




Selecting Districts and Schools for Impact Studies in Education: A Simulation Study of Different Strategies

Daniel Litwok, Austin Nichols, Azim Shivji & Robert B. Olsen


To cite this article: Daniel Litwok, Austin Nichols, Azim Shivji & Robert B. Olsen (2022): Selecting Districts and Schools for Impact Studies in Education: A Simulation Study of Different Strategies, Journal of Research on Educational Effectiveness, DOI: [10.1080/19345747.2022.2128952](https://doi.org/10.1080/19345747.2022.2128952)

To link to this article: <https://doi.org/10.1080/19345747.2022.2128952>


 [View supplementary material](#) 


 Published online: 08 Nov 2022.


 [Submit your article to this journal](#) 

 [View related articles](#) 

 [View Crossmark data](#) 

 This article has been awarded the Centre for Open Science 'Open Data' badge.

 This article has been awarded the Centre for Open Science 'Open Materials' badge.

 This article has been awarded the Centre for Open Science 'Preregistered' badge.



Selecting Districts and Schools for Impact Studies in Education: A Simulation Study of Different Strategies

Daniel Litwok^a, Austin Nichols^a, Azim Shivji^{b*}, and Robert B. Olsen^c

^aSocial & Economic Policy Division, Abt Associates, Rockville, MD, USA; ^bSan Jose, CA, USA; ^cGeorge Washington Institute of Public Policy, The George Washington University, Washington, DC, USA

ABSTRACT

Experimental studies of educational interventions are rarely based on representative samples of the target population. This simulation study tests two formal sampling strategies for selecting districts and schools from within strata when they may not agree to participate if selected: (1) balanced selection of the most typical district or school within each stratum; and (2) random selection. We compared the generalizability of the resulting impact estimates, both to each other and to a stylized approach to purposive selection (the typical approach for experimental studies in education). We found that balanced and random selection of schools within randomly selected districts were the most consistent strategies in terms of generalizability, with minimal difference between the two. Separately, for random selection, we tested two strategies for replacing districts that refused to participate—random and nearest neighbor replacement. Random replacement outperformed nearest neighbor replacement in many, but not all, scenarios. Overall, the findings suggest that formal sampling strategies for selecting districts and schools for experimental studies of educational interventions can substantially improve the generalizability of their impact findings.

ARTICLE HISTORY

Received 16 January 2022
Revised 2 August 2022
Accepted 17 August 2022

KEYWORDS

Generalizability; analysis methods; research design - experimental

Introduction

Randomized controlled trials (RCTs) are widely regarded as providing the most convincing evidence on the impacts of interventions because of their high internal validity, meaning they provide strong evidence of causal impacts. For this reason, evidence clearinghouses like the What Works Clearinghouse typically reserve their highest ratings for RCTs.¹ However, it can be difficult to conduct RCTs that have high external validity, meaning the study results are generalizable to a target population.

Problems of generalizability often arise during site recruitment. Given limited budgets, researchers can minimize recruitment costs by selecting districts and schools for

CONTACT Daniel Litwok  Dan_Litwok@abtassoc.com  Social & Economic Policy Division, Abt Associates, 6130 Executive Blvd., Rockville, MD 20852, USA.

 Supplemental data for this article is available online at <https://doi.org/10.1080/19345747.2022.2128952>.

*The University of Chicago Harris School of Public Policy, Chicago, IL, USA.

¹See <https://ies.ed.gov/ncee/wwc/>.

© 2022 Abt Associates

recruitment that offer the most “bang for their buck”—i.e., those that are large and likely to participate in an evaluation. Districts and schools also typically have weak incentives to participate, and many of them choose to opt out if recruited. As a result, the combination of researcher decisions about recruitment and district or school decisions about whether to participate in the evaluation can yield a sample that may not represent the target population of interest. For example, districts that participate in large, multi-site RCTs tend to be much larger, more urban, and more disadvantaged than the average district that could have implemented the intervention (Stuart et al., 2017). Similarly, large schools from large school districts in urban areas tend to be over-represented in cluster RCTs (Tipton et al., 2021).

Recruitment that favors certain types of sites can lead to “external validity bias,” meaning the impact estimates from the study sample are biased estimates of the impact in the population of interest (Olsen et al., 2013). Early evidence suggests that the magnitude of this bias can be sizable (Bell et al., 2016). The literature prescribes several prospective strategies for limiting external validity bias at the design phase of an education RCT (Olsen & Orr, 2016; Tipton, 2013b). This study assesses the performance of these site selection strategies in a simulated environment.²

The simulated environment is a hypothetical evaluation of an educational intervention targeting schools serving kindergarten through fifth grade (grades K–5). We used publicly available data from the U.S. Department of Education (the Common Core of Data) to define a target population of schools and generated hypothetical impacts for that entire population. We constructed the hypothetical impacts to be consistent with existing evidence on impact variation across schools for educational interventions. We varied other key parameters over plausible values to make the results more broadly informative.

Our analysis answers four research questions:

1. Which district and school selection strategies performed best in terms of producing samples with the greatest generalizability to the target population?
2. Which district and school selection strategies yielded the least recruitment burden?
3. How did the relative performance of strategies change with changes to key simulation parameters?
4. Limiting focus to one of the selection strategies (random selection), which method for replacing districts that declined to participate performed best?

Regardless of selection strategy, we stratified the population in the same way using likely impact moderators. We chose this approach to ensure our tests would capture differences in the performance of the strategies themselves rather than different stratification schemes. Specifically, we stratified using k-means clustering, which is an algorithm that partitions data into a user-selected number of clusters (k).³

²Consistent with best practices for transparent research, we preregistered an analysis plan for this simulation that described and justified decisions in advance of analysis on the Open Science Framework. We subsequently revised that plan. See <https://osf.io/fehjc> for more information on both plans and the [online appendix](#) for further discussion.

³See [Appendix A](#) for more technical details on our implementation of k-means clustering.

Within those strata, the strategies we tested included elements of three approaches: (1) a stylized version of purposive selection—sometimes referred to as convenience sampling—in which the largest districts and schools were selected; (2) random selection as used in surveys; and (3) balanced selection proposed by Tipton (2013b), which prioritizes the districts and schools that are closest to the stratum means. Throughout the article we refer to these strategies as “purposive,” “random,” and “balanced,” respectively. We tested all combinations of applying these approaches at both the district and school levels for a total of nine selection strategies.

That all strategies include stratification implies our simulation does not cover some strategies that may exist in practice, such as purposive selection without stratification. Stratification is a best practice for selecting a sample that is representative of a well-defined population (Tipton & Olsen, 2022). The strategies we tested all likely performed better than they would have in the absence of stratification.

To compare the selection strategies, we specified an “impact generating process” for the hypothetical intervention for each school in the population. We used the different site selection strategies to generate rank-ordered lists of districts and schools from the target population for recruitment; simulated decisions by districts and schools to agree or decline to participate in the evaluation, conditional on selection; and selected replacement districts and schools from the target population for those that did not agree to participate until a study sample was identified. We simulated participation decisions by districts and schools by taking a random draw from a Bernoulli distribution with agreement probabilities that could vary depending on observed and unobserved characteristics. A simple average of the impact across all schools in the target population was the Population Average Treatment Effect (PATE)—the estimand of interest. The simple average of the impact across all schools that agreed to participate in the study was the Sample Average Treatment Effect (SATE).

To answer the first research question, we compared the SATE across simulated samples to the PATE under a set of baseline parameter values. Our primary measure of performance—mean squared error—was the average squared difference between the SATE and the PATE (we also calculated and reported other measures of performance including relative mean squared error, absolute external validity bias, and standard deviation). To answer the second research question, we tabulated the number of districts and schools recruited by each of the selection strategies to achieve the target sample. To answer the third research question, we explored the extent to which the findings varied as we changed parameters from their baseline values.

Focusing specifically on random selection, the fourth research question is motivated by large-scale educational surveys (such as the National Assessment of Educational Progress (NAEP) or the Programme for International Student Assessment (PISA)) that replace schools that decline to participate with their “nearest neighbor,” or another school with similar observable characteristics. We tested this strategy for random district selection. We hypothesized that, under certain conditions, nearest neighbor replacement might result in a more representative sample than random replacement. By way of intuition, suppose the initial set of randomly selected districts included some atypical districts that had a low probability of initial selection. If replaced by a randomly selected replacement district, it is unlikely that the replacement district would have those same

atypical features. However, nearest neighbor replacement would ensure that the replacement district had similar characteristics to the initially selected district. To answer this research question, we altered the replacement algorithm to choose the nearest neighbor of the declining district (in terms of the distance metric we use in our stratification algorithm and balanced sampling) rather than the next randomly selected district and compared measures of performance.

To preview the results, strategies that rank-ordered districts randomly outperformed the other strategies for all performance measures and nearly all values of simulation parameters. Within districts, random and balanced school selection performed similarly. Focusing specifically on random selection of districts, random replacement of districts that refused to participate outperformed nearest neighbor replacement in many but not all scenarios. These findings imply that relatively simple adjustments to the process of selecting districts and schools, such as incorporating randomness, results in a sample that better represents the target population of interest.

Related Literature

Literature in many fields has documented theoretical concerns with external validity of experimental impact evaluations (Banerjee & Duflo, 2009; Heckman & Smith, 1995; Imbens & Wooldridge, 2009; Orr, 1999; Shadish et al., 2002). Nonetheless, this issue has been largely treated as a second-order concern in applied education evaluations. With recent growth in evidence-based policymaking and rigorous evidence standards, applied education researchers began to explore the threat of external validity bias (Olsen et al., 2013) and empirically test for its prevalence (Bell et al., 2016; Bell & Stuart, 2016; Fellers, 2017; Stuart et al., 2017; Tipton et al., 2016, 2021).

