errors

Because the previous analysis indicated that the weighting adjustment increased the similarity between the sample and target population, the team used the same adjustment method to estimate the population average treatment effect and its precision. The weights were incorporated into the analysis using the weight() function in the 'generalize' package in R. This produced an estimated average treatment effect in the target population of 0.26 (SE = 0.12). In comparison, the sample average treatment effect (unweighted) was 0.21 (SE = 0.10). In this case—but not always—the average treatment effect in the target population is estimated to be larger than in the sample alone. The smaller impact estimate for the sample than the target population suggests that without post-stratification, the impact estimates produced by the study would be biased downward. However, because adjustment was necessary, the estimate of the population impact has a standard error about 20 percent larger than the estimate for the sample impact.

Advanced techniques

Additional methods for addressing differences between the study sample and the target population are required under four scenarios. The first involves adjusting for differences between the schools that ultimately agreed to participate in the study and all the schools selected and invited to participate. These adjustments focus on the situation in which additional moderators were collected from all schools that were invited to be in the study (but that were unavailable in the population frame). The second involves adjusting the description of the target population when some schools that were invited to participate were determined to be ineligible to be in the study (when this eligibility data was unavailable in the population frame). The third and fourth are when the post-stratification approach described previously does not result in a representative sample. This could be because participating schools are very different from the schools in the target population. One of these approaches involves the use of propensity scores, while the other uses regression and related models.

1. *Adjust for volunteer bias when additional moderators are available in the selected sample only*

It is unlikely that every school selected and recruited will volunteer to participate in the study. In general, adjustments between the final sample and target population can be conducted as described previously. In some cases, however, during recruitment, additional information regarding potential moderators may also have been collected, both for those that volunteered and those that declined. In the example, the research team asked all schools they contacted which reading program they were currently using, and analyses comparing the volunteers and decliners indicated that the business-as-usual program differed across these groups. The general situation is described in Figure 4, which we explain below.

Incorporating these additional potential moderators into the analysis involves two steps. First, the full set of selected and recruited schools is compared to the population frame (Recommendation 2) on the original set of potential moderators (Recommendation 1). Post-stratification (Recommendation 6) is then used to create 'Sample Selection' weights that could be used in analyses (this is adjustment (a) in Figure 4). Second, the final sample of volunteers is compared to the sample selected and recruited using all of the original potential moderators *and* this new set of additional moderators collected during recruitment (for example, the business-as-usual reading program). This comparison is best conducted using propensity score methods (see *Advanced techniques,* Section 3 at the end of this section), which produce weights for each of the volunteering schools; we call these the ‘Participation’ weights (see adjustment (c) in Figure 4). To get the final weights used for analysis, multiply these weights together (Total Weight = Sample Selection Weight × Participation Weight).