With a better understanding of the problem, the literature began to propose strategies for addressing external validity bias. Two separate classes emerged: (1) retrospective strategies for generalizing results from completed evaluations to a target population; and (2) prospective strategies for designing evaluations to be generalizable to a target population. Of note, prospective and retrospective strategies need not be mutually exclusive—one can also apply a retrospective strategy to an analysis that was designed to be generalizable (Tipton & Olsen, 2018, 2022).

Retrospective strategies use information available on the study sample and the broader population to adjust results of an experiment that is already complete. A variety of strategies are available, including reweighting via poststratification or propensity scores, regression modeling of impact variation, and bounding (e.g., Andrews & Oster, 2018; Chan, 2017; Kern et al., 2016; Nguyen et al., 2017; O’Muircheartaigh & Hedges, 2014; Stuart et al., 2011). Extensive work describes the properties of these strategies (e.g., Kern et al., 2016; Tipton et al., 2017), establishes the required assumptions (e.g., Tipton, 2013a), and guides practitioners through applying the strategies (e.g., Stuart & Rhodes, 2017). While retrospective strategies are useful and important, they are not the focus of this work.

This study contributes to the literature on prospective strategies for designing RCTs to produce generalizable findings. Several different strategies have been recommended in the literature, including those that propose random selection of sites, such as schools

or districts (Allcott, 2015; Olsen & Orr, 2016) and those that propose balanced selection of sites (Tipton, 2013b; Tipton et al., 2014). As with retrospective strategies, practitioner guides are available to facilitate implementation (Tipton & Matlen, 2019; Tipton & Peck, 2017); yet no published study has compared the relative performance of these prospective strategies—a gap in the literature that this study fills.

Simulation Methods

Our approach to conducting the simulation began with constructing the population, defined as the collection of schools over which the hypothetical study aimed to estimate average impacts. Recent guidance on improving the generalizability of impact studies in education has highlighted the importance of defining the target population in terms of the schools to which the study aims to generalize (Tipton & Olsen, 2018, 2022). This section defines the population of schools for our simulations and describes the characteristics of these schools (including both characteristics that were fixed across simulations and those that we varied). Next, we describe how samples of schools were selected, including the details of stratification, initial selection of districts and schools, and replacement of districts and schools that declined to participate. Finally, we describe the simulation details used to assess the performance of different strategies, including repeated sampling, measures of performance, and the parameters we changed across iterations of the simulation.

Fixed Population Factors

Districts, Schools, and Their Characteristics

Our simulation evaluated a hypothetical intervention targeting K–5 schools. As such, we defined the target population for the simulated study to be all open, regular schools that covered grades K through 5.⁴ We identified these 42,752 schools (situated in 11,733 districts) in the 2018–2019 School District (LEA) universe survey, the 2018–19 School universe survey, and the 2016–2017 School District (LEA) finance data maintained by the National Center for Education Statistics in the Common Core of Data (CCD). The following five school- and district-level variables were extracted and used in the analysis (level of the data in parentheses): (1) total enrollment (school), (2) percentage of students eligible for free or reduced-price lunch (FRPL) (school), (3) number of eligible schools (district), (4) Census region (district), and (5) expenditures per pupil (district).⁵ We also synthetically generated an unobserved district-level variable that was not in the CCD, which we labeled “district administrator leadership.”⁶

Using the five observed variables from the CCD and the unobserved district administrative leadership variable, we specified an “impact generating process” that is a linear

⁴Specifically, we included schools that had each of the grade levels between grade K and grade 5, inclusive.

⁵A small share of these items was missing in the data. We addressed this using one of two strategies: (1) finding a non-missing value from a prior year; or (2) using a single regression-based imputation. See Litwok et al. (2021) for additional detail.

⁶We generated this variable using a random draw from the standard normal distribution for each district. We conducted the random draw a single time to make this feature a fixed component of the population of schools—we did not change the draw across iterations of the simulation.

Table 1. Summary statistics.

	Total enrollment	% FRPL eligible	Number of eligible K–5 schools per district	Expenditures per Pupil (\$)	Administrator leadership
Mean	487	57	4	12,607	0
Standard deviation	283	29	11	5,070	1
Median	458	58	1	11,421	–0.01
Minimum	50	0	1	481	–3.92
Maximum	14,306	100	476	115,932	3.94
Sample Size	42,752	42,752	11,733	11,733	11,733

Note. Total Enrollment and % FRPL Eligible reported at the school level. Number of Eligible K-5 Schools per District, Expenditures per Pupil, and Administrator Leadership reported at the district level.

Source. Author tabulations of Common Core of Data (Administrator Leadership generated by authors—see footnote 6).

Table 2. Regional distribution of study population.

Region	Share (District-Level)	Share (School-Level)
Northeast	18%	14%
Midwest	33%	22%
South	24%	36%
West	25%	29%

Note. $N = 42,752$ schools; 11,733 districts.

Source. Author tabulations of Common Core of Data.

function of five of these variables. Because these variables include a mix of school-level and district-level factors, the impact varies across districts and also across schools within the same district. In addition, we specified a probabilistic process by which districts agree or refuse to participate, and this process was a function of selected observed variables and the unobserved district administrative leadership variable. This framework was designed to mimic a realistic scenario in which unobserved characteristics influence both the impact of the intervention and the probability of inclusion in the sample. Under this scenario, external validity bias cannot generally be eliminated by the sampling methods tested in this article, and the question is which of the methods will yield the smallest bias or mean squared error.

Tables 1 and 2 report summary statistics for these six variables. The average school in the study population had 487 students, 57 percent of whom were eligible for FRPL. The average district in the study population had four eligible K–5 schools and spent roughly \$12,500 per pupil. The median district had only one eligible school—68 percent of the districts in the population had only one eligible school, and 83 percent had fewer than five eligible schools. Districts and schools in the population were distributed across all four Census regions, with the Midwest having the largest number of districts and the South having the largest number of schools.

Variable Population Factors

Generating Impacts

We generated the impact of the hypothetical intervention for all schools in the target population. Our goal was to generate impact variation across schools that is typical for educational interventions, as demonstrated by recent empirical research. We also varied

the share of that variation that could be explained by observed covariates, since that share could differ considerably across interventions and studies.

To generate a distribution with these properties, we expressed the school-specific impact (Δ_{sd}) for school s in district d as in Equation (1):

$$\Delta_{sd} = \mu_{sd} + \sum_{i=1}^3 \beta_i X_{id} + \sum_{i=4}^5 \beta_i X_{isd} \tag{1}$$

where μ_{sd} is a normally distributed school-specific random effect (with a mean and variance to be determined and assumed to be uncorrelated with the other covariates), X_{id} and X_{isd} denote each of five covariates—standardized to each have a standard deviation of 1—and β_i denotes each of five fixed coefficients that capture the conditional relationship between the impact and the standardized covariates. Three of the five covariates are at the district level: number of schools per district, expenditures per pupil, and administrator leadership; the other two are at the school level: total enrollment, and percent FRPL eligible. For simplicity, we set the five β_i coefficients equal to each other. This is a strong assumption, as the relative importance of different covariates in determining the magnitude of the impact could vary across interventions. We relaxed this assumption as a robustness check.

To simplify algebra, it is useful to convert the two summations in Equation (1) to matrix notation:

$$\Delta_{sd} = \mu_{sd} + \mathbf{X}\boldsymbol{\beta} \tag{2}$$

where the matrix \mathbf{X} contains the five covariates and the vector $\boldsymbol{\beta}$ has five coefficients. Taking the variance of both sides and rearranging terms yields Equation (3):

$$\text{Var}(\mu_{sd}) = \text{Var}(\Delta_{sd}) - \boldsymbol{\beta}'\mathbf{G}\boldsymbol{\beta} \tag{3}$$

where \mathbf{G} is the variance-covariance matrix for \mathbf{X} . Defining W as the share of impact variance not explained by covariates ($W \equiv \frac{\text{Var}(\mu_{sd})}{\text{Var}(\Delta_{sd})}$):

$$\boldsymbol{\beta}'\mathbf{G}\boldsymbol{\beta} = (1 - W)\text{Var}(\Delta_{sd}) \tag{4}$$

We generated the impact for each school in the population by: (1) varying the proportion of variation not explained by covariates, W , by setting it to the following values: 0.1, 0.25, 0.5, 0.75, 0.9, or 1; and (2) varying the unconditional variation of impact across schools, $\sqrt{\text{Var}(\Delta_{sd})}$, by setting it to the following values: 0.05, 0.1, and 0.2. We centered the values for this parameter around 0.1 standard deviations because this is the median variation in impacts for relevant studies reported in Weiss et al. (2017).⁷ Plugging these values into Equation (4) allowed us to derive values for $\boldsymbol{\beta}$ that could be used to generate impacts for the population of schools (see Equation (1)). For each combination of $\text{Var}(\Delta_{sd})$ and W , Table 3 reports the resulting value of $\boldsymbol{\beta}$.⁸

⁷Although Weiss et al. (2017) document the variation in impacts for 16 multisite RCTs, we focus on the five RCTs for which they report estimates that are relevant to our hypothetical intervention because they are based on achievement outcomes in elementary or middle school: (1) After School Reading, (2) After School Math, (3) Teach for America, (4) Charter Middle Schools, and (5) Enhanced Reading Opportunity.

⁸There are slight deviations in Table 3 from the version of this table that appeared in Litwok et al. (2021). These deviations are due to (1) standardizing district administrator leadership; and (2) updating the covariate imputation model to correctly treat region as a factor variable.

Table 3. Parameter values for impact model.