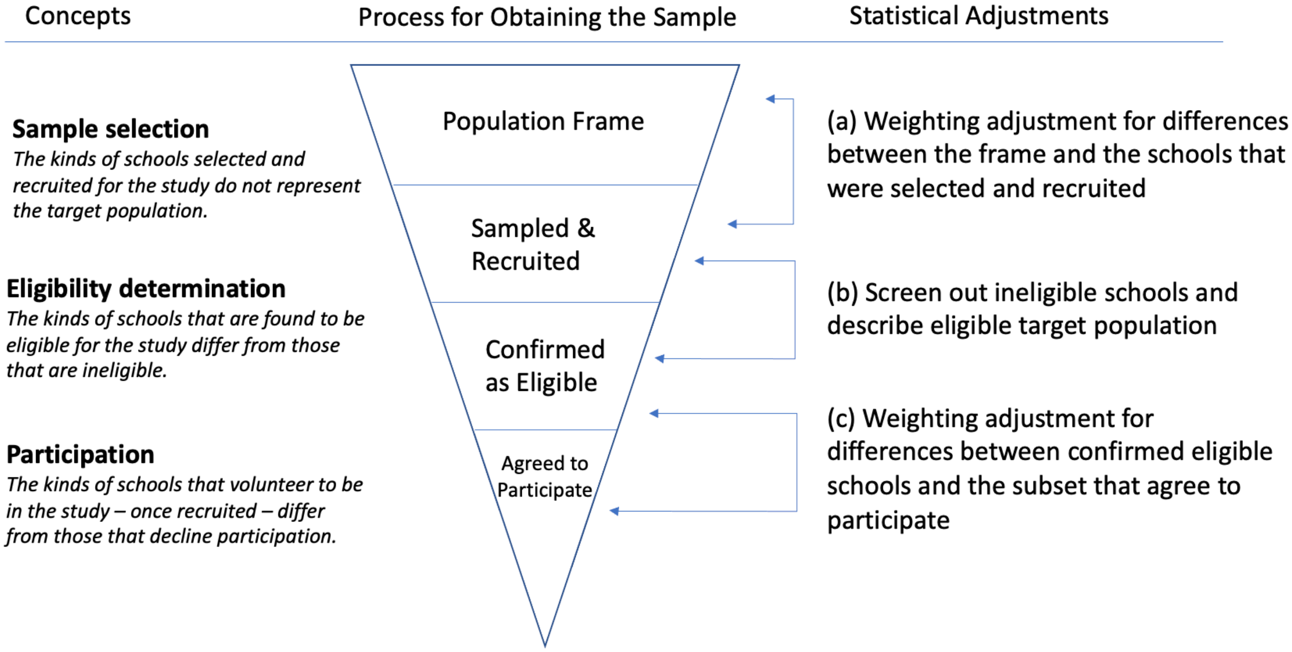
2. Adjust the population when some schools are ineligible for the study

Sometimes the population frame does not include all of the information required to determine eligibility for the study. For example, to be in the study, schools may need to be using a specific business-as-usual program or may need particular resources (for example, a school nurse). Certainly, during recruitment schools that do not meet eligibility criteria can be screened out. However, the fact that some of the schools contacted were ineligible suggests that other, non-contacted schools in the population frame may also be ineligible. This means that the description of the target population provided based upon the population frame may not be accurate.

To address this challenge, again there are two stages of adjustments. First, as before, compare the selected and recruited sample to the population frame and adjust for any differences; this results in ‘Sample Selection’ weights (see adjustment (a) in Figure 4). Second, screen out the ineligible schools from those that were selected and recruited (see adjustment (b) in Figure 4). This is akin to assigning a weight of one to the eligible schools and a weight of zero to the ineligible schools; we will call this the ‘Eligibility Determination’ weight. Now, multiply these two weights together to get the final weights for analysis (Total Weight = Sample Selection Weight × Eligibility Determination Weight). Lastly, use Total Weights to calculate the weighted mean and weighted standard deviation for each of the potential moderators for the selected and recruited schools. These new means and standard deviations can be used to describe the target population, which now accurately captures characteristics of the previous inclusion/exclusion criteria in addition to these new eligibility criteria.

Finally, if in addition to this eligibility issue, some of the selected and recruited schools that *are* eligible declined participation, and if those that declined differed from those that volunteered, then this, too, requires adjustment. This is the same adjustment as found in Section 1 above. In this case, the total weight requires that all three weights are multiplied (Total Weight = Sample Selection Weight × Eligibility Determination Weight × Participation Weight). These Total Weights are then used in analyses produced using the final sample in the study.

Figure 4. Two-stage statistical adjustments when some recruited schools are found to be ineligible


Concepts
Process for Obtaining the Sample
Statistical Adjustments
Sample selection
The kinds of schools selected and recruited for the study do not represent the target population.
Eligibility determination
The kinds of schools that are found to be eligible for the study differ from those that are ineligible.
Participation
The kinds of schools that volunteer to be in the study – once recruited – differ from those that decline participation.
Population Frame
Sampled & Recruited
Confirmed as Eligible
Agreed to Participate
(a) Weighting adjustment for differences between the frame and the schools that were selected and recruited
(b) Screen out ineligible schools and describe eligible target population
(c) Weighting adjustment for differences between confirmed eligible schools and the subset that agree to participate

3. *Propensity score methods*

Another approach to post-stratification and weighting involves propensity score methods. These are particularly useful when the generalizability index is below 0.5. Unlike the previous approach to post-stratification, however, propensity scores do not require strata to be defined in advance of selection (for development of these methods, see Stuart et al., 2011; Tipton, 2013). In general, this approach involves the following algorithm:

1. **Estimate selection probabilities**. These are also known as sampling propensity scores and predict the probability that schools were included in the sample conditional on the identified set of potential moderators. Consider estimating these using a logistic regression model in which the outcome is 1 if the school was in the study and 0 if the school was not, and the covariates include potential moderators.
2. **Assess the similarity between the sample and target population.** In this step, compare the distributions of sampling probability estimates (or their log-odds) for the sample and population. Consider comparing them visually using histograms. You may also calculate the generalizability index for the sampling probabilities.
   – If the index is medium to large (> 0.50), then proceed to Step 4.

  - If the index is low (< 0.50), the two distributions likely do not completely overlap with one another (indicating undercoverage), and generalizations to the target population would involve extrapolations beyond the study data. Proceed to Step 3.

3. **Redefine the target population.** Define a subpopulation for which generalizations would not require extrapolations beyond the study data:
   - For each of the potential moderators, compare summary statistics (min, median, max) in the sample and population. Identify a potential moderator for which the population includes some schools that fall outside of the range observed in the sample–that is, whose values are below the minimum in the sample or above the maximum in the sample.
   - Define a new subpopulation that includes only those schools in the population within the range of values found in the sample.
   - Proceed through Steps 1 and 2 for this new subpopulation.
4. **Use the estimated selection probabilities to reweight the sample.** Several approaches are available, including inverse probability weighting (Stuart, 2010; Stuart et al., 2011) and subclassification (Tipton, 2013; O'Muircheartaigh & Hedges, 2014).
   - *To use inverse probability weighting*, create weights defined as $w_i = 1/p_i$ where p_i is the probability that school i would be in the study (estimated in Step 1). Conduct analyses using these weights (for example, weighted regression).
   - *To use propensity score subclassification*, divide the population frame into five or more equally sized cells based on values of the sampling probabilities in the population, keeping in mind that more cells are better. Next, locate schools within the sample in each of the strata. Your analyses will proceed similarly to post-stratification, using the propensity score subclasses instead of the original sampling strata to construct weights.
5. **Assess the similarity between the weighted sample and (sub)population.** Assess whether the propensity score weights succeeded in balancing the sample with the population (or newly defined subpopulation, see Step 3 above).
   - If the sample and (sub)population have similar potential moderator values–that is, SMDs for moderators are all below 0.25–then proceed to step 6.
   - If the sample and (sub)population are not sufficiently similar, go back to Step 3 and consider restricting the study to a narrower subpopulation.
6. **Apply the chosen weights with the study *outcomes***. Be sure to use these weights for both the population average treatment effect and the associated standard errors. Clearly report the definition of the final target population (using inclusion criteria), the estimation strategy used, the potential moderators adjusted for, and the final results.

4. Model-based approaches

When the generalizability index is less than 0.5, model-based approaches offer possible alternatives to propensity score methods. These seek to model the outcome directly as a function of the potential moderators and to use these models to *predict* the average treatment effect in the target population. These approaches require knowledge of population means for each of the potential moderators but do not require a full population frame. The simplest of these estimators is multiple regression, including interactions between potential moderators and an indicator of the intervention. In general, using this approach involves the following algorithm:

1. In the *sample* data, center all of the potential moderators around the *population* mean (for example, $X_{ci} = X_i - \overline{X}_p$ where $\overline{X}_p$ is the population mean).
2. Fit a model including an indicator for the intervention (Trt), the centered potential moderators ($X_1, \ldots, X_p$), and interactions between the intervention and the centered moderators. Here we illustrate this in a simple randomized experiment (no nesting) with a single potential moderator:

$$Y_i = \beta_0 + \delta_p Trt_i + \beta_1 X_{ci} + \beta_2 X_{ci} * Trt_i + \in_i$$

3. In this model, δ_p is the average treatment effect in the population. To see why, note that when $X_{ci} = \overline{X}_p$, then $X_{ci} = 0$ and

$$\widehat{Y}_{\iota} = \beta_0 + \delta_p Trt_i .,$$

and thus δ_p is an estimate of the difference between treatment and control means in the population.