$\text{Var}(\Delta_{sd})$	W	$\beta'G\beta$	$\text{Var}(\mu_s)$	β
0.0025	0.1	0.00225	0.00025	0.0207
0.0025	0.25	0.001875	0.000625	0.0189
0.0025	0.5	0.00125	0.00125	0.0154
0.0025	0.75	0.000625	0.001875	0.0109
0.0025	0.9	0.00025	0.00225	0.0069
0.0025	1	0	0.0025	0
0.01	0.1	0.009	0.001	0.0414
0.01	0.25	0.0075	0.0025	0.0378
0.01	0.5	0.005	0.005	0.0308
0.01	0.75	0.0025	0.0075	0.0218
0.01	0.9	0.001	0.009	0.0138
0.01	1	0	0.01	0
0.04	0.1	0.036	0.004	0.0828
0.04	0.25	0.03	0.01	0.0756
0.04	0.5	0.02	0.02	0.0617
0.04	0.75	0.01	0.03	0.0436
0.04	0.9	0.004	0.036	0.0276
0.04	1	0	0.04	0

Note. $\text{Var}(\Delta_{sd})$ based on distribution from Weiss et al. (2017).

Source. Author calculations.

Agreement to Participate

We designed the simulations to test the performance of different selection strategies when districts and schools are not required to participate. Therefore, we also generated district and school participation decisions. A district was only included in the sample if it was selected and agreed to participate in the study. A school was only included in the sample if it was selected, its district agreed to participate in the study, and the school itself agreed to participate in the study. This conforms with our experience in conducting RCTs, where schools need district permission to participate, but they also have the latitude to opt out of the study if they do not want to participate.

We used the following model to generate district participation decisions:

$$R_d = \log\left(\frac{T}{1-T}\right) + \omega(\Delta_d - PATE) \quad (5)$$

$$P_d = \frac{e^{R_d}}{1 + e^{R_d}} \quad (6)$$

To determine if the district would agree to participate, we drew a random variable from the Bernoulli distribution with probability P_d , where a value of 1 indicated that the district agreed to participate.

In Equation (5), R_d is a latent variable that consists of a nonrandom intercept shift that is determined by the target share of districts that agree to participate (T), the district-level impact (Δ_d) relative to the PATE, and a parameter that captures the relationship between the district-level impact and the latent variable (ω).⁹ Equation (6) uses an inverse logistic function to transform this latent variable into a probability that the district agrees to participate bounded by zero and one. Equations (5) and (6) together imply that districts with impact equal to the PATE had a probability of participation equal to T ; the scaling parameter ω and district-level variation in impacts around the

⁹The district-level impact is the simple average of school impacts (with each school weighted equally).

PATE generated variation in agreement to participate around T . Districts with larger impacts were more likely to participate.

This model made the probability of agreeing to participate a function of the intervention's average impact in the district. Given the model used to generate impacts, Equations (5) and (6) implies self-selection led large districts, well-resourced districts, and districts with larger and higher-poverty schools to be overrepresented in the sample. These relationships are consistent with prior work that reported the characteristics of districts that participated in large RCTs (Stuart et al., 2017). However, that study did not separately identify whether these were characteristics associated with selection or agreement to participate.

Because districts' willingness to participate and the bias introduced by these decisions varies across studies, we varied the district-level target probability and the strength of the relationship between impacts and agreement to participate. We varied the target probability (T) from 0.1 to 0.4 in increments of 0.1.¹⁰ We also varied the strength of the relationship between participation and impact (ω) across simulations from 0 to 4.¹¹ This parameter depends in part on the correlation between impacts and inclusion probabilities, which Olsen et al. (2013) demonstrated is directly related to external validity bias.

To determine whether a school would agree to participate, we set a constant agreement probability of 85 percent. This simplification mimics studies where obtaining district approval is the major hurdle, and schools generally agree to participate if the district approves. Given the relationship between school characteristics and district-level average impacts, this framework implied that larger and higher poverty schools were more likely to be in the sample in our simulations because districts with those schools were more likely to agree to participate.

Sample Selection

The algorithm we used to select districts and schools for the evaluation combines stratification with several different sampling approaches. We begin by describing the stratification and sampling approaches before turning to the detailed procedure for sample selection.

Stratification

To ensure adequate representation of different types of schools, we first stratified the population of schools into 18 strata. This stratification used district- and school-level factors that a researcher might hypothesize would moderate impacts. For consistency

¹⁰Most impact studies do not report statistics for the number of districts recruited and that agree to participate. Limited recent evidence suggests a probability of roughly 0.1 at the district level (Gleason et al., 2019; Herrmann et al., 2019).

¹¹Given no empirical evidence on the size of the relationship, we tested various values and explored the resulting variation in district-level probability of participation (P_d) around the target probability. We determined that 0 to 4 was a reasonable range for this parameter (e.g., variability for a scenario with target probability of 0.2 and strength of relationship set to 4—the largest among the values we test—ranged from a minimum probability of 0.07 to a maximum probability of 0.99 with 10th and 90th percentiles of 0.14 and 0.29, respectively).

across strategies, we always stratified the population using the k-means clustering algorithm proposed in Tipton (2013b). Specifically, we:

- **Divided districts into six clusters.**¹² To identify these clusters, we applied k-means clustering as proposed in Tipton (2013b), using district-level averages of two school-level variables (see next bullet) and two district-level variables: (1) Census region and (2) expenditures per pupil. District size and the simulated measure of district leadership, which moderated the impact of the intervention, were omitted to mimic real-world scenarios where important moderators are unobserved or excluded.
- **Divided schools into three clusters.** To identify these clusters, we applied k-means clustering to two school-level variables: (1) the number of students enrolled in the school; and (2) the percentage of students who are FRPL-eligible.
- **Crossed the six district clusters with the three school clusters to form 18 sampling strata.** The total number of schools targeted for the study was apportioned across these 18 strata, with each having its own target number of schools for inclusion in the study.

Sampling Approaches

To select districts and schools for the study, we implemented the following approaches:

1. **Purposive selection.** To select a purposive sample of districts, we rank-ordered districts by their size—the number of eligible schools in the district—and recruited the largest districts first. To select a purposive sample of schools, we rank-ordered schools within participating districts by their size—the number of students enrolled in the school—and recruited the largest schools first. This stylized approach to purposive selection was designed to minimize the time and cost required to recruit a sample of adequate size. While this stylized approach likely ignores other characteristics that evaluators consider in their selection for recruitment, we found that this approach generated samples with districts of similar size as those that Stuart et al. (2017) document as having participated in several large RCTs in education.¹³ These findings suggest that the approach to purposive sampling tested in our study is realistic in the extent to which it favors large districts.
2. **Balanced selection.** To select a balanced sample of districts, we rank-ordered districts based on their multivariate distance from the district cluster mean and recruited the most typical districts first (those with the shortest distance from the cluster mean).¹⁴ To select a balanced sample of schools, we rank-ordered schools

¹²We determined the preferred number of clusters by comparing the “pseudo-F” statistic for different numbers of district- and school-level clusters. See [Appendix A](#) for further detail.

¹³Specifically, purposive selection of schools within purposively selected districts resulted in sampled districts with 103 total schools (not limited to K–5 schools), on average, across our simulated samples. For reference, the average for the total number of schools in our population of districts was 7. Stuart et al. (2017) report that, across 11 large RCTs, the average size of participating districts was 127 schools. See [online appendix Table O.1](#) for additional detail.

¹⁴Our distance metric is based on Gower (1971), which allows us to combine both continuous moderators (number of eligible K–5 schools in the district, expenditures per pupil, district-level average of number of students enrolled in the

based on their multivariate distance from the stratum mean and recruited the most typical schools first. This approach, which extends the balanced sampling method described in Tipton (2013b) to two-stage sampling of districts and schools, is designed to favor typical districts and schools in the selection process.

3. **Random selection.** To randomly select districts, we rank-ordered them using an approach that we developed to mimic sampling with probability proportional to size (PPS)—see [Appendix B](#) for more details. We preferred larger districts (those with a greater number of eligible K–5 schools), giving them a greater probability of appearing higher in the rank-ordered list for recruitment, to reflect practical and cost considerations typically encountered by evaluators. However, to prevent schools in large districts from being heavily overrepresented in the sample, we capped the number of schools per districts at five. To select a random sample of schools, we rank-ordered schools in participating districts purely randomly, independent of size or any other characteristic.

We tested nine different core sample selection strategies—all three approaches to district selection crossed with all three approaches to school selection—to identify the combinations that performed best.

Sample Selection Procedure

For each of the nine core sample selection strategies, we selected a sample by implementing the following steps:

1. **Set sample size targets for each of the 18 sampling strata.** The target number of schools for a particular stratum was set equal to the product of the overall target school sample size and the share of eligible schools in the population that belonged to this stratum.
2. **Rank-order districts and schools within their appropriate groups.** Districts were grouped into the six district clusters and rank-ordered within each cluster. Schools were rank-ordered within groups formed by the combination of district and school cluster (i.e., all eligible schools within a given district were grouped separately by the three school clusters). Each rank order depended on the particular selection strategy implemented.
3. **In any given district cluster, recruit the first district on the rank-ordered list and simulate its participation decision.** To simulate this decision, we generated a random variable using a Bernoulli distribution with the probability from [Equation \(6\)](#), where a value of 1 indicated that the district agreed to participate.
4. **If the district agreed to participate, recruit schools from that district within each school cluster, one at a time, in order.** We skipped a school cluster either if the sample size target for the associated stratum had already been reached or if the district had no eligible schools in the cluster. Otherwise, we simulated school participation by generating a random variable using a Bernoulli distribution with

school, and district-level average of proportion of students eligible for FRPL across schools) and categorical ones (Census region). See [Appendix A](#) for further detail.

a probability of 0.85, where a value of 1 indicated that the school agreed to participate. We continued recruiting schools in the sampling stratum and district until we either (a) exhausted the list of eligible schools in that sampling stratum and district; (b) achieved the target sample of participating schools in that sampling stratum; or (c) five schools in the district agreed to participate.¹⁵ We capped the number of participating schools in a district at five with the goal of ensuring that our simulations included enough participating districts to mimic actual RCTs in education.¹⁶

5. Repeat steps 3-4 for the remaining districts in the district cluster until the target number of participating schools in each of the three strata for this district cluster, as defined by three school clusters, was reached.
6. Repeat steps 3-5 for the other five district clusters until the target number of participating schools in each of the 18 sampling strata was reached.

Testing Replacement Strategies

Our last research question focused on approaches to replacing districts that refused to participate. In general, replacement for the nine core strategies occurred naturally by moving down to the next district or school on the rank-ordered list. However, large-scale assessments in education, such as the NAEP or PISA, use a nearest neighbor approach to replace refusing schools, where the replacement school is selected using the characteristics of the refusing school. Therefore, we also examined the performance of random district selection with nearest neighbor replacement. We implemented this replacement strategy by creating a rank-ordered list of replacement districts using the same distance measure that we used in our approach to balanced selection. However, instead of ranking districts by their distance from the cluster mean, we ranked them by their distance from the vector of characteristics for the refusing district.

Simulation Details

Repeated Sampling

Purposive and balanced district and school selection approaches used a nonrandom, deterministic process to rank-order districts and schools. If all districts and schools agreed to participate, repeated sampling would not be necessary for calculating the SATE for these approaches. However, random selection placed districts and schools in a different random order each time the sampling approach was applied. For any given

¹⁵Because most districts (86%) had fewer than six K–5 schools, the order in which we recruited schools was often immaterial. All schools were typically recruited if the district had fewer than six K–5 schools. See Table O.7 in the [online appendix](#) for an assessment of the extent to which school samples varied by strategy.

¹⁶For districts that had more than five K–5 schools and that had schools in multiple sampling strata, we had to decide how to allocate the five-school cap among the sampling strata. In other words, we had to decide from which strata to recruit first. We chose to allocate the five-school cap based on the percentage of eligible schools by stratum in the district. For example, if a district had 10 schools, 6 in sampling stratum A, 4 in stratum B, and none in stratum C, we would initially target 3 schools for selection in stratum A and 2 in stratum B (by multiplying 5 schools by 60% and 40% for the respective stratum targets). After simulating schools' decisions to participate, we adjusted the allocation if we were unable to meet the targets and if there were additional schools available for recruitment in the district. For instance, if we attempted to recruit all 6 schools in stratum A and only 2 agreed to participate, we would try to recruit an additional school in stratum B.

application of random selection, then, the relative performance of the strategy for that sample was not necessarily reflective of what one should expect on average. Instead, a better assessment of the average performance of random selection comes from taking an average over a large number of iterations—each of which sorted districts and schools differently.

Across iterations of the simulation, we also allowed district and school agreement to participate to vary. This decision implies the SATE varied across iterations for all selection strategies. We allowed agreement to participate to vary to reflect uncertainty about whether districts or schools agreed to participate, conditional on their characteristics, and to account for that uncertainty when we estimated the performance of different selection strategies.

We simulated 1,000 samples for each set of parameter values.¹⁷ Across simulated samples (for a given strategy), all that changed was the draw of the random parameters: the school-specific random effect on impact (μ_{sd}) in Equation (1) (and the corresponding effect on the probability the district agreed to participate), the draws from the Bernoulli distribution that determined agreement to participate, and a random variable we used for rank-ordering districts and schools for random selection and breaking ties in any of the selection strategies.

Measures of Performance

Our approach generated an impact for all schools in the target population. The simple average impact across these schools was the PATE. For any of the strategies, the SATE was the average impact for the selected schools that agreed to participate in the evaluation.¹⁸ We assessed generalizability by comparing the expected SATE to the PATE using the following measures of performance:

- **Mean squared error (MSE)** summarized the bias and variance across R replications together:

$$\begin{aligned} MSE &= \frac{1}{R} \sum_{k=1}^R (SATE_k - PATE)^2 \\ &= \left[\frac{1}{R} \sum_{k=1}^R (SATE_k - PATE) \right]^2 + \frac{1}{R} \sum_{k=1}^R (SATE_k - \frac{1}{R} \sum_{k=1}^R SATE_k)^2 \end{aligned}$$

where the first term equals the bias squared and the second term equals the variance across samples. We preferred this measure of performance to either bias or variance alone because it incorporated both of these components in one measure, and because the source of errors in estimating the PATE (or the relative contributions of bias or variance) is not the focus of our analysis. We answered the first and fourth research questions by

¹⁷We began by conducting 10,000 simulations at baseline parameter values and calculating simulation error for progressively smaller sample sizes. We found that conducting 1,000 simulations reduced simulation error below 0.001 (in fact, 500 or even fewer generated sufficiently small simulation error—see Litwok et al. (2021)). As a result, we report summaries of 1,000 simulations for each set of parameter values throughout the paper. Results using 500 simulations, which appeared in earlier drafts, are available in the [online appendix](#) (and are very similar).

¹⁸We weighted all schools equally for calculating the PATE and the SATE, for simplicity. Reasonable alternatives would have weighted schools by the number of students enrolled in the school or participating in the study.

comparing MSE across different strategies. We separately assessed the absolute value of the bias as well as the standard deviation (square root of variance) to explore the relative contributions of these factors as MSE changed.¹⁹

- **Relative mean squared error (RelMSE)** summarized the MSE relative to some benchmark:

$$RelMSE_s = \frac{MSE_s}{MSE_*}$$

where MSE_s was the mean squared error for strategy s and MSE_* was a benchmark MSE. This measure was useful for exploring the relative performance of strategies as parameters changed, which was the focus of the third research question. Changes to parameters caused the magnitude of MSE to change; however, our interest was in relative performance of different strategies rather than magnitude of MSE. We used random district selection and random school selection as our benchmark and calculated relative MSE to summarize strategy performance over a broad set of parameter values.

In addition to assessing performance in terms of generalizability, we also assessed the recruiting burden associated with each strategy, using the following measures:

- **Average number of districts recruited** summarized the total number of districts recruited before the target school sample size was reached. The number of districts recruited includes both districts that agreed to participate and districts that declined. This total was averaged over 1,000 simulations for each strategy.
- **Average number of schools recruited** summarized the total number of schools recruited, in districts that agreed to participate, before the target school sample size was reached. The number of schools recruited in participating districts includes both schools that agreed to participate and schools that declined. This number does not include schools in districts that declined to participate. The number of schools recruited was averaged over 1,000 simulations for each strategy.

Varying Parameters

As best we could, we aimed to simulate an evaluation that matched published RCTs targeting the K-5 population. Varying parameter values provided the opportunity to test our findings for robustness to assumed values and/or to generate evidence that might be applicable to other environments. The parameters we varied, and the values over which we varied them, were as follows:

- Proportion of variation explained by covariates: 0, 0.1, **0.25**, 0.5, 0.75, 0.9.
- Variance of impacts across schools: 0.0025, **0.01**, 0.04.
- Target school sample size²⁰: 50, **100**, 150, 200.

¹⁹Readers interested in the magnitude of all measures of performance for all parameter values can find the output in the [online appendix](#).

²⁰We explored samples that ranged from 50 schools, as we would expect for typical RCTs, to 200 schools, which was at the very high end of what a large RCT conducted by the federal government would include. For points of comparison, two recent RCTs conducted for IES's National Center for Education Evaluation and Regional Assistance—which sponsors some of the largest RCTs in education—included 146 schools (Balu et al., 2015) and 82 schools (Clark et al., 2013).

- Relationship between latent variable determining participation and impacts: 0, **1**, 2, 4.
- Probability of district agreement to participate: **0.1**, 0.2, 0.3, 0.4.

Our primary findings summarize the simulations under “baseline conditions” (indicated with bold font above). We set baseline conditions to the value most closely aligned with empirical evidence where it was available; in the absence of empirical evidence, we set the baseline condition to be around the midpoint of the range of parameter values we tested. We answered the third research question by varying simulations across all combinations of these parameter values.

Results

Research Question 1: Relative Performance of Strategies

To test our intuition about the relative performance of different selection strategies in the simplest setting, we first imposed that all selected districts and schools would agree to participate if selected (see [online appendix Table O.4](#) for results of these simulations). Simulation results under these conditions were consistent with our expectations. For instance, purposive district selection had the largest MSE due to substantially greater bias than other strategies. The source of this bias was that the largest districts were those with the greatest impact, and the strategies with purposive district selection used only those districts. Absolute bias was near zero for all strategies that used balanced or random district selection. However, as should be expected with random district selection, the variance associated with these strategies was larger than for other strategies.

Next, we repeated the simulation and allowed districts and schools to make participation decisions. Further exploration into the districts and schools selected by the strategies revealed valuable insights about the samples selected by the strategies. Balanced district selection favored schools in smaller districts. This may seem surprising because districts were selected without regard to size in balanced district selection—as opposed to purposive and random district selection, which both explicitly favored larger districts, to different degrees, in the first stage of sampling. One might expect there to be relatively little to no bias around district size for balanced district selection, and that would be mostly correct—at the district level. The average district size in samples formed through balanced district selection (7 eligible schools per district) was closer to the average in the national population of districts (4 eligible schools per district) than was the average for samples formed through random district selection (25 eligible schools per district).

However, while balanced district selection showed less bias than random district selection for district size at the district level, it showed greater bias at the school level. At the school level, the average school in the population was located in a district with 37 eligible schools. In contrast, schools in districts selected via balanced sampling were in much smaller districts (9 eligible schools on average). And schools in districts selected via random sampling were located in districts with 32 eligible schools on average, much closer to the average in the national population.

Table 4. Mean squared error, relative mean squared error, absolute bias, and standard deviation.

Strategy (District/school)	MSE	Relative MSE	Absolute Bias	Standard deviation
Purposive/purposive	0.00119	10.37	0.03222	0.01247
Purposive/balanced	0.00021	1.84	0.00954	0.01101
Purposive/random	0.00025	2.14	0.01141	0.01080
Balanced/purposive	0.00016	1.40	0.00870	0.00926
Balanced/balanced	0.00024	2.11	0.01238	0.00949
Balanced/random	0.00023	2.00	0.01204	0.00926
Random/purposive	0.00029	2.49	0.01215	0.01181
Random/balanced	0.00012	1.01	0.00034	0.01075
Random/random	0.00012	1.00	0.00020	0.01073

Notes. “MSE” = mean squared error. Relative MSE relative to “Random/random.” Parameters set to baseline conditions. Results summarize findings across 1,000 simulations for each strategy.