Although multiple regression is the simplest model-based approach, it can be difficult when there are many potential moderators. This difficulty arises when the number of potential moderators is large relative to the number of schools. A solution is to reduce the number of predictors (moderators) in the model, but selecting the right subset is challenging. In comparison, more flexible methods—like Bayesian Additive Regression Trees—can be more effective. Kern and colleagues (2016) provide an overview of these model-based methods and evaluate their efficacy using several real-data examples; however, these methods and examples all focus on larger samples than are typically found in education research (e.g., *n* > 100 schools), suggesting that more research is needed.

Recommendation 7. Report the Generalizability of Findings Appropriately

Goal: Convey where results of the study may generalize and where they may not and the assumptions required to make those generalizations

Steps:

1. Throughout the study, collect information necessary for reporting generalizability.
2. Integrate generalizability into all facets of reports on study findings.

When writing study reports or papers, you must include sufficient information to help the reader better understand where an impact estimate will generalize and where it will not. To do this, describe the target population and all of the steps you took to estimate treatment effects specifically for that population. In addition, describe any limitations that may threaten the generalizability of the findings to the target population. Finally, although not covered in this section, researchers are encouraged to report any additional information that would be helpful to consumers focused on *other* target populations. This information includes descriptive information on the students, teachers, and schools that participated in the study, impact estimates for different subgroups of students and schools based on moderator variables included in the population frame (see Recommendation 2) or collected during the course of the study, and the results from statistical tests for whether impacts vary across subgroups.

How to carry out the recommendation

1. Throughout the study, collect information necessary for reporting generalizability

Throughout the time you are conducting the study, keep records, collect data on, and report the steps that were designed to produce findings that generalize to the target population, plus information that will help readers assess the success of those efforts.

Table 10 lists reporting guidelines appropriate for each of the recommendations, as summarized from previous sections of this guide.

Table 10. Reporting guidelines

Recommendation	Information to report
Recommendation 1 Define the target population for the study	• Potential moderators of the impact • The target population of students and schools on which the evaluation will focus
Recommendation 2 Develop a population frame	• Inclusion/exclusion criteria for the study • The total number of schools included in the population frame • Summary statistics on potential moderators in the population frame
Recommendation 3 Design a sampling plan	• The variables used to construct the strata and the method used for creating the strata (such as k-means clustering)The number of strata defined, the proportion of schools from the population frame in each stratum, and the number of schools to be recruited from each stratum • The strategy for sampling schools within strata
Recommendation 4 Implement the sampling plan	• The total number of schools contacted and the total number that agreed to participate • Differences in potential moderators between schools that agreed to participate in the study and schools that were selected to participate but declined
Recommendation 5 Assess the similarity between the sample and the target population	• Differences in potential moderators between the sample and target population, as operationalized in the population frame • A measure of overall similarity (such as the generalizability index)
Recommendation 6 Adjust for differences between the sample and the target population	• The final estimation method implemented (for example, post-stratification) and the extent to which this method increases similarity between the sample and target population • The final target population for which generalization possible • The estimate of the average treatment effect for the target population as well as its standard error • Assumptions required for the estimation method to result in an unbiased estimate

2. Integrate generalizability into all facets of reports on study findings

Report information about the generalizability of the study in the appropriate sections of your report or manuscript. See below for guidance on what to report in different sections, from the study abstract through the conclusion:

- **Abstract.** The study abstract should identify the target population for which the average treatment effect applies. In many cases, this is the target population identified at the beginning of the study. In cases in which the sample and population ultimately differ because of undercoverage, this should instead be the subpopulation that was identified in the analysis.
- **Introduction.** Provide detailed information regarding the target population of focus in the study. For context, you should identify potential treatment effect moderators. Also provide the inclusion and exclusion criteria, any geographic constraints, and the total

number of schools in the population. Ideally you will explain the logic for this choice of target population, including how this target population is similar to or different from previous studies of the same type.