Source. Author calculations using simulation results.

Table 4 reports the measures of performance in this environment for 1,000 simulated samples. In describing the findings, we refer to “approach A/approach B” for the strategy that used approach A to select districts and approach B to select schools. Under baseline conditions, in which only 10% of recruited districts agreed to participate, MSE was smallest for random/random and random/balanced.

Across all nine core strategies, random/balanced and random/random also had the smallest absolute bias. The magnitude of bias in Table 4 was generally small (less than 0.05 standard deviations) for all strategies. The limited available evidence on the magnitude of external validity bias in the literature suggest bias may be larger than these estimates in practice (Bell et al., 2016).

The findings in Table 4 also demonstrate two important results pertaining to school selection approaches. First, random and balanced school selection performed similarly within districts selected through random or balanced selection. Relative MSE was slightly smaller for random school selection with some district selection approaches (e.g., balanced district selection) and slightly larger for others (e.g., purposive district selection). Second, the performance of purposive school selection was erratic—it performed moderately well for some district selection approaches and poorly for others. For example, it yielded relatively low MSE with balanced district selection; but it yielded relatively high MSE for other district selection approaches.²¹

Robustness Checks

We explored the robustness of the findings to changing some of the simplifying assumptions in the impact model specifications under baseline conditions: (a) we relaxed the linearity of the impact model by making Equation (1) a function of the natural logarithm of total school enrollment (rather than its level); (b) we relaxed the assumption of equal magnitudes of the coefficients in Equation (1) by doubling the coefficients on the number of eligible K–5 schools per district and total school enrollment; and (c) we relaxed the assumption of equal magnitudes of the coefficients in

²¹This was likely due to offsetting biases—balanced district selection systematically chose smaller districts for the evaluation, which were likely to have smaller impacts, while purposive school selection systematically chose larger schools for the evaluation, which were likely to have larger impacts. If the magnitude of the negative bias from balanced district selection was larger than the positive bias from purposive school selection, the resulting absolute bias was smaller than that from the other strategies.

Table 5. Number of districts and schools recruited by strategy.

Strategy (District/school)	Districts Recruited	Districts Agreed	Schools Recruited	Schools Agreed
Purposive/purposive	293.9	29.3	117.4	100.0
Purposive/balanced	293.9	29.3	117.7	100.0
Purposive/random	293.9	29.3	117.5	100.0
Balanced/purposive	613.1	54.7	112.6	100.0
Balanced/balanced	613.1	54.7	112.6	100.0
Balanced/random	613.1	54.7	112.5	100.0
Random/purposive	354.3	33.5	116.5	100.0
Random/balanced	354.3	33.5	116.8	100.0
Random/random	354.3	33.5	116.6	100.0

Notes. Parameters set to baseline conditions. Results summarize findings across 1,000 simulations for each strategy.

Source. Author calculations using simulation results.

Equation (1) by halving these two coefficients. Detailed output appears in the [online appendix](#).

Our main findings were robust to all three checks—the relative ranking of the strategies was nearly unchanged. Relaxing the linearity assumption had little effect on the MSE for any of the selection methods. Doubling and halving the coefficients on district and school size influenced the performance of purposive district selection, which deliberately favored large districts, purposive school selection, which deliberately favored large schools, and balanced district selection, which in practice favored schools in small districts. For each, doubling the coefficients increased absolute bias and MSE by increasing the differences in impacts between selected and unselected districts and schools; halving these two coefficients had the opposite effect. Changing the magnitude of the coefficients did not influence the performance of random district and school selection.

Research Question 2: Recruiting Burden

Separately from external validity outcome measures, the recruiting effort required to implement each of the strategies is critical to researchers who are constrained by a budget. The analysis in [Table 5](#) summarizes, under baseline conditions, the average number of districts and schools recruited to meet the evaluation’s goal of 100 schools participating in the evaluation. All the strategies recruited roughly 115 schools on average to meet the goal of 100 participating schools, which was consistent with the school participation rate of 85 percent. However, there were substantial differences in the number of districts recruited across strategies. Purposive district selection required recruitment of 293.9 districts, on average, to successfully recruit 100 schools into the study. This method yielded the smallest recruitment burden by recruiting the largest districts, which had the largest numbers of eligible schools that could participate. The recruiting burden increased to 354.3 districts for random district selection, which favored large districts, given how the probabilities were set, but to a lesser extent than purposive district selection. Finally, the recruiting burden increased substantially to 613.1 districts for balanced district selection, which did not favor larger districts at all.²²

²²The burden estimates in [Table 5](#) suggested a district agreement rate that was slightly smaller than 10 percent, particularly for balanced and random district selection. The actual rate was lower than the target under baseline conditions because, although some districts might have agreed to participate, those districts were not included in the sample if no schools within the district agreed to participate. Balanced and random selection of districts increased the

Research Question 3: Changes to Baseline Conditions

To explore the extent to which deviations from baseline conditions change the performance of different strategies, we conducted simulations for different values of the parameters. We graphed the changes in performance due to varying a single parameter at a time.

This analysis differed from the preceding analyses under baseline conditions in four ways. First, since [Table 4](#) showed that the different performance measures tended to rank the different methods similarly, this analysis focused exclusively on MSE relative to random/random.²³ This implies that the relative MSE of random/random is always equal to 1. Focusing on relative MSE as opposed to absolute MSE meant that trends in each strategy's performance were influenced by both changes in the MSE of that strategy and changes in the MSE of the benchmark, random/random. This feature is critical to proper interpretation. For instance, the MSE of a particular strategy may have declined monotonically as a parameter changed, yet still produced an increase in relative MSE because of changes in the performance of the benchmark strategy. Second, we eliminated purposive/purposive from the figures in this section. This strategy always performed worst, and the relative MSE for purposive/purposive in [Table 4](#) (10.37) is indicative of the magnitude of the relative performance of this strategy. Such large relative differences required much larger scales in the figures and obscured the differences among the other strategies. Last, we allowed the scales of the figures to vary across parameters. Changes to parameters led to different degrees of variation in relative MSE. Since our focus was on the relative performance of each of the strategies as a single parameter changed rather than the magnitude of relative MSE or comparing across figures, we allowed the scales to change across figures to clarify that relative performance.

The figures that follow change parameters associated with two dimensions of the simulation environment: parameters reflecting impact and parameters reflecting recruitment. [Figures 1](#) and [2](#) focus on parameters reflecting impact—the cross-school impact variance and the share of impact variation explained by covariates—and [Figures 3](#), [4](#), and [5](#) focus on parameters reflecting recruitment—the number of schools, the district participation rate, and the strength of the relationship between impact and district participation decisions.

[Figure 1](#) summarizes how estimates of relative MSE change with changes to the variance of impact across schools. The baseline condition for cross-school impact variance was 0.01 (obtained by squaring a standard deviation of 0.1). [Figure 1](#) demonstrates that the relative MSE for random/random and random/balanced were consistently small and often the smallest across all values of cross-school impact variance. With a smaller value of cross-school impact variance (0.0025) the relative performance of all other strategies was similar. As cross-school impact variance grew (0.04), the relative performance of all strategies with balanced district selection improved. This relationship is explained by changes to the absolute MSE for random/random. As cross-school impact variance increased, so did the absolute MSE for random/random (largely due to increases in variance). The relative performance for strategies that used balanced district selection

likelihood of selecting smaller districts, in which a single school choosing not to participate might have resulted in the district not participating in the study.

²³Results for other measures of performance appear in the [online appendix](#).

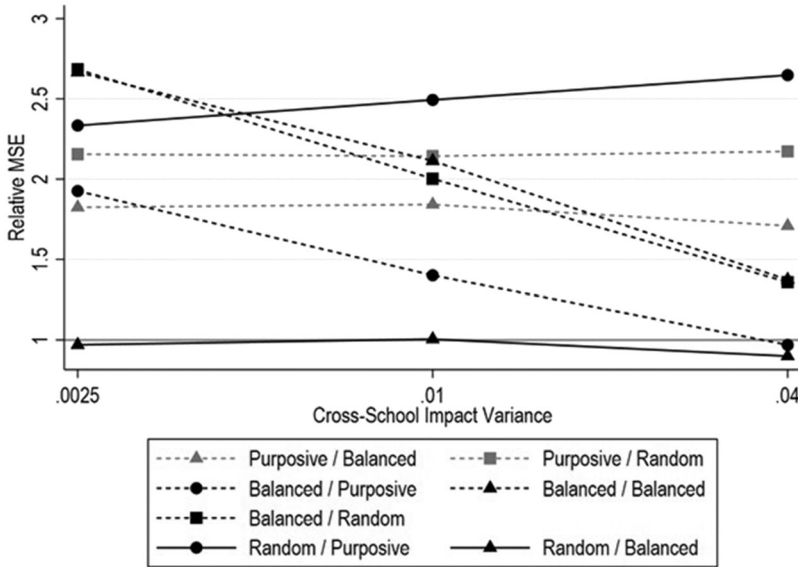


Figure 1. Changes to relative MSE with cross-school impact variance. *Note.* Figure reports changes to relative MSE as cross-school impact variance changed. All other parameters fixed at baseline conditions. Solid gray line at relative MSE of 1 reflects performance of random/random. Figures exclude purposive/purposive because the relative MSE for this strategy is much larger than all other strategies and would distort the figure. See [online appendix](#) for full output for all strategies. *Source.* Author calculations using simulation results.

improved because, while the absolute MSE also increased for these strategies, it increased at a smaller rate.