- **Methods.** The methods section should include information regarding the development of the population frame, including the data sources for potential moderators and the creation of strata. Explain the method for within stratum recruitment and your recruitment procedures. Additionally, introduce the procedure for assessing similarity and adjusting for differences.
- **Analysis.** The analysis section should include an analysis of recruitment, including information on the total number of schools contacted, the number that agreed to participate, the reasons given for not participating, and differences in the characteristics between volunteers and refusals. Additionally, this section should include a comparison of the final sample to the target population with regard to potential moderators, before and after statistical adjustments are made for any differences. Clearly state the assumptions regarding these adjustments and any limitations of the analyses.
- **Conclusion.** When discussing the broader implications of the study results for both science and policy, take care to constrain generalizations to the target population for the study, as originally defined or restricted to support stronger generalizations. If it is possible that the results extend more broadly than the target population, clearly state this as a hypothesis, not a conclusion.

Example

In what follows, we provide sample language regarding how generalizability could have been included in an academic paper or study report for the example used throughout this guide. We only include here the parts related to generalization, keeping in mind that each section of the paper would also report other facets of the study as well.

Abstract

“This paper focuses on the evaluation of this reading program in the population of public, Title I ‘regular’ elementary schools serving majority non-White students in the southeastern United States.”

Introduction

“The development and prior studies of this program were conducted in large, urban elementary schools. In defining the population for this study, we include elementary schools serving majority non-White and low-SES students, regardless of their size and locale, to test whether the intervention is effective in a broader population. But we restrict the population to schools in the Southeast so that the study could be feasibly conducted in a reasonable time frame. Overall, this population included about 4,100 ‘regular’ public Title I schools.”

Methods

“In order to represent this target population, we used *The Generalizer* to develop a recruitment plan. First, we identified a broad set of potential moderators, including student characteristics and contextual factors. *The Generalizer* then used cluster analysis to divide schools in the population frame into five strata based on the moderators included in the frame. Stratum membership explained 44 percent of the variance in these moderators across schools. Based on the proportion of the schools in the target population in each stratum and the total number of schools required for the sample (from our power analysis), recruitment targets were provided for each stratum.

“Before the study began, the PI already had relationships with 10 schools that agreed to be in the study; these schools were primarily in two of the strata. The recruitment process focused on filling in the remaining 30 schools necessary to meet the stratum recruitment goals. Researchers randomly ordered schools in the strata and recruiters were given lists of schools to recruit. Ten of the 120 schools recruited were found to be ineligible for the study. These schools were found to be similar in their characteristics to other schools in the population frame, suggesting that the population frame provides an adequate description of the target population for the study. Thirty of the remaining 110 recruited schools agreed to participate in the study. The total sample of 40 participating schools, including the 10 schools that agreed initially and the 30 schools recruited once the study began, was compared to the population frame in terms of potential moderators; we summarize these differences in terms of a generalizability index (Tipton, 2014a) and standardized mean differences. Based upon these results, we adjusted for slight differences in composition by using post-stratification weights.”

Analysis

“Schools that agreed to take part in the study tended to serve more students and larger percentages of low-SES students (indicated by eligibility for free or reduced-price lunch) and Black students compared to those that declined. The predominant reasons schools gave for not participating were largely exogenous to the study (for example, change in school leadership). However, schools that declined were more likely than schools that joined the study to use one of two well-established reading programs, neither of which included the core feature of the tested intervention, which was to ensure representation of Black and Hispanic students in the intervention’s written content.

“Recruitment was more difficult in some strata than others, resulting in a few differences between the sample and population. The generalizability index, which summarizes the differences, was calculated to equal 0.82, suggesting a high degree of similarity (Tipton, 2014a). But schools in the study differed from those in the population in terms of school size, student SES, student racial composition, and urbanicity. We thus used a post-stratification adjustment when estimating the population average treatment effect. This adjustment involved up-weighting schools in some strata (2 and 5) and down-weighting those in others (3 and 4). The post-stratification adjustment greatly improved the similarity between the sample and population (reweighted index of 0.93). Importantly, this adjustment affected both the estimate of the population impact and the standard error of this estimate."

Conclusion

"This study indicated that, on average, the reading program led to an estimated increase of 0.26 (SE = 0.12) standard deviations in reading scores for K-2 students in 'regular' public Title I elementary schools serving predominately non-White students in the Southeast. The estimated impact was statistically significant and corresponds to three months of additional learning. These results are promising, given the diversity of schools found in the southeast, including both rural and urban schools, small and large schools, and schools serving moderate and large percentages of Black and Hispanic students. However, this study does not include the full diversity of schools found in the broader United States, and researchers and policymakers should keep this in mind when applying the results of this study beyond the characteristics of this population."

References

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, *38*(2), 318-335.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness, 10*(4), 817-842.

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a culture of replication: An examination of education and special education research grants funded by the Institute of Education Sciences. *Educational Researcher, 47*(9), 594-605.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*(2), 295-313.

Coyne, M. D., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education: A framework of systematic, conceptual replications. *Remedial and Special Education*, *37*(4), 244-253.

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24-67.

Gleason, P., Crissey, S., Chojnacki, G., Zukiewicz, M., Silva, T., Costelloe, S., & O'Reilly, F. (2019). *Evaluation of support for using student data to inform teachers' instruction* (NCEE 2019-4008). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review, 37*(6), 445-489.

Hedges, L. V., & Rhoads, C. (2009). *Statistical power analysis in education research* (NCSER 2010-3006). U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research.

Heppen, J. B., Kurki, A., & Brown, S. (2020). *Can texting parents improve attendance in elementary school? A test of an adaptive messaging strategy* (NCEE 2020-006a). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Herrmann, M., Clark, M., James-Burdumy, S., Tuttle, C., Kautz, T., Knechtel, V., Dotter, D., Wulsin, C. S., & Deke, J. (2019). *The effects of a principal professional development program focused on instructional leadership: Appendices* (NCEE 2020-0002). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103-127.

Olsen, R. B., & Orr, L. L. (2016). On the "where" of social experiments: Selecting more representative samples to inform policy. *New Directions for Evaluation*, *2016*(152), 61-71.

O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C: Applied Statistics, 63*(2), 195-210.

Orr, L. L., Olsen, R. B., Bell, S. H., Schmid, I., Shivji, A., & Stuart, E. A. (2019). Using the results from rigorous multisite evaluations to inform local policy decisions. *Journal of Policy Analysis and Management, 38*(4), 978-1003.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start impact study. Final report*. Administration for Children & Families, U.S. Department of Health and Human Services.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*(3), 279-292.

Rubin, D.B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2,* 169-188.

Ruel, T., Ackerman, B., Coburn, K., Chao, B., & Tipton, B. (2022). generalizeR: Design a sample recruitment plan and assess its generalizability to broader populations. https://nustat.github.io/generalizeR.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90-100.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, 33(1), 62-87.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" software*. William T. Grant Foundation.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 25*(1), 1-21.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of Research on Educational Effectiveness*, *10*(1), 168-206.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(2), 369-386.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38,* 239-266.

Tipton, E. (2014a). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics, 39*(6), 478-501.

Tipton, E. (2014b). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, *37*(2), 109-139.

Tipton, E. (2021) Beyond the ATE: Designing impact evaluations to understand treatment effect heterogeneity. *Journal of the Royal Statistics Society: Series A,* 184(2), 504-521.

Tipton, E. (2022) Sample selection in randomized trials with multiple target populations. *American Journal of Evaluation,* doi:10.1177/1098214020927787.

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Ruiz de Castilla, V. (2016). School selection in experiments: An assessment of school recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness*, *9*(Suppl. 1), 209-228.

Tipton, E., & Miller, K. (2021). *The Generalizer* [Web tool]. https://thegeneralizer.org.

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

Tipton, E., Spybrook, J., Fitzgerald, K. G., Wang, Q., & Davidson, C. (2021). Toward a system of evidence for all: Current practices and future opportunities in 37 randomized trials. *Educational Researcher, 50*(3), 145-156.

Tipton, E., Yeager, D., Iachan, R., & Schneider, B. (2019). Designing probability samples to study treatment effect heterogeneity. *Experimental Methods in Survey Research: Techniques That Combine Random Sampling with Random Assignment*, 435-456.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management, 33*(3), 778-808.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness, 10*(4), 843-876.

Wright, C., Ellis, S. E., Hicks, S. C., & Peng, R. D. (2021). *Tidyverse skills for data science*. https://jhudatascience.org/tidyversecourse/.