Figure 2 summarizes how relative MSE changed with the proportion of impact variation explained by covariates. Relative MSE was close to 1 for all strategies when covariates did not explain any variation in impacts. As the proportion of variation explained by covariates grew, the relative MSE grew for all strategies other than random/balanced and random/random. That relative MSE grew in this way—and fairly consistently across strategies—implies the findings under baseline conditions (where 25 percent of impact variation was explained by covariates) were robust to changes in this parameter. This growth in MSE was due primarily to growth in bias, rather than variance (see [online appendix Tables O.10a and O.10b](#)). This was consistent with our expectations—because balanced and purposive strategies intentionally excluded characteristics that determined impacts (balanced excluded administrator leadership and district size while purposive excluded everything other than district/school size, see *Sampling Approaches*”), we expected bias to grow as those omitted variables explained a larger and larger share of cross-school impact variation.

Next, we examined the importance of factors related to selecting an adequate sample—specifically, the target school sample size, the district participation rate, and the strength of the relationship between district-level impacts and agreement to participate.

Figure 3 varies the number of schools targeted for selection. As with [Figures 1 and 2](#), the strong performance of random/random and random/balanced was robust to the

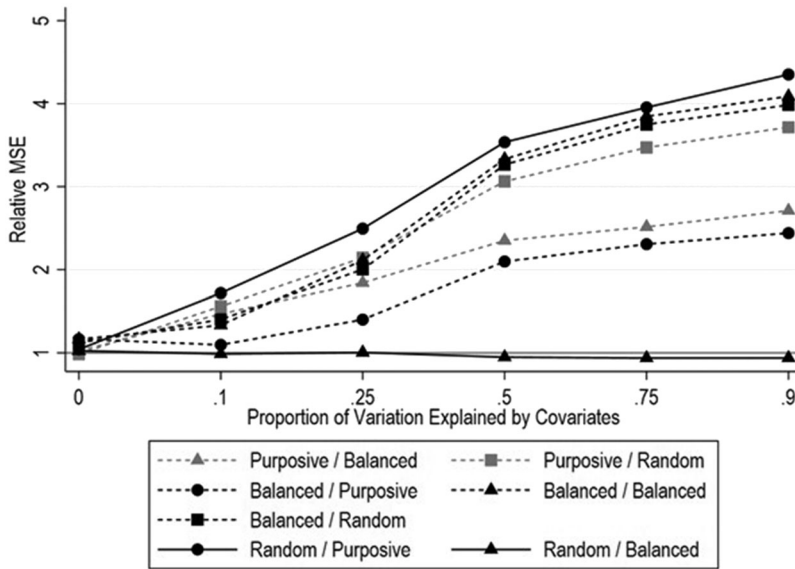


Figure 2. Changes to relative MSE with proportion of variation explained by covariates. *Note.* Figure reports changes to relative MSE as proportion of variation explained by covariates changed. All other parameters fixed at baseline conditions. Solid gray line at relative MSE of 1 reflects performance of random/random. Figures exclude purposive/purposive because the relative MSE for this strategy is much larger than all other strategies and would distort the figure. See [online appendix](#) for full output for all strategies. *Source.* Author calculations using simulation results.

total number of schools included in the evaluation. At the smallest number of schools (50), [Figure 3](#) shows similar relative MSE for all strategies other than purposive/balanced, purposive/random, and random/purposive. As the number of schools grew, the MSE of the three balanced district selection strategies increased relative to the MSE of the benchmark selection method, random district selection with random selection of schools. This is because the variance of the impact estimates declined by more for random district selection than for balanced district selection as the number of districts in the sample increased to recruit additional schools ([online appendix, Table O.9b](#)). Meanwhile, the relative performance of random/purposive did not change substantially, and the relative MSE of purposive/balanced and purposive/random declined as the target number of schools increased—presumably because the average size of participating districts declined as the algorithms worked further down the recruitment list to smaller districts, in order to meet the demands of a greater sample size.²⁴

[Figure 4](#) varies the district participation rate. Unlike other parameters, the relative performance for most strategies was largely unaffected by changes to the district participation rate (meaning the overall findings were also robust to changes in this parameter). There were two clear exceptions: purposive/balanced and purposive/random.

²⁴Average district size declines sharply with the target number of schools. On average, participating districts recruited through purposive district selection had 61 eligible schools for a target of 50 schools, 47 eligible schools for a target of 100 schools, 40 eligible schools for a target of 150 schools, and 35 eligible schools for a target of 200 schools.

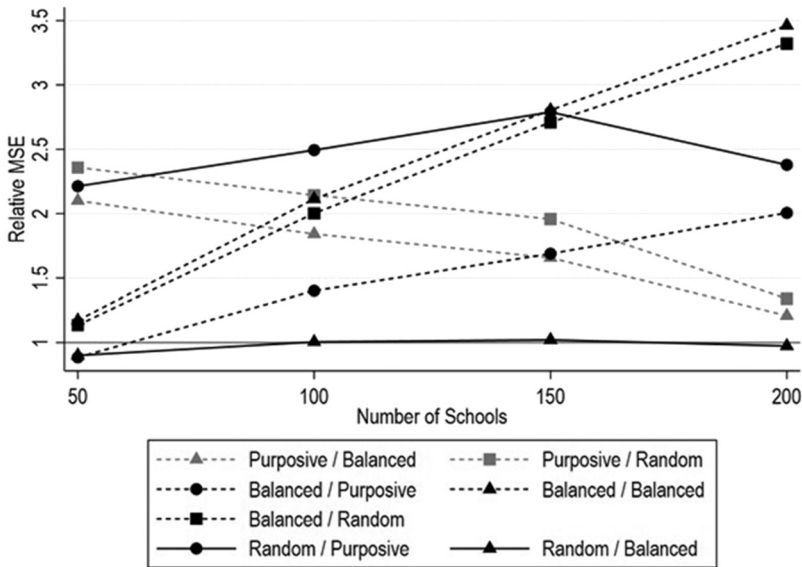


Figure 3. Changes to relative MSE with number of schools. *Note.* Figure reports changes to relative MSE as number of schools changed. All other parameters fixed at baseline conditions. Solid gray line at relative MSE of 1 reflects performance of random/random. Figures exclude purposive/purposive because the relative MSE for this strategy is much larger than all other strategies and would distort the figure. See [online appendix](#) for full output for all strategies. *Source.* Author calculations using simulation results.

For these strategies relative MSE grew as the district participation rate grew. This finding can be attributed to the fact that as the district participation rate increases, the study needs to recruit fewer districts to reach the target number of schools. And recruiting fewer districts purposively means recruiting larger districts—districts higher on the list ranked in descending order of size—which are less typical of districts nationwide.

Figure 5 shows how relative MSE varied with changes to the final parameter—the strength of the relationship between district participation and average district impact. The behavior of random/random and random/balanced remained among the strongest across all values of the parameter. However, the relative performance of balanced district selection improved as the relationship between participation and impact strengthened (and not just in relative terms—balanced selection was the only district selection strategy where the MSE consistently decreased in absolute terms as the relationship between participation and impact strengthened). Indeed, for the largest value of this relationship (4), the performance of the balanced strategies was stronger than random/random and random/balanced. Increasing the value of this parameter implies districts with larger average impacts are more likely to agree to participate; the results in Figure 5 imply balanced district selection counterbalanced this bias by favoring smaller districts (which tended to have smaller impacts).

In summary, the primary findings from our analysis—that random/random and random/balanced strategies performed best in terms of relative MSE—were robust to

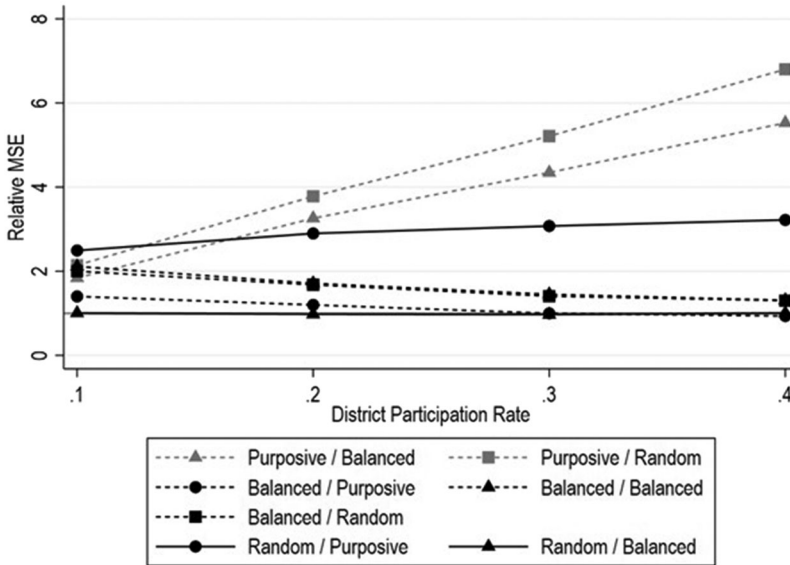


Figure 4. Changes to relative MSE with district participation rate. *Note.* Figure reports changes to relative MSE as district participation rate changed. All other parameters fixed at baseline conditions. Solid gray line at relative MSE of 1 reflects performance of random/random. Figures exclude purposive/purposive because the relative MSE for this strategy is much larger than all other strategies and would distort the figure. See [online appendix](#) for full output for all strategies. *Source.* Author calculations using simulation results.

changes in baseline conditions. Under certain conditions (such as a small sample size or a strong relationship between impact and participation), other strategies had similar or slightly better performance, but the relative MSE for those strategies did not drop much below 1. These findings indicate that random/random and random/balanced were the most consistent performers across the combinations of parameters we tested.

Research Question 4: District Replacement

To explore the performance of methods for replacing districts that declined to participate, we limited the focus to the strategies that used random district selection, modified the algorithm for replacing declining districts, and produced the same evidence we used to answer the first research question.

Table 6 reports performance measures across 1,000 simulated samples for random district selection with random replacement and with nearest neighbor replacement (analogous to Table 4). When paired with balanced or random school selection, Table 6 shows that nearest neighbor replacement of nonparticipating districts did not improve performance relative to random replacement. Comparing these rows in Table 6 demonstrates that nearest neighbor replacement reduced the variance across samples; however, it increased the magnitude of bias such that it increased the relative MSE. The increase in bias may be explained by selection of nearest neighbor districts without regard for

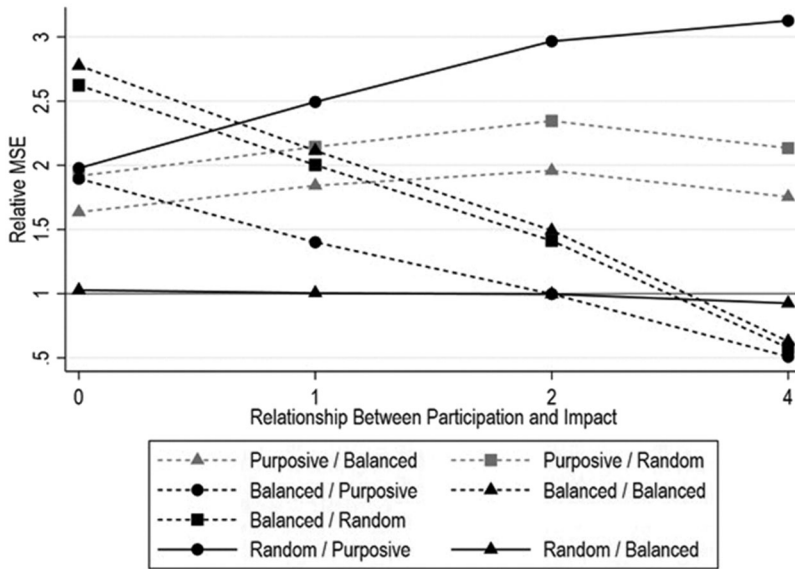


Figure 5. Changes to relative MSE with Relationship between District Participation and Impacts. *Note.* Figure reports changes to relative MSE as relationship between district participation and impact changed. All other parameters fixed at baseline conditions. Solid gray line at relative MSE of 1 reflects performance of random/random. Figures exclude purposive/purposive because the relative MSE for this strategy is much larger than all other strategies and would distort the figure. See [online appendix](#) for full output for all strategies. *Source.* Author calculations using simulation results.

Table 6. Mean squared error, relative mean squared error, absolute bias, and standard deviation by replacement strategy.

Strategy (District/school)	MSE	Relative MSE	Absolute Bias	Standard deviation
Random (random replacement)/purposive	0.00029	2.49	0.01215	0.01181
Random (random replacement)/balanced	0.00012	1.01	0.00034	0.01075
Random (random replacement)/random	0.00012	1.00	0.00020	0.01073
Random (nearest neighbor replacement)/purposive	0.00012	1.08	0.00113	0.01108
Random (nearest neighbor replacement)/balanced	0.00016	1.36	0.00681	0.01048
Random (nearest neighbor replacement)/random	0.00015	1.32	0.00658	0.01041

Notes. “MSE” = mean squared error. Relative MSE relative to “Random (random replacement)/random.” Parameters set to baseline conditions. Results summarize findings across 1,000 simulations for each method.

Source. Author calculations using simulation results.

their size—a characteristic that mediated impact by construction.²⁵ This result differed from pairing nearest neighbor replacement of nonparticipating districts with purposive school selection, which reduced both the magnitude of bias and variance relative to random replacement leading to a large reduction in relative MSE.²⁶

²⁵Nearest neighbor replacement of nonparticipating districts used the same distance metric as balanced district selection. Therefore, since balanced district selection performed somewhat worse than random district selection, it is perhaps not surprising that nearest neighbor district replacement performed somewhat worse than random district replacement.

²⁶The surprisingly strong performance of purposive school selection—when paired with random district selection with nearest neighbor replacement—can likely be attributed to offsetting biases from a sampling process that favored large schools, generating a positive bias, in small districts, generating a negative bias. Purposive school selection favored large schools by design. Nearest neighbor replacement favored small districts for the same reason as balanced district

Recruiting burden was also relevant for nearest neighbor replacement. When randomly selected districts that refuse to participate were replaced with their nearest neighbor, the number of districts recruited to meet the same school target (i.e., the analog to Table 5) was considerably larger than the number for random selection with random replacement—an evaluation would need to recruit 598 districts with nearest neighbor replacement rather than 354 with random replacement. This could be because nearest neighbor districts were selected without regard for their size, while random replacements were selected with probabilities proportional to their size, like the districts they replaced. Alternatively, it could be the case that nearest neighbors for districts unlikely to agree to participate are also unlikely to agree to participate (whereas a randomly selected replacement district may be more likely to agree to participate).

Last, we also explored the extent to which the findings in Table 6 were robust to changes in baseline conditions. As with the earlier analysis of the third research question, we found that the stronger performance of random replacement over nearest neighbor replacement was robust to most changes in baseline conditions. The exceptions were for the highest level of cross-school impact variance (0.04) and the strongest relationship between impact and district agreement to participate (4). In these cases, nearest neighbor replacement had lower MSE. Detailed results for these analyses appear in the [online appendix](#).

Discussion and Conclusion

This article generated empirical evidence on the performance of prospective strategies for selecting a sample of schools to participate in an impact study—with performance measured in terms of generalizability to a target population of schools. We tested these strategies in an environment where researchers first selected districts and then selected schools to participate. The strategies consisted of a variety of approaches, both those recommended in the literature—random and balanced selection—and a stylized approach that matched the characteristics of districts that typically participate in large RCTs—purposive selection. Importantly, random selection of districts was not a simple random sample within strata from the target population (which evaluators may not be comfortable with). Our approach favored larger school districts by imposing that their probability of selection was proportional to district size.

In an environment where districts chose whether to participate in the evaluation, the results of our analysis were nearly unequivocal. Strategies that combined random district selection with either random or balanced school selection fared better than other strategies for all external validity outcome measures we explored (mean squared error, relative mean squared error, absolute bias, and standard deviation). Among those other strategies purposive/purposive clearly had the worst performance, and balanced district selection required so many more districts than random district selection (without meaningfully improving external validity outcomes) as to be impractical for any evaluation with a budget constraint. The findings held among nearly all combinations of parameter

selection: neither approach deliberately favored small districts, but the cap of five schools per district constrained the representation of large districts in the sample.

values we tested and a wide range of possible values. These parameters captured characteristics of impact—cross-school variance in impacts and share of impact variation explained by covariates—as well as characteristics of the evaluation—target number of schools, district participation rate, and relationship between impact and participation.

We also tested whether replacing districts that refused to participate in the evaluation with their nearest neighbors performed better than random replacement. While nearest neighbor replacement sharply reduced the variance of the SATE across samples, the corresponding increase in bias resulted in worse performance in terms of relative MSE. Similar to the overall findings comparing selection strategies, this finding held among many combinations of parameter values we tested.

These results have important implications for the common practice of favoring large districts (and large schools within those districts) in the design of RCTs in education. Favoring large districts may be appealing for cost reasons because the study can achieve its target number of schools or students by recruiting only a few districts. But our simulations demonstrate that favoring the largest districts can lead to impact estimates with a high degree of external validity bias, consistent with Bell et al. (2016; though bias is of a smaller magnitude in our simulations than in Bell et al. (2016)). Our findings suggest this bias can be reduced by favoring large districts to a lesser degree (by selecting districts with probabilities proportional to size) or not at all (by selecting a balanced sample of districts to match the population, which may include a lot of small districts). Reducing external validity bias results in evaluation evidence that is more relevant to policymakers because the findings generalize to a well-defined target population.

At present, the two-stage sampling approaches tested in our simulations—first selecting districts, and then selecting an uncertain number of schools within those districts—may be challenging for education researchers to implement. The practical challenges to implementing the strategies that were found to improve generalizability in our simulations imply that the field would benefit from additional training and resources, including computer code, to help implement these strategies. In that regard, we have made our statistical code available in our repository on the Open Science Framework.²⁷

The questions addressed in our simulations warrant additional research to determine the extent to which our findings generalize to scenarios not covered in our simulations. For example, we defined the target population as the full national population of schools, but we recognize that many impact studies may have a different target population in mind (such as rural elementary schools). A researcher may even prefer to define the target population in terms of students rather than schools. Our simulation explored biases relative to one plausible target population, and the performance of different strategies for generalizing to other populations is unknown.

We also made a variety of other choices in designing our simulation that were plausible, but where alternative choices would also be plausible. Our simulation focused on scenarios where the main threat to the generalizability of the study findings stemmed from district decisions about whether to participate in the study. However, scenarios exist in which the main threat stems from schools. We also chose to associate larger schools with larger impacts for the hypothetical intervention we studied—a choice that

²⁷See <https://osf.io/fehjc>.

aimed to create external validity bias consistent with what has been observed in prior studies, but one that could have led our simulations to overestimate the negative consequences of purposive sampling in practice. For example, the surprisingly strong performance of balanced district selection (and random district selection with nearest neighbor replacement) with purposive school selection may also be overstated by the choices we made. This selection strategy appeared to benefit from directionally opposed biases since balanced district selection favored smaller districts (which, by design, were associated with smaller impacts), and purposive school selection favored larger schools (which, by design, were associated with larger impacts). Additional testing of the sensitivity of these choices as well as deeper analysis of the simulation results (e.g., of the interaction between changes in multiple parameters at once) are areas for future research. Furthermore, future research might combine the prospective strategies we tested with the retrospective strategies that are gaining prominence in the literature to see whether differences in the performance of different strategies can be addressed at the analysis stage.

In closing, the field of applied educational research has moved toward incorporating experimental methods that minimize bias for internal validity. Our findings demonstrated that researchers' decisions about how to select and recruit districts and schools can reduce the amount of external validity bias in the study's impact estimate. Our findings also showed that relatively simple adjustments to the process of selecting schools and districts, such as incorporating randomness, resulted in a sample that better represented the target population of interest.

Open Research Statements

Study and Analysis Plan Registration

The study and analysis plan are registered on the Open Science Framework (<https://osf.io/jcy8a>, <https://osf.io/j2p9v>).

Data, Code, and Materials Transparency

The data and code that support the findings of this study are publicly available on the Open Science Framework (<https://osf.io/fehjc>). An immutable version can be found here: <https://osf.io/9w6ed>.

Design and Analysis Reporting Guidelines

This manuscript was not required to disclose use of reporting guidelines, as it was initially submitted prior to JREE mandating open research statements in April 2022.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects

of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Replication Statement

This manuscript reports an original study.

Open Scholarship



This article has earned the [Center for Open Science](#) badges for Open Data, Open Materials and Preregistered through Open Practices Disclosure. The data and materials are openly accessible at <https://osf.io/9w6ed>, <https://osf.io/jcy8a>, and <https://osf.io/j2p9v>.

Funding

The research reported here is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D190020 to Westat. This work benefitted from substantive discussions with Stephen Bell, Larry Orr, Matthew Soldner, Elizabeth Stuart, and Elizabeth Tipton. Utsav Kattel provided outstanding research support.

References

- Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, 130(3), 1117–1165. <https://doi.org/10.1093/qje/qjv015>
- Andrews, I., & Oster, E. (2018). *Weighting for external validity* (No. w23826). National Bureau of Economic Research.
- Litwok, D., Nichols, A., Shivji, A., & Olsen, R. (2021). *Selecting districts and schools for impact studies in education: A simulation study of different strategies*. Revised Analysis Plan. Available at: <https://osf.io/ja6ne>
- Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). *Evaluation of response to intervention practices for elementary school reading (NCEE 2016-4000)*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics*, 1(1), 151–178. <https://doi.org/10.1146/annurev.economics.050708.143235>
- 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. <https://doi.org/10.3102/0162373715617549>
- Bell, S. H., & Stuart, E. A. (2016). On the “where” of social experiments: The nature and extent of the generalizability problem. *New Directions for Evaluation*, 2016(152), 47–59. <https://doi.org/10.1002/ev.20212>
- Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, 3(1), 1–27. <https://doi.org/10.1080/03610927408827101>
- Chan, W. (2017). Partially identified treatment effects for generalizability. *Journal of Research on Educational Effectiveness*, 10(3), 646–669. <https://doi.org/10.1080/19345747.2016.1273412>
- Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The effectiveness of secondary math teachers from Teach for America and the Teaching*

- Fellows Programs (NCEE 2013-4015)*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Fellers, L. A. (2017). *Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences* [Doctoral dissertation]. Columbia University.
- 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: Appendices. NCEE 2019-4008*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857–871. <https://doi.org/10.2307/2528823>
- Halpin, B. (2016). *Cluster analysis stopping rules in Stata*. University of Limerick Department of Sociology Working Paper Series. Working Paper WP2016-01.
- Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, 9(2), 85–110. <https://doi.org/10.1257/jep.9.2.85>
- 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*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5–86. <https://doi.org/10.1257/jel.47.1.5>
- 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. <https://doi.org/10.1080/19345747.2015.1060282>
- Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., & Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, 11(1), 225–247. <https://doi.org/10.1214/16-AOAS1001>
- 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. <https://doi.org/10.1002/ev.20207>
- Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of policy analysis and management*, 32(1), 107–121. <https://doi.org/10.1002/pam.21660>
- O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society*, 63(2), 195–210.
- Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods*. SAGE.
- Shadish, W. R., Cook, T. D., & Cmapbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- StataCorp. (2019). *Multivariate statistics reference manual*. Stata Press.
- Stuart, E. A., & Rhodes, A. (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation Review*, 41(4), 357–388. <https://doi.org/10.1177/0193841X16660663>
- 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*, 174(2), 369–386. <https://doi.org/10.1111/j.1467-985X.2010.00673.x>
- 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. <https://doi.org/10.1080/19345747.2016.1205160>
- Tipton, E. (2013a). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3), 239–266. <https://doi.org/10.3102/1076998612441947>

- Tipton, E. (2013b). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, 37(2), 109–139. <https://doi.org/10.1177/0193841X13516324>
- Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Ruiz de Castilla, V. (2016). Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness*, 9(sup1), 209–228. <https://doi.org/10.1080/19345747.2015.1105895>
- Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, 41(5), 472–505. <https://doi.org/10.1177/0193841X16655665>
- Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, 7(1), 114–135. <https://doi.org/10.1080/19345747.2013.831154>
- Tipton, E., & Matlen, B. J. (2019). Improved generalizability through improved recruitment: Lessons learned from a large-scale randomized trial. *American Journal of Evaluation*, 40(3), 414–430. <https://doi.org/10.1177/1098214018810519>
- Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, 47(8), 516–524. <https://doi.org/10.3102/0013189X18781522>
- Tipton, E., & Olsen, R. B. (2022). *Enhancing the generalizability of impact studies in education*. (NCEE 2022-003). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Tipton, E., & Peck, L. R. (2017). A design-based approach to improve external validity in welfare policy evaluations. *Evaluation Review*, 41(4), 326–356. <https://doi.org/10.1177/0193841X16655656>
- Tipton, E., Spybrook, J., Fitzgerald, K., 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. <https://doi.org/10.3102/0013189X20960686>
- 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. <https://doi.org/10.1080/19345747.2017.1300719>

Appendix A. Balanced site selection details

This appendix describes two components of how we implemented balanced site selection: the distance metric and the number of strata.

Distance Metric

The variables we used in balanced selection of districts were both continuous (expenditures per pupil, district-level average of number of students enrolled in the school, and district-level average of proportion of students eligible for FRPL across schools) and categorical (Census region). The variables we used in balanced selection of schools were continuous (number of students enrolled in the school and proportion of students eligible for FRPL).

When the desired set of characteristics includes both continuous and categorical measures, Tipton (2013b) recommends a distance measure based on Gower (1971). We used this measure as our distance metric at both the district and school levels.

Selecting Number of Strata

Tipton (2013b) recommends using k-means clustering to generate the strata for balanced site selection. Implementing k-means clustering requires the researcher to select a number of strata

(or the value of k). One option for determining the value of k is to test several different values, generate results under each of the values, and compare the findings.

The “pseudo-F” statistic is one measure Tipton (2013b) recommends for determining the optimal number of clusters. This approach, attributed to Caliński and Harabasz (1974), reports the “variance ratio criterion” for each value of k . The variance ratio criterion compares the sum of squared distances within the partitions to that in the unpartitioned data, taking account of the number of clusters and cases (Halpin, 2016). Values of k with large variance ratio criterion, which is analogous to the pseudo-F statistic, have the most distinct cluster structure (StataCorp, 2019).

To determine the optimal number of district clusters, we computed the pseudo-F statistic for values of k ranging from 2 to 10 for district-level clustering based on Census region, expenditures per pupil, district-level average of number of students enrolled in the school, and district-level average of proportion of students eligible for FRPL across schools. We repeated the analysis for school-level clustering based on number of students enrolled in the school and proportion eligible for FRPL. Results from those analyses appear in Tables O.2 and O.3 in the [online appendix](#).

At the district level, the evidence indicated the most distinct cluster structure when k was set to 2 or 6. Although it was not the largest value of the pseudo-F statistic, we chose to set k equal to 6. We felt that two strata were too few to adequately balance these variables. Furthermore, Tipton and Olsen (2018, 2022) argue that in education experiments defining between four and six strata is often a good compromise. At the school level, we set k equal to 3. We chose this value because it both maximized the pseudo-F statistic and we felt it was a sufficient number of strata for our school-level analysis.

Appendix B. Operationalizing random district selection

This appendix describes our approach to random district selection. We based this approach on random sampling of districts with probability of selection proportional to district size (PPS sampling).

Traditional PPS sampling is designed to select a fixed number of districts at random. Because the target number of districts is unknown (it is a function of the rates of agreement to participate at the district and school levels), traditional PPS sampling was not feasible for our analysis. Instead, we aimed to develop an approach that similarly selects districts randomly with a probability proportional to district size and generates a rank-ordered list of districts for recruitment (so the evaluators could work down the list until the target number of schools is reached).

Our approach used a district-level lottery. Within each district cluster, we determined the total number of eligible schools for each district (N_d), calculated the total number of eligible schools across all districts ($N = \sum_{d=1}^D N_d$), randomly ordered districts, and assigned a range of integers to each district based on the number of eligible schools in the district (e.g., $(1, 2, \dots, N_1)$ for the first district, $(N_1 + 1, N_1 + 2, \dots, N_1 + N_2)$ for the second district, and so on). For example, if a district cluster consisted of two districts—district A with 4 schools and district B with 12 schools, then we would create a unique list of numbers from 1 to 16, and district A would be assigned the range of 1–4 and district B would be assigned the range of 5–16.

Next, we generated a rank-ordering for districts by randomly selecting from the range of integers (without replacement) and matching the integer to its district. This random sorting resulted in the same district appearing on the list more than once. If we reached a district that had already been selected as we proceeded down the list, we moved on to the next district in the list. Returning to the example from the prior paragraph, if the first random draw was 15, district 2 would be the first district selected. If the second draw was 7, we would proceed down the list because we already selected district 2. If the third draw was 2, district 1 would be the second district selected.

We tested this approach using a small simulation in which we sampled 50 districts from a population of 998. We imposed that each district had a different preset selection probability—in

one case using PPS sampling and in a second case using the lottery approach described above.²⁸ We repeated both algorithms for 1,000 samples and calculated actual selection rates for each method as the number of times selected divided by the number of samples. Figure O.1 in the [online appendix](#) graphs these selection rates for the two approaches. In addition to the graphical similarity, we calculated the correlation between the selection rates and the predetermined selection probability to be 0.95 for both approaches. Given these results, we felt comfortable proceeding with the lottery approach as a method to produce an ordered list akin to PPS sampling.

²⁸We created this population by splitting the range from 0 to 1 into 1,000 unique values (inclusive of 0 and 1). The preset selection probability for each district corresponded to one of the values from this range (though we omitted 0 and 1, resulting in a total population of 998).